Sex Differences in Mathematical Ability: Fact or Artifact?

Author(s): Camilla Persson Benbow and Julian C. Stanley

Source: *Science*, Dec. 12, 1980, New Series, Vol. 210, No. 4475 (Dec. 12, 1980), pp. 1262-1264

Published by: American Association for the Advancement of Science

Stable URL: https://www.jstor.org/stable/1684489

dysfunction or disorder. While negative findings in such examinations of children with behavioral or learning problems must be considered tentative and inconclusive, positive findings would justify referral for more exhaustive evaluations.

H. AHN
L. PRICHEP, E. R. JOHN
Brain Research Laboratories, Department of Psychiatry, New York University Medical Center, New York 10016

H. BAIRD
St. Christopher's Hospital for Children, Department of Pediatrics, Temple University School of Medicine, Philadelphia, Pennsylvania 19122

M. TREPETIN
Neurometrics, Inc., New York 10038

H. KAYE
Department of Psychology, State University of New York at Stony Brook, Stony Brook 11790

References and Notes

1. E. R. John, H. Ahn, L. Prichep, M. Trepetin, D. Brown, H. Kaye *Science* **210**, 1255 (1980).
2. The equations are first-order functions whose coefficients are published in (*1*). Frequency bands were defined as delta (1.5 to 3.5 Hz), theta (3.5 to 7.5 Hz), alpha (7.5 to 12.5 Hz), and beta (12.5 to 25 Hz). The derivations were F_7T_3, F_8T_4, T_3T_5, T_4T_6, C_3C_z, C_4C_z, P_3O_1, and P_4O_2, according to the nomenclature of the international 10/20 system.
3. These terminals correspond to that described in (*4*, *5*), but we used an LSI-11 microprocessor instead of a PDP 11/10 and recorded digital data on floppy disks instead of magnetic tape. Frequency and amplitude calibration signals were recorded regularly to guarantee standardized recording conditions.
4. E. R. John *et al.*, *Science* **196**, 1393 (1977).
5. E. R. John, *Functional Neuroscience*, vol. 2, *Neurometrics: Clinical Applications of Quantitative Electrophysiology* (Erlbaum, Hillsdale, N.J., 1977).
6. Board of Cooperative Educational Services (BOCES) District III, James E. Allen Learning Center, Dix Hills, N.Y.
7. Pediatric Neurology Service, Handicapped Children's Unit, St. Christopher's Hospital for Children, Philadelphia. Supported in part by NIH general CRC grant RR-75. Disks from the neurometric examinations of 474 neurological patients referred to this service were sent to NYU for analysis with no information other than the age of the patient. This terminal was constructed by Neurometrics, Inc., under license from NYU.
8. This terminal was used to gather data on a sample of 129 children who were exposed to malnutrition in the first year of life. They were matched by age, grade, gender, and handedness to a control sample of 129 children who had not suffered from malnutrition. The study (manuscript in preparation) was conducted in collaboration with F. Ramsey, J. Galler, and G. Solimano and was supported by the Ford Foundation, grant 770-0471. The analyses reported here refer only to a subset of the control population (see group 2, Barbados normals).
9. Data gathered at the Rockland Psychological and Educational Center in Spring Valley, N.Y.
10. Data gathered at the Applied Neuroscience Institute, University of Maryland, Eastern Shore, Princess Anne.
11. The normal children examined at BOCES (*6*) and at NYU were studied in a project supported by National Science Foundation grant DAR 78-18772, formerly APR 76-24662, intended to provide part of the data base for construction of an EEG/evoked response discriminant function capable of separating learning disabled from normal children.
12. The learning disabled children were examined in a project cited in (*11*) and in a project supported by the Office of Education, Bureau of Education for the Handicapped (grant G007604516), in which neurometric methods are used to diagnose and help remediation of the learning disabled child.
13. All U.S. normal children (group 1) had scores of 90 or higher on the Peabody Picture Vocabulary Test (PPVT) and standard scores of 90 or higher on all sections (reading, spelling, and arithmetic) of the Wide Range Achievement Test (WRAT). The Barbados normal children (group 2) were the subset of the control group (*8*) which had full-scale Wechsler Intelligence Scale for Children (WISC) scores of 85 or higher and appropriate grade level for age. The WISC was modified by J. Galler to make it culturally relevant for Barbadian children. The learning disabled children (group 4) had IQ scores between 65 and 84 on the WISC-R and WRAT standard scores below 90 in language or arithmetic skills, or both. The specific learning disabled children (group 5) had IQ scores above 85 on the WISC-R, and WRAT standard scores below 90 in language and or arithmetic skills. In groups 4 and 5, PPVT scores were used when WISC-R scores were not available. Most of the children in groups 4 and 5 were attending a special school (*6*) for children unable to learn satisfactorily in their local schools.
14. With the exception of a few children in group 3 with convulsive disorders, none of the children in any group received medication for at least 72 hours prior to examination.
15. On the basis of relative power, one of four frequency bands is a linear combination of the others. Thus, using exact probabilities of χ^2 distributions, we calculated Bonferroni significance levels (*16*) based on 24 independent measures. The critical levels of exact probabilities corresponding to nominal P values ranging from $P \leq .05$ to $P \leq .0001$ were as follows: $P \leq .05$ = 2E-3; $P \leq .01$ = 4E-4; $P \leq .001$ = 4E-5; and $P \leq .0001$ = 4E-6.
16. W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1977), vol. 1.
17. F. A. Gibbs and E. L. Gibbs, *Atlas of Encephalography*, vol. 3, *Neurological and Psychological Disorders* (Addison-Wesley, Reading, Mass., 1964).
18. O. Eeg-Olafsson, *Acta Paediatr. Scand. Suppl. 208* (1970).
19. The percentage of hits for each of the 32 measures was averaged. The mean percentage of individual hits was 20, 21, and 20 percent at the $P \leq .05$ level and 12, 11, and 10 percent at the $P \leq .01$ level for groups 3, 4, and 5, respectively.
20. H. Ahn, thesis, University of Iowa (1977).
21. We acknowledge the assistance of L. Valencia, M. Flanders, S. Lobel, E. Mason, P. Clark, A. Toro, and S. Balbontin.

17 September 1979; revised 28 August 1980

Sex Differences in Mathematical Ability: Fact or Artifact?

Abstract. *A substantial sex difference in mathematical reasoning ability (score on the mathematics test of the Scholastic Aptitude Test) in favor of boys was found in a study of 9927 intellectually gifted junior high school students. Our data contradict the hypothesis that differential course-taking accounts for observed sex differences in mathematical ability, but support the hypothesis that these differences are somewhat increased by environmental influences.*

Huge sex differences have been reported in mathematical aptitude and achievement (*1*). In junior high school, this sex difference is quite obvious: girls excel in computation, while boys excel on tasks requiring mathematical reasoning ability (*1*). Some investigators believe that differential course-taking gives rise to the apparently inferior mathematical reasoning ability of girls (*2*). One alternative, however, could be that less well-developed mathematical reasoning ability contributes to girls' taking fewer mathematics courses and achieving less than boys.

We now present extensive data collected by the Study of Mathematically Precocious Youth (SMPY) for the past 8 years to examine mathematical aptitude in approximately 10,000 males and females prior to the onset of differential course-taking. These data show that large sex differences in mathematical aptitude are observed in boys and girls with essentially identical formal educational experiences.

Six separate SMPY talent searches were conducted (*3*). In the first three searches, 7th and 8th graders, as well as accelerated 9th and 10th graders, were eligible; for the last three, only 7th graders and accelerated students of 7th grade age were eligible. In addition, in the 1976, 1978, and 1979 searches, the students had also to be in the upper 3 percent in mathematical ability as judged by a standardized achievement test, in 1972 in the upper 5 percent, and in 1973 and 1974 in the upper 2 percent. Thus, both male and female talent-search participants were selected by equal criteria for high mathematical ability before entering. Girls constituted 43 percent of the participants in these searches.

As part of each talent search the students took both parts of the College Board's Scholastic Aptitude Test (SAT)—the mathematics (SAT-M) and the verbal (SAT-V) tests (*4*). The SAT is designed for able juniors and seniors in high school, who are an average of 4 to 5 years older than the students in the talent searches. The mathematical section is particularly designed to measure mathematical reasoning ability (*5*). For this reason, scores on the SAT-M achieved by 7th and 8th graders provided an excellent opportunity to test the Fennema and Sherman differential course-taking hypothesis (*2*), since until then all students had received essentially identical formal instruction in mathematics (*6*). If their hypothesis is correct, little difference in mathematical aptitude should be seen between able boys and girls in our talent searches.

Results from the six talent searches are shown in Table 1. Most students

0036-8075/80/1212-1262$00.50/0

Table 1. Performance of students in the Study of Mathematically Precocious Youth in each talent search (N = 9927).

Test date	Grade	Number		SAT-V score* ($\bar{X}$ ± S.D.)		SAT-M scores†				Percentage scoring above 600 on SAT-M	
						$\bar{X}$ ± S.D.		Highest score			
		Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
March 1972	7	90	77			460 ± 104	423 ± 75	740	590	7.8	0
	8+	133	96			528 ± 105	458 ± 88	790	600	27.1	0
January 1973	7	135	88	385 ± 71	374 ± 74	495 ± 85	440 ± 66	800	620	8.1	1.1
	8+	286	158	431 ± 89	442 ± 83	551 ± 85	511 ± 63	800	650	22.7	8.2
January 1974	7	372	222			473 ± 85	440 ± 68	760	630	6.5	1.8
	8+	556	369			540 ± 82	503 ± 72	750	700	21.6	7.9
December 1976	7	495	356	370 ± 73	368 ± 70	455 ± 84	421 ± 64	780	610	5.5	0.6
	8‡	12	10	487 ± 129	390 ± 61	598 ± 126	482 ± 83	750	600	58.3	0
January 1978	7 and 8‡	1549	1249	375 ± 80	372 ± 78	448 ± 87	413 ± 71	790	760	5.3	0.8
January 1979	7 and 8‡	2046	1628	370 ± 76	370 ± 77	436 ± 87	404 ± 77	790	760	3.2	0.9

*Mean score for a random sample of high school juniors and seniors was 368 for males and females (*8*). †Mean for juniors and seniors: males, 416; females, 390 (*8*). ‡These rare 8th graders were accelerated at least 1 year in school grade placement.

scored high on both the SAT-M and SAT-V. On the SAT-V, the boys and girls performed about equally well (*7*). The overall performance of 7th grade students on SAT-V was at or above the average of a random sample of high school students, whose mean score is 368 (*8*), or at about the 30th percentile of college-bound 12th graders. The 8th graders, regular and accelerated, scored at about the 50th percentile of college-bound seniors. This was a high level of performance.

A large sex difference in mathematical ability in favor of boys was observed in every talent search. The smallest mean difference in the six talent searches was 32 points in 1979 in favor of boys. The statistically significant t-tests of mean differences ranged from 2.5 to 11.6 (*9*). Thus, on the average, the boys scored about one-half of the females' standard deviation (S.D.) better than did the girls in each talent search, even though all students had been certified initially to be in the top 2nd, 3rd, or 5th percentiles in mathematical reasoning ability (depending on which search was entered).

One might suspect that the SMPY talent search selected for abler boys than girls. In all comparisons except for two (8th graders in 1972 and 1976), however, the girls performed better on SAT-M relative to female college-bound seniors than the boys did on SAT-M relative to male college-bound seniors. Furthermore, in all searches, the girls were equal verbally to the boys. Thus, even though the talent-search girls were at least as able compared to girls in general as the talent-search boys were compared to boys in general, the boys still averaged considerably higher on SAT-M than the girls did.

Moreover, the greatest disparity between the girls and boys is in the upper ranges of mathematical reasoning ability. Differences between the top-scoring boys and girls have been as large as 190 points (1972 8th graders) and as low as 30 points (1978 and 1979). When one looks further at students who scored above 600 on SAT-M, Table 1 shows a great difference in the percentage of boys and girls. To take the extreme (not including the 1976 8th graders), among the 1972 8th graders, 27.1 percent of the boys scored higher than 600, whereas not one of the girls did. Over all talent searches, boys outnumbered girls more than 2 to 1 (1817 boys versus 675 girls) in SAT-M scores over 500. In not one of the six talent searches was the top SAT-M score earned by a girl. It is clear that much of the sex difference on SAT-M can be accounted for by a lack of high-scoring girls.

A few highly mathematically able girls have been found, particularly in the latest two talent searches. The latter talent searches, however, were by far the largest, making it more likely that we could identify females of high mathematical ability. Alternatively, even if highly able girls have felt more confident to enter the mathematics talent search in recent years, our general conclusions would not be altered unless all of the girls with the highest ability had stayed away for more than 5 years. We consider that unlikely. In this context, three-fourths as many girls have participated as boys each year; the relative percentages have not varied over the years.

It is notable that we observed sizable sex differences in mathematical reasoning ability in 7th grade students. Until that grade, boys and girls have presumably had essentially the same amount of formal training in mathematics. This assumption is supported by the fact that in the 1976 talent search no substantial sex differences were found in either participation in special mathematics programs or in mathematical learning processes (*6*). Thus, the sex difference in mathematical reasoning ability we found was observed before girls and boys started to differ significantly in the number and types of mathematics courses taken. It is therefore obvious that differential course-taking in mathematics cannot alone explain the sex difference we observed in mathematical reasoning ability, although other environmental explanations have not been ruled out.

The sex difference in favor of boys found at the time of the talent search was sustained and even increased through the high school years. In a follow-up survey of talent-search participants who had graduated from high school in 1977 (*10*), the 40-point mean difference on SAT-M in favor of boys at the time of that group's talent search had increased to a 50-point mean difference at the time of high school graduation. This subsequent increase is consistent with the hypothesis that differential course-taking can affect mathematical ability (*2*). The increase was rather small, however. Our data also show a sex difference in the number of mathematics courses taken in favor of boys but not a large one. The difference stemmed mainly from the fact that approximately 35 percent fewer girls than boys took calculus in high school (*10*). An equal proportion of girls and boys took mathematics in the 11th grade (83 percent), however, which is actually the last grade completed before taking the SAT in high school. It, therefore, cannot be argued that these boys received substantially more formal practice in mathematics and therefore scored better. Instead, it is more likely that mathematical reasoning ability influences subsequent differential course-taking in mathematics. There were also no significant sex differences in the grades earned in the various mathematics courses (*10*).

A possible criticism of our results is that only selected mathematically able,

highly motivated students were tested. Are the SMPY results indicative of the general population? Lowering qualifications for the talent search did not result in more high-scoring individuals (except in 1972, which was a small and not widely known search), suggesting that the same results in the high range would be observed even if a broader population were tested. In addition, most of the concern about the lack of participation of females in mathematics expressed by Ernest (*11*) and others has been about intellectually able girls, rather than those of average or below average intellectual ability.

To what extent do girls with high mathematical reasoning ability opt out of the SMPY talent searches? More boys than girls (57 percent versus 43 percent) enter the talent search each year. For this to change our conclusions, however, it would be necessary to postulate that the most highly talented girls were the least likely to enter each search. On both empirical and logical grounds this seems improbable.

It is hard to dissect out the influences of societal expectations and attitudes on mathematical reasoning ability. For example, rated liking of mathematics and rated importance of mathematics in future careers had no substantial relationship with SAT-M scores (*6*). Our results suggest that these environmental influences are more significant for achievement in mathematics than for mathematical aptitude.

We favor the hypothesis that sex differences in achievement in and attitude toward mathematics result from superior male mathematical ability, which may in turn be related to greater male ability in spatial tasks (*12*). This male superiority is probably an expression of a combination of both endogenous and exogenous variables. We recognize, however, that our data are consistent with numerous alternative hypotheses. Nonetheless, the hypothesis of differential course-taking was not supported. It also seems likely that putting one's faith in boy-versus-girl socialization processes as the only permissible explanation of the sex difference in mathematics is premature.

CAMILLA PERSSON BENBOW
JULIAN C. STANLEY
Department of Psychology,
Johns Hopkins University,
Baltimore, Maryland 21218

References and Notes

1. E. Fennema, *J. Res. Math. Educ.* **5**, 126 (1974); "National assessment for educational progress," *NAEP Newsl.* **8** (No. 5), insert (1975); L. Fox, in *Intellectual Talent: Research and Development*, D. Keating, Ed. (Johns Hopkins Univ. Press, Baltimore, 1976), p. 183.
2. For example, E. Fennema and J. Sherman, *Am. Educ. Res. J.* **14**, 51 (1977).
3. W. George and C. Solano, in *Intellectual Talent: Research and Development*, D. Keating, Ed. (Johns Hopkins Univ. Press, Baltimore, 1976), p. 55.
4. The SAT-V was not administered in 1972 and 1974, and the Test of Standard Written English was required in 1978 and 1979.
5. W. Angoff, Ed., *The College Board Admissions Testing Program* (College Entrance Examination Board, Princeton, N.J., 1971), p. 15.
6. C. Benbow and J. Stanley, manuscript in preparation.
7. This was not true for the accelerated 8th graders in 1976. The N for the latter comparison is only 22.
8. College Entrance Examination Board, *Guide to the Admissions Testing Service* (Educational Testing Service, Princeton, N.J., 1978), p. 15.
9. The t-tests and P values for 7th and 8th graders, respectively, in the six talent searches were 2.6, $P < .01$; 5.3, $P < .001$; 5.1, $P < .001$; 5.2, $P < .001$; 4.9, $P < .001$; 7.1, $P < .001$; 6.6, $P < .001$; 2.5, $P < .05$; 11.6, $P < .001$; and 11.5, $P < .001$.
10. C. Benbow and J. Stanley, in preparation.
11. J. Ernest, *Am. Math. Mon.* **83**, 595 (1976).
12. I. MacFarlane-Smith, *Spatial Ability* (Univ. of London Press, London, 1964); J. Sherman, *Psychol. Rev.* **74**, 290 (1967).
13. We thank R. Benbow, C. Breaux, and L. Fox for their comments and help in preparing this manuscript. Supported in part by grants from the Spencer Foundation and the Educational Foundation of America.

21 March 1980; revised 14 August 1980

Human Sleep: Its Duration and Organization Depend on Its Circadian Phase

Abstract. *Two- to threefold variations in sleep length were observed in 12 subjects living on self-selected schedules in an environment free of time cues. The duration of polygraphically recorded sleep episodes was highly correlated with the circadian phase of the body temperature rhythm at bedtime and not with the length of prior wakefulness. Furthermore, the rate of REM (rapid eye movement) sleep accumulation, REM latency, bedtime selection, and self-rated alertness assessments were also correlated with the body temperature rhythm.*

Forty years ago, Kleitman wrote that "the time between going to bed at night and getting up in the morning is one of the easiest characteristics of sleep to study" (*1*). Despite this ease of measurement, the processes that control the length of "ad libitum sleep" (that is, sleep not truncated by an alarm clock or disturbance) have remained undefined. A number of studies have contradicted the intuitive assumption that the length of sleep is determined by the length of prior wakefulness. "Recovery" sleep after 3 to 10 days of total sleep deprivation rarely exceeds 11 to 16 hours (*2*), while both longer (15 to 20 hours) and shorter (6 to 10 hours) sleep episodes have been observed in subjects not deprived of sleep who lived on a self-scheduled routine (*3-6*). In fact, the wide variation in sleep duration reported in such "free-running" subjects has been characterized as random and irregular. We now report that such variations in sleep duration occur in a consistent and predictable manner which depends on when subjects go to sleep, rather than how long they have been awake beforehand.

We polygraphically recorded the sleep of 12 male subjects (21 to 53 years old), each living separately for 16 to 189 days (total of 562 days) on a self-scheduled routine in an environment free of time cues (*3, 4*). These subjects developed free-running, non-24-hour sleep-wake, body temperature, and neuroendocrine cycles. In one group of subjects, all those free-running rhythms remained internally synchronized with nearly identical periods, although their waveshapes and phase relationships were different from those during entrainment to a 24-hour day. For example, the decrease in body temperature that has long been associated with the daily sleep episode (*1*) began several hours before sleep in those free-running subjects, reaching its nadir near sleep onset and then rising throughout the rest of the sleep episode (*3-5*).

However, six of our subjects had a number of sleep-wake cycles of extraordinary duration—up to 50 hours in length—with a persisting near-24-hour rhythm in body temperature. This state, which several others have observed (*5, 6*), has been termed "internal desynchronization" by Aschoff and Wever, a concept emphasizing the uncoupling of rhythms that are normally linked in close temporal order.

Examination of the bedrest-activity pattern, when plotted in a raster format (*7*), led us to recognize—even in such "desynchronized" subjects—regularly recurring clusters of short (6 to 10 hours) sleep episodes, interrupted by long episodes of sleep which were not "in phase" with those clusters (triple plotted example from subject CA shown in Fig. 1 on experimental days 35 to 83). The visible line along which those clusters recurred had a consistent phase relationship to the ongoing near-24-hour cycle of body temperature (stippled area in Fig. 1); that is, the short sleep episodes usually began just at or after the mid-trough of the temperature cycle. This phase relationship was very similar to that already

0036-8075/80/1212-1264$00.50/0