INTELLIGENCE 7, 129–152 (1983)

Structure of Intelligence in Intellectually Precocious Children and in Their Parents*

CAMILLA PERSSON BENBOW
Study of Mathematically Precocious Youth (SMPY)
The Johns Hopkins University

JULIAN C. STANLEY
Study of Mathematically Precocious Youth (SMPY)
The Johns Hopkins University

MARSHALL K. KIRK
Harvard University

ALAN B. ZONDERMAN
The Johns Hopkins University

Students representing the top 0.03% of their age group in intellectual ability, who were identified by the Study of Mathematically Precocious Youth (Benbow & Stanley, 1980), were tested along with their parents using a battery of specifically designed cognitive tests. These highly intelligent children had less intelligent, but yet quite bright parents. Vernon's (1961) model of intelligence best fits our results. His following two factors explained most of the variance in the performance of the students and parents: verbal-educational and practical-spatial-mechanical. Moreover, there was potential evidence for a general factor. Among the children, who were mostly past puberty, age related to development of verbal abilities, but not spatial or mechanical abilities. Sex differences favoring the males were found on the spatial ability and mechanical comprehension tests.

From 1972 to 1979, the Study of Mathematically Precocious Youth (SMPY) identified approximately 10,000 junior high school students who reasoned quite well mathematically (Benbow & Stanley, 1980). As seventh and eighth graders, several of those students scored *exceptionally* high on the College Board Scho-

*We thank Lynn M. Daggett, Lloyd G. Humphreys, Lola L. Minor, and Robert Plomin for their assistance. Correspondence and requests for reprints should be addressed to Camilla Persson Benbow, Department of Psychology, The Johns Hopkins University, Baltimore, MD 21218

lastic Aptitude Test (SAT); mathematics (SAT-M) and verbal (SAT-V) portions (Angoff, 1971; Messick & Jungeblut, 1981). SMPY is interested in discovering the bases and associated characteristics of their extreme intellectual precocity.

Thus, in this study we tried to decipher how overall extreme intellectual precocity relates to aptitude on difficult tests of certain specific mental abilities. Furthermore, we tried to learn which factor analytic model of intelligence could best account for their performance on these various tests? Is there strong evidence for a *g* or for specific factors? How do age and education relate to the development of various specific abilities? Finally, among the extremely precocious, are there sex differences in some specific abilities? Moreover, some of the students' parents were also tested with the same tests as the students. Their performance is discussed and compared.

Aspects of assortative marriage and familiality of cognitive abilities in families of these extremely gifted students are discussed by Benbow, Zonderman, and Stanley (this issue). That study shows that the parents resembled one another to a higher degree than less able spouses did and that the children resembled their parents to a lesser extent than less able children resembled their parents.

Because the students in this investigation were so bright, some mental tests had to be designed specifically in order to provide an appropriate measure of these students' abilities. This circumvented the problem of "ceiling" that most tests have when used with this type of population. As far as we are aware, a battery of tests of this kind for such a population has not yet been developed. Some piloting of the battery had already been done (Kirk, 1978b, 1979, 1980); this is the first fairly large-scale investigation validating those tests.

METHOD

Subjects

The student participants in this study were selected on the basis of their high scores on SAT-M and, when available, SAT-V, made at the time of their participation in a talent search.[2] These talent-search participants ranged in age, at that time, from 9 to 14. To control for the effects of age on performance, SAT-scaled scores were converted to age-adjusted scores. Stanford-Binet mental age equivalents were estimated for each SAT-M or SAT-V + SAT-M score from the 25th to the 99th percentile of a random sample of high school juniors taking the SAT, whose mean Stanford-Binet IQ is approximately 106 (Kirk, 1978a, 1980). For

[1]The experimental tests are available upon request from Marshall K. Kirk, Department of Psychology, Harvard University, Cambridge, MA 02138.

[2]The six talent searches were conducted in March 1972, January–February 1973, January 1974, December 1976, January 1978, and January 1979. The SAT-M and SAT-V portions were administered every year except 1972 and 1974, when only the SAT-M was administered. The Test of Standard Written English was also taken by the students participating in 1978 and 1979.

example, a V + M score of 900 was equated to a Stanford-Binet mental age of 18.25 years because both values define the 75th percentile of high school juniors (Kirk, 1980). The estimated mental age for each talent-search participant was divided by his or her chronological age at testing, then multiplied by 100 to obtain a "Precocity Quotient," quantitatively and conceptually similar to the Stanford-Binet IQ. "Precocity Quotients" for approximately 10,000 talent-search participants obtained by computer search ranged from the 120s to 235. Students with "Precocity Quotients" of 170 or more (estimated frequency less than 1 in 3,000 of their age group) were asked to participate in the present study.[3] One hundred boys and 14 girls, out of a 43% female pool qualified. This is approximately the upper 1% of the talent-search participants, who themselves were approximately the upper 3% in ability. They thus represent the top .03% of their age group in intellectual ability.

All of these students and their parents were invited to participate. Three testing sessions were conducted in the Summer and Fall of 1979 (a given examinee came to only one of these). Of the students invited, 12 girls (86%) and 60 boys (60%) attended. Of the parents, 46 mothers (40%) and 45 fathers (39%) attended. Thirty-five complete familites (i.e., father, mother, and child) attended the testing sessions.

The average ages of the boys and girls in the Summer and Fall of 1979 were 15.1 and 13.5 years, respectively. They ranged from 10.5 to 19.8 years for the boys, and 12.3 to 18.5 years for the girls. The girls as a group were younger than the boys because few girls from the early talent searches qualified for this study. In the early years only the mathematics part of the SAT was administered, and very few girls scored high on it (Benbow & Stanley, 1980, 1981). As a result, girls bright enough to be part of this study were identified in the later talent searches when SAT-V scores were available. Therefore, they were young at the time of testing for this study. With regard to years of education, the boys had completed an average of 10 school grades beyond kindergarten and the girls 8.

For the parents, the average ages at the time of testing were 45.3 for the fathers and 42.5 for the mothers. They ranged from 35 to 68 years for the fathers and 32 to 62 for the mothers. Their mean number of years of education were 18 and 16, respectively. For both the mothers and the fathers, the range in education was from completing grade 12 (high school diploma) to 20 years (doctorate). Thus, this was a younger and (as yet) somewhat less well educated sample of female than male parents.

Instruments

Descriptions of the nine tests administered during an approximately five-hour-long testing session (not including breaks) are provided in the next section. The first four are the standardized tests, and the last five are the experimental tests.

[3]One exception was made for an individual with a "precocity quotient" of 161.

The experimental tests tended to be somewhat speeded for this group. Therefore, we could not calculate the internal consistency reliabilities of these tests. A *lower* bound estimate of the reliabilities could be obtained, however.

Lower bound estimates of the reliability coefficients for the experimental tests were estimated by using the well-known formula (Stanley, 1971, p. 400, Formula 38) for correcting coefficients of correlation for attenuation due to errors of measurement. For the population of examinees the formula is:

$$\rho_{T_xT_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\quad\sqrt{\rho_{yy'}}},$$

where $\rho_{T_xT_y}$ is the correlation coefficient between true scores on the two variables, ρ_{xy} is the correlation coefficient between obtained scores on the standardized test X and the experimental test Y, $\sqrt{\rho_{xx'}}$ is the reliability coefficient of the standardized test, and $\sqrt{\rho_{yy'}}$ is the reliability coefficient of the experimental test.

$$\rho_{T_xT_y} \leqslant 1,$$

the coefficient of correlation between true scores on parallel forms of the same test. The equation becomes

$$1 \geqslant \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}\quad\sqrt{\rho_{yy'}}}.$$

Thus, multiplying both sides by $\sqrt{\rho_{yy}}$, one obtains

$$\sqrt{\rho_{yy'}} \geqslant \frac{\rho_{xy}}{\sqrt{\rho_{xx'}}}.$$

The lower bound estimate was figured for the actual r_{xy} for every correlation between an experimental and a standardized test. The highest estimated reliability coefficient for each experimental test resulting from these computations was accepted as the lower-bound; of course, this capitalizes somewhat on chance.

For the spatial ability tests we could not get an accurate lower bound estimate of the reliability coefficients. Because a spatial ability measure with a known reliability coefficient was not administered, the correlation coefficients in the numerator of our formula for the lower bound estimate of the reliability coefficient of the experimental spatial ability test is the *r* between two tests of different abilities (i.e., spatial ability and ability on QED). The *r*s between the experimental spatial tests and the QED are undoubtedly considerably lower than would be obtained between experimental and standardized spatial ability measures.

The reliabilities of the standardized tests for the samples in this study were estimated by use of the following formula (see Stanley, 1971, p. 362):

$$\text{reliability} = 1 - \frac{s_e^2}{s_x^2}$$

where s_e^2 = variance of errors of measurement as inferred from the test manual,

and s_x^2 = obtained variance of group on test.

The Tests

The *Quantitative Evaluative Device (QED),* Form D, was designed to predict competence in the quantitative aspects of graduate research and study (Stake, 1962). Each of the 60 items has five choices, four of which share a common feature. The object is to select, by inductive reasoning, the one that does not share in the similarity. The score is corrected for guessing and is the number correct minus one-fourth the number incorrect ($R - \frac{1}{4}W$). The norm group (graduate education majors) yielded a matched-half reliability coefficient of .80. The best estimates for the reliabilities of the participants in our testing are .84 for the boys, .88 for the girls and the mothers, and .92 for the fathers.

Form CC of the Owens and Bennett *Mechanical Comprehension Test* was designed "to measure the ability of an individual to understand various kinds of physical and mechanical relationships" (Owens & Bennett, 1949). This untimed test consisting of 60 problems, with five options each, is perhaps the most difficult of its type. Each item is illustrated by a large, clear drawing on which the question and answer are based. Since the authors suggest that the test not be given to women, except when the individual would be entering a field in direct competition with men, the norms are based on men only. The reliability corrected by the Spearman-Brown formula was, long ago, .80 for a combined group of male freshmen engineering majors and .75 for first-term engineering majors at Princeton. For the participants in our testing, the best estimates of reliability are .86 for the boys, .77 for the girls, and .91 for the fathers and the mothers.

The *Concept Mastery Test,* Form T was designed by the late Lewis M. Terman to test his gifted group[4] as adults. The test measures the ability to deal with abstract ideas at a high level.[5] This untimed test consists of two parts: Part I (Synonyms and Antonyms) with 115 two-option items requiring the identification of pairs of words as the same or opposite in meaning, and Part II (Analogies) with 75 three-option items requiring the recognition of relationships in order to complete the analogies. Scoring for both parts is corrected for guessing: Part I, $R - W$; Part II, $R - \frac{1}{2}W$. Test-retest correlations reported in the test manual vary

[4]As children, most of the 1,528 members of this group scored at least 140 on the original (i.e., 1916) version of the Stanford-Binet Intelligence Scale (see Oden, 1968).

[5]This test is viewed by some as measuring chiefly vocabulary (Part I) and general information (Part II).

from .86 (Air Force Captains) to .94 (undergraduate and graduate students and teaching assistants at Stanford University and the University of California). The Terman group's reliability coefficient was .87. Reliability estimates for the testing participants are .90 for the boys, .94 for the girls, and .93 for the fathers and mothers.

The three language subtests of the *California Test of Mental Maturity* (CTMM), Advanced Level, were designed to be used with adolescents and adult populations to obtain a measure of mental maturity. This test is not a specific ability test. It was used in order to study overall intellectual functioning of the children and their parents. The score used is the total correct for tests 14, 15, and 16. Test 14 consists of 15 arithmetic reasoning items in word-problem form; Test 15 consists of 15 items requiring logical reasoning; and Test 16 has 50 vocabulary definition items. The reliabilities on the complete CTMM are reported as .90 and above. If we can assume that these three subtests have about the same reliability as the complete test, then the reliability estimates for the testing participants are about .92 for the girls and about .91 for the boys, fathers, and mothers. The reliability of the shorter version is probably smaller than the complete test, but not so different as to change these numbers radically.

Kirk's *Synonym-Antonym Test,* Form Cel, is a measure of the individual's knowledge of high-level vocabulary gained more from general reading (personal and educational) than from specific technical literature. The test consists of 100 two-option items similar to Part I of the Concept Mastery Test (see Figure 1), but more difficult.

Although the test is untimed, the directions state that 10 minutes should be sufficient for completion. Scoring is corrected for chance and is the number correct minus the number wrong (*R*—*W*). The internal consistency reliability of a previous version (Form Be) was .93, based on Harvard-Radcliffe undergraduates and graduates (Kirk, 1978b, 1979). Our best estimated lower bound for the reliability coefficient of Form Cel is .79.

Kirk's *General Information Test,* Form Cel, is a measure of an individual's knowledge of facts, concepts, and terminology considered academically important. The 120 three-option items are divided equally among three content areas: mathematics and hard sciences, social sciences, and humanities (see Figure 1). Since no course of study could encompass the whole, scores are influenced more by general reading than by formal education. A bias against older adults, despite their lengthier education is probable since the test content contains relatively recent references (Kirk, personal communication). The test is untimed with a suggested completion time of 30 minutes. The score is based on the number correct minus one-half the number incorrect (*R*—½*W*). The reliability of an earlier timed version (Form Be) was reported as .86 (Kirk, 1978b, 1979). Our estimate for Form Cel is a reliability coefficient of at least .71.

Kirk's *Test of Semantic Comprehension,* Form Cel, consists of Part I, 20 five-option sentence completion items, and Part II, 30 six-option proverb interpretation items (see Figure 1). The test was designed to measure an individual's ability to understand conceptually complex passages of written English in a high-school level vocabulary (Kirk, 1978b, 1979). Part I, with a 10-minute time limit, is scored on the number correct minus one-fourth the number incorrect (R—¼W). Part II, also with a 10-minute time limit, is scored on the number correct minus one-fifth fthe number incorrect (R—1/5W). The total score is the sum of the scores for both Parts I and II. The reliability estimate is .66. We could perform an independent check on our lower bound estimate of the reliability coefficient for this test. Using the correlation coefficient between the two subtests, we could estimate the reliability of the 20-item subtest. Then using the Spearman-Brown step-up formula, we estimated the reliability of the 50-item complete test. This came out to be .65, which agrees almost exactly with the first lower bound estimate. The correlation of the students' scores between Part I and Part II is .41, significant beyond the .001 level.

Kirk's *Cubes Test,* Form Be, is a spatial relationship scale that measures an individual's ability to form and manipulate mental images of objects (Kirk, 1978b, 1979). Each of the 30 two-option items shows three faces of each of the two members of a pair of cubes. A cube has a different figure on each of its six faces. The object is to determine whether or not the second cube could be any rotation of the first (see Figure 2).

There is a 5-minute time limit, and the score is the number of items marked correctly minus the number marked incorrectly (R—W). Our best minimum estimate is a reliability coefficient of at least .53, with Kirk estimating the reliability for his examinees at about .80 (Kirk, personal communication). Our estimate is almost without a doubt considerably too small, since a standardized spatial ability measure was not administered.

Kirk's *Rotation-Inversion Test,* Form Cel, is a test of spatial ability more complex than the *Cubes Test*; it also measures the ability to manipulate figures mentally (Kirk, 1978b, 1979). Each item consists of a lozenge with a heavy black line on one side and a figure in one corner. Answering the item correctly requires mentally inverting the lozenge and then rotating it until the heavy black line rests on the line provided (see Figure 2). The test consists of four separately timed pages (2 minutes each) containing 16 items each for a total of 64 items. The score is the number correct. Our best reliability estimate is at least .65, with Kirk estimating the reliability for his examinees at about .80 (Kirk, personal communication). As with the *Cubes Test,* our estimate of the reliability is probably too small, since a standardized spatial ability measure was not administered.

Data Analysis

The nine tests described above were hand-scored, and the scoring was verified by a different scorer. Discrepancies in marking were checked by a third scorer. The statistical analyses were performed by computer, using the SPSS program

TEST OF SEMANTIC COMPREHENSION AND MANIPULATION

Part I: Sentence Completion

Example: I. He was not (   ) by his stupidity; indeed, he was much too (   ) to know that he was stupid.

(a) bothered-stupid
(b) made wealthy-poor
(c) delighted-exhausted
(d) enlightened-intelligent
(e) made famous-obscure

In the example above, only choice (a) renders the sentence meaningful, coherent, and consistent. The other choices either do not "follow" from the rest of the sentence, or make the sentence contradictory.

Part II: Proverbs

Example: I. "Strike while the iron is hot."

(a) Opportunity knocks but once.
(b) Fate favors those who seize the moment.
(c) The shoemaker's children go barefoot.
(d) A blow struck in anger is repented at leisure.

In the example above, (a) and (b) clearly mean that you should take advantage of an opportunity when the time is ripe. Choice (d) means the opposite. Choice (c) has nothing to do with the other proverbs. You would mark *a* and *b* as your answers.

TEST OF GENERAL INFORMATION

MATHEMATICS AND HARD SCIENCES ITEM

1—Animal life includes ________ alpha-amino and -imino acids, of which ________ are 'essential' to Man.

(a) 18–8 (b) 22–8 (c) 18–4

HUMANITIES ITEM

9—In *Animal Farm*, Snowball corresponds symbolically to

(a) Lenin (b) Trotsky (c) Stalin

SOCIAL SCIENCES ITEM

11—North America during the period 1700–1800 was most politically like ________

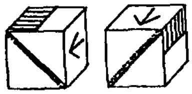
(a) Africa, 1950–65 (b) China, 1850–65 (c) India, 1948

SYNONYM-ANTONYM TEST

36 antipodean..propinquitous
37 malodorous ....nocent
38 demotic...ultramontane
39 acuminated ....truncated
40 cavalier...obsequious

FIG. 1. Sample items from the verbal experimental tests

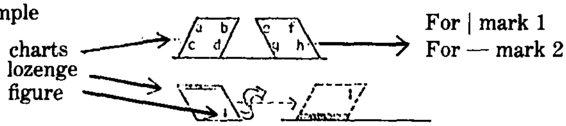
CUBES



For each pair of Cubes, you are to determine whether the two Cubes could possibly be rotations of the same Cube.

ROTATION-INVERSION

Here is an example



The little line will be in position *b*, and it will be *vertical*. So you should write *b1* as your answer.

FIG. 2. Sample items from the Cubes and Rotation-Inversion Tests Illustrating the Tests

(Nie, Hull, Jenkins, Skinbrenner, & Bent, 1975). Since age had a relationship with performance on the tests, as will be discussed later, we controlled for it statistically when analyzing relationships among the tests.

Results

Descriptive statistics from the testing sessions are shown by sex for the students and parents in Table 1. The mean "Precocity Quotients" were 186 for both boys and girls in the study. The analogous Stanford-Binet IQ would represent a frequency of less than 1 in 25,000 (Kirk, 1980). Clearly this group of students was highly precocious.

Quantitative Evaluative Device

On the QED both students and parents attained high scores (see Table 1). The children and parents scored at approximately the 97th and 86th percentiles, respectively, of potential University of Nebraska graduate students in education. Thus, both parents and students were well equipped to handle the quantitative aspects of research, even though some were only 12-years old at the time and one was 10.

QED appears to have a substantial verbal component for this group, as can be judged from its pattern of relationship with the other tests administered (Table 2). For the parents, QED ability correlated highly with all the other mental abilities

TABLE 1

Mean Scores by Sex on the Nine Cognitive Tests Administered to the Exceptionally Gifted Students and Their Parents and the Mean Precocity Quotients of the Students

	Precocity Quotient		Q.E.D.		Mechanical Comprehension		CMT		CTMM-Language Mental age years		Synonym-Antonym	
	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*	$\bar{X}$	*SD*
Boys (60)	186	13	40	8	38	9	97	30	23.5	2.1	18	15
Girls (12)	186	18	37	9	28	7	98	40	22.7	3.0	20	17
t of					$t = 4.32$							
difference	ns		ns		$p < .001$		ns		ns		ns	
Fathers (43)	—	—	33	11	40	11	118	38	24.1	2.6	35	19
Mothers (45)	—	—	33	9	26	11	120	38	23.5	2.6	35	17
t of					$t = 5.5$							
difference			ns		$p < .001$		ns		ns		ns	

	General Information		Semantic Comprehension		Cubes		Rotation Inversion	
Boys (60)	32	13	17	6	21	8	28	15
Girls (12)	27	13	17	6	16	6	24	12
t of					$t = 2.2$		ns	
difference	ns		ns		$p < .05$			
Fathers (43)	25	15	15	6	15	7	18	15
Mothers (4)	23	10	14	6	11	7	12	11
t of					$t = 2.1$		$t = 2.3$	
difference	ns		ns		$p < .05$		$p < .05$	

TABLE 2
Partial Correlations of the Performances on the Various Mental Ability Tests for Children (N=72), Fathers (N=43), and Mothers (N=45), Controlling for Age

	QED	Mechanical Comprehension	CMT	CTMM Language MA	Synonym-Antonym	General Information	Semantic Comprehension	Cubes	Rotation-Inversion
Precocity Quotient Children	.40***	.23*	.42***	.55***	.32**	.36***	.40***	.21*	.23*
QED									
Children		.19	.30**	.39***	.34***	.34**	.31**	.40**	.18
Fathers	----	.71***	.65***	.62***	.43**	.64***	.57***	.58***	.71***
Mothers		.67***	.64***	.79***	.56***	.52***	.47***	.44**	.54***
Mechanical Comprehension									
Children			.04	.26*	-.03	.35**	.21*	.33**	.47***
Fathers	----	----	.48***	.53***	.30**	.67***	.50***	.53***	.69***
Mothers			.46**	.66***	.40**	.67***	.51***	.49***	.64***
CMT									
Children				.81***	.71***	.64***	.56***	.10	.07
Fathers	----	----	----	.86***	.78***	.71***	.72***	.35*	.46**
Mothers				.80***	.83***	.63***	.70***	.19	.40**
CTMM Language MA									
Children					.61***	.74***	.60***	.20*	.24*
Fathers	----	----	----	----	.67***	.71***	.69***	.40**	.41*
Mothers					.70***	.50***	.66***	.37**	.58***

Synonym-Antonym									
Children						.51***	.54***	.03	.05
Fathers	----	----	----	----	----	.70***	.52***	.08	.28*
Mothers						.57***	.70***	.22	.37**
General Information									
Children							.49***	.28**	.31**
Fathers	----	----	----	----	----	----	.62***	.36*	.48***
Mothers							.47***	.33*	.49***
Semantic Comprehension									
Children								.07	.22*
Fathers	----	----	----	----	----	----	----	.51***	.46**
Mothers								.27*	.61***
Semantic Comprehension									
Children									.55***
Fathers	----	----	----	----	----	----	----	----	.68***
Mothers									.60***

*$p < .05$
** $p < .01$
***$p < .001$

tested (range of the 16 partial *rs*, .44 –.80). For the children, these relationships were much less strong (.18 –.40), especially with ability on the mechanical comprehension test (.19) and the Rotation-Inversion test (.18). Particularly for the parents, the QED may serve as a reasoning test.

Mechanical-Comprehension Test

Males performed more than a standard deviation better than females (see Table 1), but within each sex the children did not perform significantly better or worse than the parents. Males performed at the 32nd percentile and females at about the 4th percentile of first-term Princeton University freshmen in engineering. The norms, however, are old and probably demanding. Thus, these scores indicate relatively good performance for the males. For the females, the test was somewhat too difficult.

Concept Mastery Test

On the CMT (see Table 1) the students scored at about the 11th percentile and the parents at about the 23rd percentile of the Terman gifted group. The Terman gifted group is also an elite group against which to compare performance. As children, most of the 1,528 members of Terman's group had scored at least 140 on the Stanford-Binet Intelligence scale. The norms for CMT were based on the testing results for the group when its members were, on the average, 41 years old. The Terman cohort had been tested 10 years earlier, and their scores rose considerably during the 10-year time period. Taking this into account, and the fact that a rather select group of Air Force captains scored, on the average, only 60 points out of the possible 190, the performance of this group, especially that of the students, was extremely high (see Bayley & Oden, 1955). The students scored almost 40 points more than the Air Force captains, and the parents almost 60 points more. When these students become as old as their parents were when tested, undoubtedly they will score higher than them on CMT.

CTMM-Language Factor

The point scores on the sum of these three tests were translated into mental ages.[6] It is clear that verbally both the parents and students are extremely able (see Table 1). Since the students received approximately the same mental age score as the parents, they are abler relative to their age than the parents, as would be expected from regression to the mean (i.e., the students, but not their parents, were selected on the basis of ability). The boys, girls, fathers, and mothers averaged, respectively, the following verbal IQs: 164 (SD = 16), 171 (SD = 15),

[6]It should be noted that, when comparing mental ages, a usual convention is that all individuals 16 years or older are assigned a chronological age of 15.

151 (SD = 17), and 148 (SD = 17). This test was probably too easy for the group, especially for some of the parents. We do not know how highly several of the parents could have scored on a more difficult test of this type.

Synonym-Antonym Test

Out of 100 possible questions corrected for chance, the children scored, on the average, 18 and the parents 35. Approximately two-thirds of the children's scores fell in the range 4 through 32. For the parents, this range was 12 through 54. For the examinees as a group, this test was too difficult at this time.

The strong partial *r*s, controlling for age, of Synonym-Antonym Test scores with performance on the CMT and CTMM Language (.71—.84) indicate that this test measures verbal ability (see Table 2). Scores on the Synonym-Antonym Test also correlated rather highly with those on the other experimental verbal tests.

General Information Test

The mean corrected-for-chance score on this test, out of a possible 120, was 31 for the students and 24 for the parents. The range in which two-thirds of the scores fell was 20 to 44 for the students and 13 to 36 for the parents. The higher student scores support Kirk's claim (Kirk, personal communication) that this test is probably biased against older adults. Also, it is too difficult for the group as a whole.

From the pattern of intercorrelations in Table 2, it does *not* appear that this test measures only verbal ability or information obtained through verbal means. Mechanical and spatial ability related substantially to performance on the General Information Test, but less so than the verbal tests did. This provides some evidence for the validity of the test as a measure of knowledge of information in 35 different areas.

Semantic Comprehension Test

Out of a possible corrected-for-chance score of 50, the students' mean was 17 and the parents' 14. Thus, the children were somewhat abler verbally than the parents ($p < .001$), which was also found with the CTMM and the General Information Test, but not the CMT.

The Semantic Comprehension Test correlated rather strongly (.56–.76) with the CMT and the CTMM—Language factor (Table 2), indicating that it does measure verbal ability.

Cubes Test

On this 5-minute 30-item spatial ability test there appeared a significant sex difference in favor of the males, even for this high ability group, for both children and parents (see Table 1), as is commonly found in the general population (see

Burnett, Lane & Dratt, 1979; Wittig & Petersen, 1979). Furthermore, the boys scored considerably better than their fathers, and the girls better than their mothers on this appropriately difficult spatial ability test.

Unfortunately, standardized spatial ability tests were not administered to be used as markers to see what aspects of spatial ability the *Cubes Test* does indeed measure. It does, however, correlate rather highly (.56 — .68) with the other experimental spatial ability test, *Rotation-Inversion*. The Cubes Test has face validity for measuring rotational ability in three-dimensional space (see Figure 2).

Rotation-Inversion Test

This 64-item spatial ability test was far too difficult for the parents, but only slightly too difficult for the typical student in this study (Table 1). Again, a potential sex difference emerged in favor of males, but was statistically significant only for the parents.

Since standardized spatial ability tests were not administered as markers for spatial ability, we cannot be sure that the test performs its function. The test has face validity (see Figure 2). It may have measured, in part, especially for the parents, the examinees' ability to understand the instructions. The somewhat stronger verbal component in this test than in the *Cubes Test* would seem to support this contention (see Table 2).[7] Better instructions than those that were supplied with the test were needed.

Sex Differences

The most striking sex difference was that so few females qualified to participate in this study (i.e., 14 girls vs. 100 boys). This occurred because girls in SMPY's talent searches tended to score lower than boys on the test of mathematical reasoning ability (Benbow & Stanley, 1980, 1981), especially in the earlier of the six searches. As in the Terman studies (Burks, Jensen, & Terman, 1947; Terman, 1925), we found many more highly able boys than girls.

Although no sex differences were seen in the precocity quotients and performance on the verbal tests, a distinct sex difference favoring the males emerged on the spatial ability tests and the test of mechanical comprehension. Actually, the girls may have been slightly abler verbally than the boys. Because of sample size, these results are only suggestive of possible sex differences on difficult spatial and mechanical comprehension tests among extremely able students.

---

[7]Because the parents were tested separately from their children and, therefore, necessarily by different examiners, one cannot rule out the possibility that one examiner gave more effective instructions for taking some of these unusual tests than the other did. This would probably affect scores on the Cubes Test and the Rotation-Inversion Test more than the others.

Age and Education

The relationship of age and education to performance on the mental tests is shown in Table 3. It is seen that, for the parents, age was negatively correlated (usually significantly) with performance on all the tests. Education, however, was (usually significantly) positively correlated with performance for the mothers and fathers on all tests. For the children, age and education were correlated usually positively with performance (median .49), except with the Cubes Test.

TABLE 3
Zero-order *r*s of the Childrens' ($N=72$), Fathers' ($N=43$), and Mothers' ($N=45$) Performance on the Cognitive Tests with Their Respective Age and Education

		Age	Education (in years)
QED	Children	.51***	.58***
	Fathers	–.50***	.32**
	Mothers	–.41**	.49***
Mechanical Comprehension	Children	.32**	.31**
	Fathers	–.07	.21
	Mothers	–.21	.49***
CMT	Children	.56***	.61***
	Fathers	–.19	.30**
	Mothers	–.27*	.60***
CTMM Language MA	Children	.51***	.58***
	Fathers	–.39**	.40**
	Mothers	–.30**	.54***
Synonym-Antonym	Children	.52***	.46***
	Fathers	–.34*	.41**
	Mothers	–.09	.58***
General Information	Children	.44***	.45***
	Fathers	–.25	.36*
	Mothers	–.44***	.40**
Semantic Comprehension	Children	.51***	.51***
	Fathers	–.29*	.32**
	Mothers	–.42**	.51***
Cubes	Children	–.09	–.03
	Fathers	–.35**	.22
	Mothers	–.31*	.29*
Rotation-Inversion	Children	.36***	.33**
	Fathers	–.29**	.38*
	Mothers	–.35**	.24

*$p \le .05$
*$p \le .01$
***$p \le .001$

Because of the negative correlation of age with ability for the parents, parents were grouped by age and then their performance was compared. There was a clear trend for the older parents to do more poorly on the tests. Thus, it was not some parents extremely outside the range in age who performed poorly and caused the negative correlations with age.

The relationship of age to test performance is different for these adults than for these children. For the latter, age and education were highly correlated (r = .86, $p < .001$), while negatively correlated for the mothers ($r = -.18$) and the fathers ($r = -.41$, $p < .01$) Thus, understandably, for the students the age and education variables are measuring virtually the same thing, but not for adults. This occurs partly because many of these youths (age range 10 to 20 years) are still progressing educationally one school grade per year, whereas their parents are aging without completing another grade.

There is a large literature on correlations of abilities with age (e.g., Eisdorfer & Wilkie, 1973; Jarvik, Eisdorfer, & Blum, 1973; Matarazzo, 1972; Owens, 1966; Schaie & Strother, 1968; Wechsler, 1944). In longitudinal research there is little or no evidence of decreasing ability with age. Cross-sectional studies, however, found results similar to ours. Generational differences in nutrition, education, and other opportunities can account for most of the differences found between age cohorts. That high verbal ability does not diminish with age was clearly demonstrated in the Terman group (Oden, 1968).

Our results suggest that during the adolescence of brilliant youths, verbal ability improves with age. Development of spatial and mechanical comprehension abilities may be mostly complete by the time adolescence is reached.

Intercorrelation of the Performance on the Exams

The students', mothers', and fathers' performance on the exams were intercorrelated within themselves, with age partialled out (Table 2). Because of the small number of girls (12), their correlation coefficients are not presented separately from the boys. It appeared that the parents' high performance on one test meant an overall high level of performance on the other cognitive tests, which was also true, but to a lesser degree, for the students. This suggests that the general factor, g, of Spearman (1904) was operating, especially for the parents. Interestingly, the separate-by-sex partial correlations were indicative of a stronger influence of g among the girls than among the boys.

The above conclusions were supported by component analyses performed separately for the students, fathers, and mothers on age-adjusted scores on the specific mental ability tests.[8] In Table 4 the factor matrix, using principal components having eigenvalues greater than one, and the communalities are

[8]The CTMM test is more similar to a general intelligence test than to a specific mental ability test. Thus, it was excluded from this analysis.

repesented. For all three groups (children, mothers, and fathers) two components with eigenvalues greater than one emerged. For the children, they accounted for 63% of the variance in age-adjusted scores, for the fathers 76%, and for the mothers 73% (see Table 4).

The communality of each variable indicates the extent to which its variance can be accounted for on the basis of the common factors. Clearly for the children, some of the variables, especially the QED and the mechanical comprehension tests, do measure a great amount that is unique. For both the fathers and mothers, however, the communalities indicated that the tests appeared to be measuring something in common to all the tests, probably general intelligence.

The principal components were then rotated by the Kaiser normalization procedure. The resulting factor matrix is shown in Table 5. The first factor for the children, which loaded on everything but the three spatial-mechanical tests, seems similar to Vernon's (1961) verbal-educational factor (V:ed). Their second factor, which loaded most highly on the spatial-mechanical tests, seems similar to Vernon's practical-spatial-mechanical factor (K:m). The correlation between factors was .25.

The rotated factor structure for the fathers and mothers was similar to, but not the same as, that obtained for the children (see Table 5). The first factor appears to be the verbal-educational factor, since it loads most highly on the verbal tests. The second factor may also be the practical-spatial-mechanical factor of Vernon, since it loads most highly on the spatial and mechanical tests and the QED. The two factors were correlated (approximately .40) which provides evidence that Spearman's *g* may be influencing the performance of the parents.

As a cautionary note, the presence of sex differences in means on the spatial-mechanical tests attenuates the correlations between the spatial-mechanical tests and the verbal tests. This increases the separation between the two factors for the children as compared to the separate factor analyses for fathers and mothers by sex. The separate-by-sex correlations for the students, however, indicate that perhaps the boys' factors would be less correlated than the girls' would be. Unfortunately, our sample size is too small to test this hypothesis.

DISCUSSION

This study was conducted to further test the most brilliant students the Study of Mathematically Precocious Youth (SMPY) had identified over a 7-year period. Some of their parents were also tested. Difficult, high-level tests of several specific mental abilities were used. Because of the extreme precocity of the group, some of the tests had to be designed specifically for use with such subjects.

As expected, it was found that selecting for highly able students also selects for parents who are highly able, though somewhat less so than their brilliant offspring. Regression towards the mean predicts this. Previous studies have shown that selection of high-ability parents yields children who score closer to the mean

TABLE 4
First Two Principal Components for Age-Adjusted Scores on Eight Mental Ability Tests and Their Communalities Separately for Children (N=72), Fathers (N=43), and Mothers (N=45)

	Children			Fathers			Mothers		
	Component 1	Component 2	Communality	Component 1	Component 2	Communality	Component 1	Component 2	Communality
QED	.58	.10	.35	.82	−.26	.74	.77	.12	.61
Mechanical Comprehension	.48	.50	.48	.80	−.27	.71	.81	.29	.74
Concept Mastery test	.74	−.53	.82	.79	−.46	.83	.76	−.55	.88
Synonym-Antonym	.73	−.44	.72	.63	.68	.86	.74	−.54	.84
General Information	.83	−.09	.70	.80	.35	.77	.77	−.05	.59
Semantic Comprehension	.72	−.24	.57	.80	.13	.65	.78	−.21	.66
Cubes	.43	.71	.69	.64	−.59	.76	.54	.65	.71
Rotation Inversion	.49	.69	.72	.74	−.50	.80	.74	.46	.76
Eigenvalue	3.3	1.8		4.6	1.6		4.4	1.4	
Percent of Variance Explained	41.	22.		57.	19.		55.	17.	

TABLE 5
Rotated Factor Matrix of the First Two Principal Components in Table 4 Using Kaiser Normalization Procedure Separately for Children (N=72), Fathers (N=43), and Mothers (N=45)

	Children		Fathers[a]		Mothers	
	Factor 1_c	Factor 2_c	Factor 1_f	Factor 2_f	Factor 1_m	Factor 2_m
QED	.41	.33	.24	.74	.39	.54
Mechanical Comprehension	.08	.67	.21	.73	.26	.72
Concept Mastery Test	.93	–.20	.88	.08	.97	–.10
Synonym-Antonym	.87	–.12	.99	–.22	.95	–.10
General Information	.74	.26	.78	.19	.53	.38
Semantic Comprehension	.74	.06	.58	.38	.69	.23
Cubes	–.09	.85	–.17	.92	–.22	.91
Rotation-Inversion	.03	.86	–.03	.90	.07	.84

[a]These factors were renumbered to be consistent with the order obtained for the mothers and children.

of the population. This study demonstrated that the same principle, inherent in the coefficient of correlation itself, operates when going from children to parents.

The group scored quite well on the various standardized tests administered. For the most part, however, the experimental tests ranged from slightly to considerably too difficult even for this highly able group at this point in time. Yet, for some individuals the tests were quite appropriate. As a result, the tests did partially perform their function of separating the students according to their ability. For the group as a whole, though, the Cubes Test appeared to be the only suitably difficult experimental test at this juncture.

The experimental verbal test correlated highly with the standardized verbal tests and loaded most highly on the same factor. The spatial-ability tests did not correlate highly with the verbal tests, but correlated highly with the mechanical reasoning tests and among the various tests themselves. This would be expected if these tests measure spatial ability. Therefore, even though the experimental tests tended to be difficult for the group, they seem to be measuring what they are intended to measure.

Because the experimental tests were somewhat speeded, we could not calculate internal consistency reliabilities for them. A *lower* bound estimate of the reliability of these tests could, however, be obtained by use of the formula for the coefficient of correlation corrected for attenuation due to errors of measurement. It appeared that the tests were reliable, and about equally so, for the group. Support was found for the caution in the Mechanical Comprehension Test (Form CC) Manual that it may be too difficult for women.

Performance on the tests was associated with age and education. Education related positively with performance on all the tests, but least with the spatial ability and mechanical comprehension tests. For the parents, age related negatively to performance, a result previously found for cross-sectional studies on development of abilities, but not usually found in longitudinal studies. It is probably an artifact of differences in education and opportunities. For the children, however, performance on the verbal tests improved with age. This clearly indicates that the students' verbal abilities are still developing.

It is interesting to note that age and education related least, or not at all, to ability on the spatial and mechanical comprehension tests. Development of spatial ability has been studied extensively (e.g., Wittig & Petersen, 1979). Puberty seems to halt the development of spatial ability (Waber, 1976, 1977). Our behavioral data are consistent with this viewpoint.

Component analyses were performed on the age-adjusted partial *r*s separately for children, fathers, and mothers. The factor structure was similar for the three groups. Two principal factors were derived for each. For the parents, the first two principal factors correlated about .40. This is evidence that a general factor (Spearman, 1904) may be operating. Thus, for the parents, overall high ability on one test tended to relate to high ability on the other tests. Upon rotation of the parents' principal factors, the first factor appeared to be Vernon's verbal-educational (V:ed) and the second factor Vernon's practical-spatial-mechanical (K:m). They could explain 76% of the variance in the fathers' scores and 73% of the variance in the mothers' scores.

For the children, the evidence for Spearman's (1904) *g* was less compelling, although it seemed perhaps to be more so for the girls than the boys. The two factors, which accounted for 63% of the variance, correlated only .25. The first factor appeared to be Vernon's verbal-education and the second factor Vernon's practical-spatial-mechanical.

Thus, our data seem to fit the model of intelligence postulated by Vernon (1961). First, some evidence for *g* was found. Since the students were selected for high ability and the parents were also highly able, it would be expected that the effect of *g* would be somewhat reduced in this group. In samples not selected for intellectual ability, *g* can account for as much as 50% of the variance in the matrix. The parents were less able than the children. As would be expected by this reasoning, the evidence for *g* was stronger for the parents than for the children. We also found factors resembling Vernon's V:ed and K:m. In conclusion, Vernon's (1961) model of intelligence seems to best explain the performance of our high-ability subjects.

Even among such highly intelligent individuals, sex differences were apparent. The most conspicuous result was the number of girls qualifying for this study of children with "precocity quotients" of 170 or more. For every girl qualifying, seven boys qualified. This was most likely the result of the girls' lower mathematical reasoning scores (Benbow & Stanley, 1980, 1981), which were an important component in ascertaining the "precocity quotients." Other sex differences in mental ability also occurred for the children and their parents. As has been found for populations unselected in ability (e.g., Burnett, Lane, & Dratt, 1979; Wittig & Petersen, 1979), high ability females do not perform as well as high ability males on mechanical comprehension or spatial ability tests. The girls tended to be younger than the boys, as few extremely precocious girls were found in the early talent searches. Age was positively related to performance on the tests, but much less so for the spatial ability and the mechanical comprehension tests. Thus, the age difference cannot account for these sex differences.

Another problem that the data in this article help address is assessing the degree of familiality of cognitive abilities and also of assortative mating. Benbow, Zonderman, and Stanley (this issue) did such an analysis. They showed that the parents resembled one another to a higher degree than the parents of less able children, and that the children resembled their parents to a lesser extent than less able children resemble their parents. This indicated that the etiology of individual differences at the high end of the distribution of intellectual abilities may *not* be similar to that in the rest of the distribution.

In conclusion, this study demonstrated that highly able children tend to have highly able parents. Moreover, Vernon's (1961) model of the structure of intelligence appears to best fit our data. Verbal ability continues to develop during adolescence, but mechanical comprehension and spatial ability seem not to increase much during this period.

REFERENCES

Angoff, W. (Ed.) *The College Board admissions testing program.* Princeton, NJ: College Entrance Examination Board, 1971.

Bayley, N., & Oden, M. H. The maintenance of intellectual ability in gifted adults. *Journal of Gerontology,* 1955, *10,* 91–107.

Benbow, C. P., & Stanley, J. C. Sex differences in mathematical ability: Fact or artifact? *Science,* 1980, *210,* 1262–1264.

Benbow, C. P., & Stanley, J. C. Mathematical ability: Is sex a factor? *Science,* 1981, *212,* 118–121.

Burks, B. S., Jensen, D. W., & Terman, L. M. The promise of youth: Follow-up studies of a thousand gifted children. *Genetic studies of genius* (Vol. 3). Stanford, CA: Stanford University Press, 1947.

Burnett, S. A., Lane, D. M., & Dratt, L. M. Spatial visualization and sex differences in quantitative ability. *Intelligence,* 1979, *3,* 345–354.

Eisdorfer, C., & Wilkie, F. Intellectual changes with advancing age. In L. F. Jarvik, C. Eisdorfer, & J. E. Blum (Eds.), *Intellectual functioning in adults.* New York: Springer, 1973.

Jarvik, L. F., Eisdorfer, C., & Blum, J. E. *Intellectual functioning in adults.* New York: Springer, 1973.

Kirk, M. K. Personal Communication, Nov. 1979.

Kirk, M. K. *The form of the curve of distribution of intellectual superdeviance: An empirical study.* Unpublished manuscript, 1978. (a)

Kirk, M. K. *The standardization of the ASMT, Form AE: Technical report.* Unpublished manuscript, 1978. (b)

Kirk, M. K. *Field testing of the ASMT—BE: Preliminary results, interpretations, and revision.* Unpublished manuscript, 1979.

Kirk, M. K. *Generic end-testing of the intellectually brilliant at early maturity.* Unpublished bachelor's thesis, Harvard University, 1980.

Matarazzo, J. D. *Wechsler's measurement and appraisal of adult intelligence* (5th ed). Baltimore, MD: Williams & Wilkins, 1972.

Messick, S., & Jungeblut, A. Time and method in coaching for the SAT. *Psychological Bulletin,* 1981, *89,* 191–216.

Nie, N. H., Hull, C. H., Jenkins, J. G., Skinbrenner, K., & Bent, D. H. *SPSS: Statistical package for the social sciences* (2nd ed). New York, NY: McGraw-Hill Book Co., 1975.

Oden, M. H. The fulfillment of promise: 40-year follow-up of the Terman gifted group. *Genetic Psychology Monographs,* 1968, *77,* 3–93.

Owens, W. A. Age and mental abilities: A second adult follow-up. *Journal of Educational Psychology,* 1966, *57,* 311–325.

Owens, W. A., & Bennett, G. K. *Manual for the Mechanical Comprehension Test—Form CC.* New York, NY: The Psychological Corporation, 1949.

Schaie, K. E., & Strother, C. R. A cross-sectional study of age changes in cognitive behavior. *Psychological Bulletin,* 1968, *57,* 326–336.

Spearman, C. General intelligence, objectively determined and measured. *American Journal of Psychology,* 1904, *15,* 201–293.

Stake, R. E. A non-mathematical quantitative aptitude test for the graduate level: The QED. *Journal of Experimental Education,* 1962, *1,* 81–83.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed). Washington, DC: American Council on Education, 1971.

Terman, L. M. Mental and physical traits of a thousand gifted children. *Genetic studies of genius* (Vol. 1). Stanford, CA: Stanford University Press, 1925.

Vernon, P. E. *The structure of human abilities* (2nd ed). London: Methuen, 1961.

Waber, D. P. Sex differences in mental abilities: A function of maturation rate? *Science,* 1976, *192,* 572–574.

Waber, D. P. Sex differences in mental abilities, hemispheric lateralization, and rate of physical growth at adolescence. *Developmental Psychology,* 1977, *13,* 29–38.

Wechsler, D. *The measurement and appraisal of adult intelligence.* Baltimore, MD: Williams & Wilkins, 1958.

Wittig, M. A., & Petersen, A. *Sex-related differences in cognitive functioning: Developmental issues.* New York: Academic Press, 1979.