Journal of Educational Psychology
1992, Vol. 84, No. 1, 51–61

Academic Achievement in Mathematics and Science of Students Between Ages 13 and 23: Are There Differences Among Students in the Top One Percent of Mathematical Ability?

Camilla Persson Benbow
Iowa State University

The predictive validity of the Scholastic Aptitude Test–Mathematics subtest (SAT-M) was investigated for 1,996 mathematically gifted (top 1%) 7th and 8th graders. Various academic achievement criteria were assessed over a 10-year span. Individual differences in SAT-M scores obtained in junior high school predicted accomplishments in high school and college. Among students in the top 1% of ability, those with SAT-M scores in the top quarter, in comparison with those in the bottom quarter, achieved at much higher levels through high school, college, and graduate school. Of the 37 variables studied, 34 showed significant differences favoring the high SAT-M group, which were substantial. Some gender differences emerged; these tended to be smaller than the ability group differences; they were not observed in the relationship between mathematical ability and academic achievement. The predictive validity of the SAT-M for high-ability 7th and 8th graders was supported.

"Standardized testing is much in the news. New testing programs, test results, and criticisms of standardized testing all are regular fare in the popular media today" (Haney, 1981, p. 1021). Moreover, "with the possible exception of evolution, no area in the sciences has been as filled with emotional and confusing mixtures of science, politics, and philosophy as the field of mental testing" (Carroll & Horn, 1981, p. 1012). These remarks portray quite well the status of mental testing at the beginning of the 1980s, yet they seem to be equally appropriate for describing mental testing at the beginning of the 1990s. Some might perceive this as a rather recent development. However, concern over standardized testing has been voiced ever since the introduction of the Stanford–Binet Intelligence Scale and the Army Alpha test (Cronbach, 1975; Haney, 1981).

The concerns over mental testing primarily have been threefold: test bias against certain groups (primarily women and minorities at present, but children from families of low socioeconomic status in earlier decades), the role testing might play in perpetuating social and economic injustice, and the utility of test information (Cleary, Humphreys, Kendrick, & Wesman, 1975; Cole, 1981; Gottfredson & Crouse, 1986; Haney, 1981; Jensen, 1980; Scarr, 1981). The questionable value of test information has been a particularly frequent criticism levied against college admissions tests, such as the College Board Scholastic Aptitude Test (SAT; see Linn, 1982b, for a review). This study was conceptualized to address the latter concern, namely, the predictive validity of the SAT for a special population. I assess the value of the SAT, not for high school seniors and the college admissions process, but rather for identifying highly mathematically gifted seventh and eighth graders and making predictions about their achievement over a 10-year period following their SAT-Mathematics assessment. Specifically, I asked whether the SAT-M can detect individual differences in the top 1% of the ability continuum that bear on subsequent academic achievement in mathematics and science.

The use of the SAT to identify intellectually precocious students in Grades 7 and 8 dates to 1972 when Julian Stanley launched the first talent search (Keating & Stanley, 1972). Stanley was interested in students who ranked in the top 1% in mathematical ability. Because considerable variance in academic ability is found among students in the 99th percentile and because Stanley was interested in differentiating among such students, out-of-level testing (i.e., using tests designed for older age groups) was required. For that reason among others (see Stanley & Benbow, 1986), Stanley chose the SAT as the instrument with which to screen highly gifted students. Since 1972 more than 1,000,000 seventh and eighth graders have been tested with the SAT, and more than 100,000 such students now take the SAT annually through various talent search programs across the United States. The distribution of scores of such students on the SAT is about the same as found for a random sample of high school students (Benbow, 1988). The scores tend to maintain their ordinal ranking over time, increasing 40 to 50 points per year (Benbow & Stanley, 1982; Brody & Benbow, 1990; Olszewski-Kubilius, 1990). Thus, from a psychometric viewpoint, the use of the SAT with seventh and eighth graders seems justified.

It has not been demonstrated, however, whether use of the SAT with young but academically competent students has utility. Is the SAT a valid tool for assessing individual differences in current development, and can this instrument be used to refine predictions of exceptional academic achieve-

Financial support for this study was provided by grants from the College Board and the National Science Foundation (MDR 8855625).

I would like to thank David Lubinski for helpful comments and suggestions.

Correspondence concerning this article should be addressed to Camilla Persson Benbow, Department of Psychology, Iowa State University, Ames, Iowa 50011-3180.

ments? As Cronbach (1971) pointed out, "validation is the process of examining the accuracy of a specific prediction or inference made from a test score" (p. 471). In assessing the validity of the SAT for highly gifted 7th and 8th graders, I evaluated whether academic achievement, especially in mathematics/science, during the 10-year period after these students were identified is much higher for those students with exceptionally high SAT Mathematics subtest (SAT-M) scores (top quarter of the top 1%) than for those with comparatively "low" SAT scores who were nonetheless in the top 1% in ability (i.e., the bottom quarter of the top 1%). I hypothesized that meaningful differences would be detected. Several studies have revealed that individuals with the most potential for high academic achievement in mathematics and science are generally considered to be those students with high ability, particularly, high mathematical ability (Davis, 1965; Green, 1989; Walberg, Strykowski, Rovai, & Hung, 1984; Werts, 1967). Moreover, Kuhn (1962) noted that an overwhelming majority of "scientific revolutions" can be ascribed to the works of mathematically brilliant persons.

Nevertheless, many researchers and educators, most notably Renzulli (1986), have argued that there is a threshold effect for ability. According to this argument, after a certain point, there is a decline in the power of ability to influence academic achievement and other variables, such as motivation and creativity, become increasingly important. The precise location of this threshold for ability has not been determined. However, it is thought to be at some point well below the top percentile for ability. If Renzulli and others of this viewpoint are correct, then there should be no statistically significant differences in mathematics/science achievement between the two high-ability groups. All students in the top 1% should achieve highly, and placement within the top 1% should not affect the results.

The reasoning in the above paragraph assumes that there is only one threshold for ability. Yet there could be a threshold effect for ability within a certain range (e.g., between the 90th and 98th percentiles) but not within the top 1%. That is, differences in ability within the 90th and 98th percentiles may not relate much to subsequent academic achievement in mathematics/science. This view is reasonable given that the possible differences in ability within a range, for example, within the 90th–98th or 80th–89th percentile ranges, are small and not reliable in comparison with the ability differences found within the top 1% when out-of-level testing is used. I do not test this possibility in this study. If, however, one is interested in scientific eminence or productivity, and a threshold effect of ability for this level of achievement, it is within the top percentile of ability that one must focus.

Although my prediction is contrary to Renzulli's position, it should be noted that there are data that support the validity of Renzulli's position. For example, students who were in the top 1% in mathematical ability in the 7th and 8th grades were studied at 23 years of age to identify those factors that affect the ways in which childhood potential or ability is translated into adult achievement (Benbow & Arjmand, 1990). As a group these students had achieved academically at a very high *level but not uniformly* so. When those students who were classified as high academic achievers in mathematics/science areas (i.e., those who were attending graduate school in mathematics/science or medical school; n = 261) were compared with those students in the sample who were classified as low academic achievers in those areas (those who were not attending college or had withdrawn, those who graduated with mathematics/science major but with low grades; n = 95), a difference in previous ability between the two groups was found (the ability difference approximated two thirds of a standard deviation on the SAT-M). The canonical correlation (from the discriminant analysis) between (a) 7th-grade/8th-grade SAT-M and (b) high school SAT-M, SAT Verbal subtest (SAT-V), and achievement group membership was .30 for male students and .29 for female students. (Too few cases had 7th-grade/8th-grade SAT-V scores to allow inclusion in the analysis.) Nonetheless, ability exhibited the weakest relationship with academic achievement in mathematics/science as compared with variables in the areas of educational opportunity, family characteristics, and attitudes. Similarly, Sanders, Benbow, and Albright (1991) found that among mathematically talented female students, previous ability on SAT-M was not a primary factor relating to choice of mathematics/science career or to educational aspirations.

Thus, the aforementioned studies indicate that among those students in the top 1%, SAT-M performance was a factor but not the major factor predicting the students' academic success. That is, a bright mind will not make its own way. The educational opportunities provided to gifted children make a difference in the children's development. In the present study, I ask the central question: Do individual differences within the top 1% in ability make a difference in the eventual display of achievement?

In sum, I examine whether use of the SAT in out-of-level testing of highly gifted students yields useful information for the prediction of academic achievement up to 10 years after assessment. That is, is it useful to diagnose level of talent within the top 1%, as is currently being done with well over 100,000 seventh- and eighth-grade students on an annual basis? More succinctly, is there a benefit to knowing where in the top 1% a student's ability lies? It has been popularly assumed that such information is not helpful. In essence, I assess the predictive validity of the SAT for use with gifted 7th and 8th graders.

Method

Subjects

Intellectually talented students were identified by the Study of Mathematically Precocious Youth (SMPY), in which the SAT was administered to intellectually able 12- and 13-year-olds in the 1970s and early 1980s (Keating & Stanley, 1972). During that 12-year period, more than 10,000 preadolescents (mostly 7th graders) participated in SMPY "talent searches." (Since that time more than 1 million students have taken the SAT through other talent search programs.) About 3,500 of the students in the talent searches were included in the SMPY 50-year longitudinal study. As part of this study, researchers in the SMPY are currently tracking four cohorts of students and studying their development longitudinally.

Students in Cohort 1 comprised the sample in this investigation; they were drawn from the first three talent searches of the SMPY (i.e., those conducted in 1972, 1973, and 1974). In those talent searches, 7th and 8th graders in Maryland were eligible to participate if they had scored in the upper 5% (1972) or the upper 2% (1973, 1974) nationally on any standardized mathematics achievement test. Qualified students took the SAT-M and, in 1973, the SAT-V also. These tests are designed to measure developed mathematical and verbal reasoning ability, respectively, of high school students. However, the SAT is believed to be a more potent measure of reasoning for 7th and 8th graders than for 11th and 12th graders (Minor & Benbow, 1986; Stanley & Benbow, 1986).

A score of at least 390 on the SAT-M or 370 on the SAT-V in the 7th or 8th grade was required for inclusion in Cohort 1 of the longitudinal study. These SAT criteria resulted in the selection of 2,118 of 2,582 students who, as 7th or 8th graders, scored as well as the average high school female; the criteria also provided a wide range of talent to study. SAT scores had been grade adjusted (7th-grade scores had been adjusted upward to be comparable to 8th-grade scores, with the procedure outlined in Angoff, 1971). Mean SAT scores at age 13 were as follows for male students, 556 ($SD = 73$) on SAT-M and 436 ($SD = 85$) on SAT-V, and for female students, 519 ($SD = 59$) on SAT-M and 462 ($SD = 88$) for SAT-V. Approximately 4 years later, in high school the mean scores had increased to 695 ($SD = 70$) on SAT-M and 593 ($SD = 88$) on SAT-V for male students. For female students the mean scores had increased to 650 ($SD = 71$) on SAT-M and 599 ($SD = 89$) on SAT-V.

In this study, as detailed below, there were 2,118 students participating at age 13 years; 1,996 students at age 18 years; and 1,247 at age 23 years. I estimate that the students' abilities are approximately in the top 1%. This estimate is derived from the three screenings used to select students for this study (i.e., the talent search cutoff, the self-selection of students for the talent search, and the selection criteria for the longitudinal study).

On the basis of SAT scores in the 8th grade, each student was placed in one of three groups. Only two of these groups were targeted for study. The high SAT-M group or top quarter group included students in the top quarter of SAT-M scores, whereas the low SAT-M group consisted of students in the bottom quarter of SAT-M scores. The SAT scores in 8th grade and again at the end of high school, as well as the number of students in each of the two groups, are shown in Table 1. It is clear that even students in the low SAT group had scores in the 8th grade that were comparable to those of college-bound seniors and thus were highly able.

Procedure

All talent search participants completed a brief background questionnaire before they took the initial SAT at 12–14 years of age. Students were first surveyed longitudinally at age 18 with an 8-page questionnaire (Benbow, 1983; Benbow & Stanley, 1982). A second follow-up survey with a 24-page printed questionnaire was administered at age 23 years (Benbow & Arjmand, 1990).[1] In both follow-up surveys, participants were first mailed the questionnaire in late fall, along with a letter encouraging them to participate. For the survey administered to 18-year-olds, it was possible to offer a monetary inducement (i.e., $5 or $6). Nonrespondents were reminded by letter, and then by postcard. Those individuals who did not respond by the following summer were telephoned and eventually asked to provide responses orally.

With 1,996 students responding, a 91% response rate was obtained for the survey of 18-year-olds. Initial response rate to the second follow-up at age 23 years was 65%. Because viability of a longitudinal study depends on the retention of a large proportion of the original sample, nonrespondents were surveyed by telephone with 20 critical questions. This increased the response rate to about 70%. The sample at age 23 years included 786 male students and 461 female students.

Discriminant analyses were computed separately for male and female students to determine whether nonrespondents at age 23 differed from respondents on the basis of 8th-grade SAT-M score, high school SAT-M and SAT-V scores, college attendance, quality of college attended, parental educational levels, number of siblings, and father's occupational status. (Too few students had completed SAT-V tests in 8th grade to allow the inclusion of these scores.) No statistically significant differences existed between respondents and nonrespondents. The largest difference, which favored the nonrespondents, was 0.18 S.D. for father's occupational status.

Statistical Analyses

Responses to essentially all questions on the post–high school and post–college questionnaires that pertained to academic achievement were selected for analysis. Most of these variables were in the mathematics/science area, as that is a focus of the SMPY. Specifically, I analyzed self-reported course taking in mathematics/science, course grades, honors or awards, outside-of-class academic achievement (i.e., math contests, science fairs, working on a research project, publishing a paper), achievement test scores in mathematics and science, educational aspirations, graduate school attendance, field of study, and career goal. The intellectual and status level of colleges attended were ascertained by using the Astin (1965) scale.

First, correlations between SAT-M in 8th grade and the continuous criterion variables were computed by gender. Because of the large sample size (e.g., $n = 1{,}996$ for some of the high school data), I set the significance level at .01. Cohen (1988) classified correlations as small effects if $.1 \leq r < .3$, as medium effects if $.3 \leq r < .5$, and as large effects if $r \geq .5$. A medium effect size is described by Cohen (1988) as the "degree of relationship [that] would be perceptible to the naked eye of a reasonably sensitive observer" (p. 80), whereas Cohen categorized large effect sizes as effects "about as high as they come" (p. 81). Of course, not all researchers would accept his interpretation. In this study the power to detect a medium effect size, which I viewed as important or useful, was greater than .995.

Next, for the continuous criteria, I compared students who were in the top and bottom quarter using analyses of variance (ANOVA), with SAT group (high vs. low) and gender as variables. Because of the unequal sample sizes of the subgroups, the ANOVAs were nonorthogonal. I decided to retain a nonorthogonal design (because the larger the sample size, the greater is the statistical power) and to use the regression or simultaneous approach for decomposing sums of squares in the ANOVA. I analyzed categorical data using chi-squares, in separate analyses for group and sex. In this phase of analysis, alpha was set at .05. (Our sample size was considerably smaller for this portion of the study.) For all statistically significant differences we calculated effect sizes: for means, $d = (M_1 - M_2)/SD$, average SD used (Cohen, 1988). The effect size for proportions, h, is determined by an arcsine transformation of each proportion, followed by calculation of the difference. Cohen (1988) arbitrarily classified effect sizes as small if $.2 \leq h, d < .5$; medium if $.5 \leq h, d < .8$; and large if $h, d \geq .8$. The power to detect a medium effect size was at least .80.

[1] Many students also supplied us with transcripts for our use in determining grades and coursework patterns.

Table 1
Scholastic Aptitude Test (SAT) Scores in 7th Grade/8th Grade and at the End of High School of Students Who Are in the Top 1% in Ability, by Gender and SAT Group

Sample/	Top quarter: Male students			Top quarter: Female students			Bottom quarter: Male students			Bottom quarter: Female students		
measure	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
8th-grade[a] SAT-M	640	48	367	622	31	100	462	22	248	462	24	282
8th-grade[a] SAT-V	463	89	109	514	86	29	401	70	75	436	87	93
High school SAT-M	747	42	362	714	49	96	631	71	216	613	63	250
High school SAT-V	631	79	361	654	76	96	540	85	216	564	84	250

Note. M = Mathematics subtest; V = Verbal subtest.
[a] 7th-grade scores were adjusted upward to be comparable to 8th-grade scores.

Results

Scholastic Aptitude Test Scores and Achievement: Correlations

Tables 2 and 3 present the correlations between 8th grade SAT-M score and various variables reflecting achievement in high school and college, respectively. In cases in which the criterion variable was an actual achievement test score in math/science or reflected the student's decision to take a math/science achievement test, the correlations tended to be sizable for the most part, ranging from .16 to .57 (all but two were greater than .3). The highest correlations tended to occur with physics achievement scores, followed by mathematics test scores. For the remaining nontest variables, the correlations tended to be positive and small, but significant nonetheless. Some of the largest correlations with 8th-grade SAT-M score occurred with reported grade point average (GPA) in college, GPA in college math/science courses, and intellectual level of college attended.

In sum, there is a relationship between SAT-M score in 8th grade and subsequently reported measures of achievement. For 20 of the 36 correlations, this relationship is small, with the remaining correlations being medium or large.[2]

Achievement: High Versus Low Scholastic Aptitude Test–Mathematics Groups

High school achievement. Shown in Table 4 are the descriptive statistics for the high school variables by SAT-M group (i.e., top quarter vs. bottom quarter of the select group) and by gender. It is clear that the achievement of all groups is high, as would be expected given their relative ability. Moreover, there is considerable variability in achievement within each ability group. It is also evident that the higher SAT-M group had achieved much more than the lower SAT-M group in high school. In ANOVAs that were computed on the continuous high school achievement variables (i.e., on the measures for which means and standard deviations are reported in Table 4), the effect of SAT-M group was significant at the .001 level (with a range in F values from 11.3 to 177.9) in every instance, except College Board Biology Achievement Test score, for which $F(1, 80) = 5.2$, $p < .05$. The effect sizes for the differences between the two groups tended to be large. The range of d values was .30 to 1.69, with 8 of 12 values greater than .8 (see Table 4).

Significant differences between the SAT-M groups on the categorical high school variables (i.e., those variables for which a percentage is reported in Table 4) were tested with a chi-square analysis. Group differences were all significant at the .01 level, with effect sizes ranging from .41 to .82. The only

[2] Correlations between the variables in Table 3 and College Board Math Achievement tests also were computed. Students took the College Board Math Achievement tests an average of 4 years after they took the 7th-grade/8th-grade SAT-M. Nonetheless, the resulting correlations were generally of the same magnitude as for 7th-grade/8th-grade SAT-M.

Table 2
Correlations Between 8th-Grade Scholastic Aptitude Test–Mathematics (SAT-M) Score and Variables Assessing Academic Achievement in High School for Male and Female Students

Measure/ sample	No. semesters math		No. science courses		No. math/ science achieve-ment/ AP tests		College Board Achievement Test										AP calculus test				No. AP exams		No. academic awards	
							Math 1		Math 2		Biology		Chemistry		Physics		AB		BC					
	r	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*
8th-grade SAT-M																								
Male students	.15	1,185	.08	1,185	.42	1,185	.39	233	.47	411	.39	93	.50	218	.52	161	.38	139	.38	204	.33	1,186	.11	1,186
Female students	.11	749	.08	749	.31	749	.49	174	.47	134	.16	67	.32	74	.57	24	.44	61	.42	34	.21	749	.13	749

Note. All correlations are significant at least at the .01 level, except for number of high school science courses ($p < .05$) among female students. AP = Advanced Placement; AB = form that measures one semester of college calculus; BC = form that measures two semesters of college calculus.

Table 3
Correlations Between 8th-Grade Scholastic Aptitude Test–Mathematics (SAT-M) Score and Variables Assessing Academic Achievement in College for Male and Female Students

Measure/ sample	Intellectual level of college		Status level of college		College GPA		No. honors in college		No. college math/science courses		College math/science GPA	
	r	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*
8th-grade SAT-M												
Male students	.25	1,069	.09	1,069	.18	689	.14	663	.19	761	.26	176
Female students	.22	682	.18	682	.15	379	.19	400	.09	454	.19	76

Note. All correlations are significant at least at the .01 level, except for number of college math/science courses and college math/science grade point average (GPA) among female students.

Table 4
Academic Achievement in High School of Students in the Top 1% in Ability, by Group and by Gender

	Top quarter: Male students (*n* = 367)			Top quarter: Female students (*n* = 100)			Bottom quarter: Male students (*n* = 248)			Bottom quarter: Female students (*n* = 282)			Effect size[a]	
Measure	*M*	*SD*	%	*M*	*SD*	%	*M*	*SD*	%	*M*	*SD*	%	Ability group	Gender
No. high school semesters of math	9.6	2.6		9.0	2.6		8.5	2.7		8.2	2.3		0.42	0.29
No. high school science courses	4.1	1.3		3.7	1.2		3.6	1.4		3.6	1.2		0.31	0.21
Took high school calculus			85			60			51			33	0.82	0.64
Took high school physics			84			72			67			54	0.47	0.39
Completed precalculus curriculum with essentially all *A* grades			65			54			27			37	0.63	0.18
Took biology, chemistry, and physics with all *A* grades			38			37			20			17	0.45	0.21
No. math/science achievement or AP tests	2.4	1.8		1.5	1.4		0.7	1.2		0.5	0.9		1.11	0.66
College Board Math 1 Achievement Test	726	60		698	71		647	71		615	64		1.31	0.75
College Board Math 2 Achievement Test	774	43		750	49		700	64		662	69		1.56	0.96
College Board Biology Achievement Test	688	62		638	74		626	86		614	100		0.71	—
College Board Chemistry Achievement Test	711	69		668	87		609	100		613	57		1.18	—
College Board Physics Achievement Test	706	61		664	53		609	89		553	102		1.69	1.01
AP calculus exam score														
AB	4.1	0.9		4.1	0.9		3.0	1.1		2.8	1.1		1.24	—
BC	4.1	1.0		4.1	1.0		3.1	1.2		3.0	0.7		0.96	—
No. AP exams taken	1.5	1.7		0.9	1.2		0.4	0.8		0.3	0.8		0.86	0.51
Took college course as high school student			28			27			9			15	0.41	—
Academic awards in high school	2.9	3.3		2.9	2.6		1.9	2.6		2.2	2.6		0.30	—
Participated in math contests			36			21			11			5	0.65	0.46
Participated in science fairs			15			16			22			17	—	—

Note. AP = Advanced Placement; AB = form that measures 1 semester of college calculus; BC = form that measures 2 semesters of college calculus.
[a] Effect sizes were not computed for cases in which $p > .05$, as indicated by dashes.

exception was participation in science fairs, for which no statistically significant group differences were revealed.

The average effect sizes for the group differences on the variables during the high school years were .97 ($SD = .47$) for the continuous variables and .57 ($SD = .16$) for the categorical variables.

College achievement. Descriptive statistics for the variables characterizing the college years are displayed in Table 5. Again, high achievement is displayed by both ability groups (e.g., in both groups a large number of students earned college degrees with strong academic records); there is variability in achievement within both groups; and a series of clear-cut differences favoring the higher SAT-M group is evident. ANOVAs were computed for the continuous measures of achievement. The effect of group was statistically significant at the .001 level for all continuous variables, except one (mathematics/science GPA; $p < .01$), with the F values ranging from 7.6 to 60.0. The associated effect sizes for the group differences that were statistically significant ranged from .23 to .69 (see Table 5). The largest differences were found for intellectual level of college attended and mathematics/science GPA.

For the categorical variables, all differences were significant at the .01 level in a chi-square analysis except the variables of graduating in the top 10% of the student's class ($p > .05$) and participating in special college-level mathematics/science programs ($p > .05$). The range of effect sizes for statistically significant differences was .21 to .61. The largest difference occurred with the variable mathematics/science major. The average effect sizes for the statistically significant group differences on the achievement variables in college were .47 ($SD = .18$) for the continuous variables and .39 ($SD = .13$) for the categorical variables.

Graduate school achievement. Because the latest follow-up survey was completed when students were approximately 23 years old, relatively little information was available for the graduate school stage. The information that was available is displayed in Table 6. More students in the high ability group than in the low ability group were attending graduate school, aspiring toward a doctorate, specializing in mathematics/science areas, and possessed career goals in mathematics/science areas. All differences were significant at the .01 level, with effect sizes ranging from .35 to .49. The average effect size for the ability group achievement differences in graduate school was .41 ($SD = .06$).

Achievement by type of task. I decided to analyze the data from a different perspective. The academic achievement variables were categorized according to task rather than developmental stage (i.e., high school, college, graduate school). I computed an average effect size for the difference between the top quarter and the bottom quarter groups on the variables comprising each category. It should be noted that some variables did not fit into any category and were thus excluded from this analysis of the data. The categories and mean effect sizes for the group differences, were, in order of magnitude, as follows: standardized test scores ($d = 1.24$); grades earned for coursework ($d = .50$, $h = .54$); course taking ($d = .42$, $h = .65$); mathematics/science career goals ($h = .48$); educational aspirations ($h = .41$); out-of-class academic experiences, such as research participation ($h = .41$); and prizes and awards ($d = .27$, $h = .41$). Clearly, SAT-M scores predict future standardized test scores best and awards and honors worst.

Gender Differences in Achievement

The relationship between comparatively higher versus lower SAT groups and achievement variables in high school, college, or graduate school did not appear to vary as a function of gender (see Tables 2–5). That is, similar relationships between ability group and achievement were found for both male and female students. Not surprisingly then, the interaction terms in the ANOVAs generally were not statistically significant. There were three exceptions: number of college mathematics/science courses completed, $F(1, 627) = 5.4$, $p < .05$; number of Advanced Placement (AP) examinations taken, $F(1, 992) = 7.9$, $p < .01$; and number of mathematics/science achievement/AP tests taken in high school, $F(1, 992) = 12.7$, $p < .001$. These significant interactions occurred because the gender difference in the dependent variable was much larger for the high-ability group than for the low-ability group.

Although there were no apparent gender differences in the relationship between ability group and achievement, for many of the variables, gender differences in means or proportions were noted (see Tables 3–5). Female students tended to exhibit better classroom performance, as reflected in grades and academic honors, whereas male students tended to participate in the mathematics/science areas to a greater extent, to exhibit better performance on standardized mathematics/science achievement tests, and to have higher educational aspirations. The gender differences in the variables were much smaller than the ability group differences, as judged by their associated effect sizes (see Tables 3–5). The average effect sizes for the 34 statistically significant ability group differences were .80 ($SD = .45$) for the 18 continuous variables and .46 ($SD = .15$) for the 16 categorical variables. In contrast, the average effect sizes for the 24 statistically significant gender differences favoring male students were .57 ($SD = .30$) for the 10 continuous variables and .34 ($SD = .16$) for the 14 categorical variables. Interestingly, the only gender difference that was larger than the ability group difference occurred in the percentage of students majoring in mathematics/science areas (62% of male students and 30% of female students).

Discussion

This 10-year longitudinal study addressed the predictive validity of the SAT-M among 7th- and 8th-grade students who were known to be in the top 1% of students of their age group in ability. The students' placement within the top 1% was then used to make predictions about their subsequent academic achievement. Specifically, I investigated the following question: Among students, about whom the only thing known is that they are in the top 1% in mathematical ability, should predictions of subsequent academic achievement be substantially higher for those who are in the top quarter than for those who are in the bottom quarter of this truncated segment? The answer is clearly affirmative. Of the 37 variables

Table 5
Academic Achievement in College of Students in the Top 1% in Ability, by Group and by Gender

	Top quarter: Male students (*n* = 250)			Top quarter: Female students (*n* = 62)			Bottom quarter: Male students (*n* = 154)			Bottom quarter: Female students (*n* = 168)			Effect size[a]	
Measure	*M*	*SD*	%	*M*	*SD*	%	*M*	*SD*	%	*M*	*SD*	%	Ability group	Gender
Intellectual level of college attended	62.7	11.9		57.7	15.2		53.5	15.1		49.1	17.5		0.69	0.50
Status ranking of college attended	57.6	11.5		56.3	12.7		53.8	13.5		50.7	15.4		0.39	0.29
Earned bachelor's degree			92			89			75			82	0.34	—
College GPA	3.4	0.5		3.5	0.4		3.2	0.5		3.4	0.4		0.34	—
Top 10% of class			36			42			32			32	—	—
No. honors in college	1.9	1.6		2.5	1.7		1.3	1.4		1.6	1.6		0.23	—
Math/science major			72			40			47			26	0.61	0.65
No. math/science courses in college	13.9	10.6		8.8	8.5		8.6	9.5		7.5	8.1		0.53	0.43
Math/science GPA	3.4	0.5		3.5	0.4		3.0	0.6		3.3	0.4		0.65	—
Worked on research project			32			26			23			20	0.21	0.16
Published journal article			17			10			6			6	0.33	0.20
Award in math/science			10			3			2			1	0.41	0.34
Participated in math/science contest			13			0			1			2	0.44	0.36
Participated in special math/science program			5			5			1			4	—	—

Note. GPA = grade point average.

[a] Effect sizes were not computed when $p > .05$, as indicated by dashes.

Table 6
Achievement at the Graduate School Level of Students in the Top 1% of Ability, by Group and by Gender

Measure	Top quarter: Male students (*n* = 250)	Top quarter: Female students (*n* = 62)	Bottom quarter: Male students (*n* = 154)	Bottom quarter: Female students (*n* = 168)	Effect size	
					Ability group	Gender
Attending graduate school	58%	42%	32%	34%	0.45	0.24
Aspiring toward a doctorate	44%	34%	27%	24%	0.36	0.24
If attending graduate school, specializing in math/science	71%	39%	43%	39%	0.49	0.42
Possessing a career goal in math/science	56%	24%	32%	34%	0.35	0.33

measuring academic achievement in primarily mathematics/science areas during a 10-year period, 34 showed statistically significant differences favoring the high SAT-M group and were substantial. Differences were somewhat larger in high school than in college or in graduate school. This was primarily a result of SAT-M scores in 7th and 8th grade being especially good predictors of future standardized test performance. The largest differences between the two ability groups occurred in subsequent standardized test performance. Yet the average effect size for the remaining differences approached a medium level of magnitude. Therefore, I conclude that just as the predictive validity of the SAT-M has been demonstrated for the general population (Linn, 1982a), the SAT-M also appears to have predictive validity for differentiating highly able 7th-grade/8th-grade students. This test can identify a pool of future scientists who might meet our nation's technological needs. This indicates that there is usefulness in having high-ability students take the SAT-M at an early age, as more than 100,000 students now do annually.

The results of this study also address the issue of whether there is an "intellectual threshold" for academic achievement, beyond which higher levels of ability are irrelevant. Because there are differences in subsequent achievement between those in the top quarter and those in the bottom quarter of the top 1%, the results did not support the notion of a threshold effect for ability. Of course, I did not rule out the possibility that threshold effects could be operating within the top 10%, after excluding the top 1%. That is, there might be no predictive difference between the 90th percentile and the 98th percentile, yet there is a group of highly able students in the top percentile who distinguish themselves. Moreover, the present finding does not imply that motivation, creativity, self-management skills, educational opportunity, and so forth cannot compensate for lower levels of ability, as argued by Renzulli (1986). Indeed, the variability in achievement within each ability group was sizable. Moreover, the results of Benbow and Arjmand (1990), Phye and Benbow (1991), and Sanders, Benbow, and Albright (1991) are consistent in that respect. These investigations reveal that, among those in the top 1% in ability, educational opportunity, family background, and attitudes have stronger relationships with subsequent academic achievement than does ability. One might interpret these findings to mean that ability on the SAT-M is a measure of potential, whereas educational opportunity, family characteristics, and attitudes are some of the factors that determine whether childhood potential is translated into adult achievement. (For a recent discussion of the importance of these nonintellectual factors, see Lubinski & Humphreys, 1990.)

There appeared to be no gender differences in the relationship between ability on the SAT-M and subsequent academic achievement. Nonetheless, gender differences in mean achievement were noted. In general, the female students tended to exhibit better classroom performance, as reflected in grades and academic honors. In contrast, the male students tended to participate in the math/science areas to a greater extent, to exhibit better performance on standardized math/science achievement tests, and to have higher educational aspirations (cf. Lubinski & Humphreys, 1990). There were also many more male than female students in the high SAT-M group. Such achievement and ability differences have been documented previously for this group (e.g., Benbow & Arjmand, 1990; Benbow & Minor, 1986; Benbow & Stanley, 1980, 1982, 1983). Yet it was distressing to note that gender differences in mathematics/science achievement exist even among students in the top quarter of the top 1% in SAT-M ability; at the graduate school level, the gender differences were even larger in the top quarter than in the bottom quarter of the top 1%. Thus, it appears that gender differences in mathematics/science achievement favoring male students are especially large among the most able students. To put these differences in perspective, however, one should note that the magnitude of the gender differences was smaller than the magnitude of the differences between the high-ability and low-ability groups.

In this study I addressed academic achievement in primarily the mathematics/science areas. Therefore, the results do not shed much light on the role that mathematical ability might play in creative accomplishments or other significant adult achievements. In fact, for those domains ability might be less important. (Our study is also limited by a less-than-perfect response rate and by the self-report nature of our high school and college data.)

In conclusion, the utility of the SAT-M for differentiating among students of extremely high ability was affirmed. Within the top 1%, there is no threshold effect for ability and its relationship to subsequent academic achievement. Therefore, this study does not contradict the practice of differentiating expectations and educational programming for students in the top 1%. Useful in this context is the well-established finding that students with superior intellectual development seem to profit from acceleration (Benbow, 1991) and from instruction that gives students considerable responsibility for organizing and interpreting information rather than from tightly structured lessons (Cronbach, 1989).

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 425–435). Washington, DC: American Council on Education.

Astin, A. W. (1965). *Who goes where to college*? Chicago: Science Research.

Benbow, C. P. (1983). Adolescence of the mathematically precocious: A five-year longitudinal study. In C. P. Benbow & J. C. Stanley (Eds.), *Academic precocity: Aspects of its development* (pp. 9–29). Baltimore: Johns Hopkins University Press.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability among the intellectually talented: Their characterization, consequences, and possible explanations. *Behavioral and Brain Sciences, 11,* 169–232.

Benbow, C. P. (1991). Meeting the needs of gifted students through use of acceleration. In M. C. Wang, M. C. Reynolds, & H. J. Walberg (Eds.), *Handbook of special education,* (Vol. 4, pp. 23–36). Elmsford, NY: Pergamon Press.

Benbow, C. P., & Arjmand, O. (1990). Predictors of high academic achievement in mathematics and science by mathematically talented students: A longitudinal study. *Journal of Educational Psychology, 82,* 430–441.

Benbow, C. P., & Minor, L. L. (1986). Mathematically talented males and females and achievement in the high school sciences. *American Educational Research Journal, 23,* 425–436.

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

Benbow, C. P., & Stanley, J. C. (1982). Consequences in high school and college of sex differences in mathematical reasoning ability. *American Educational Research Journal, 19,* 598–622.

Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: More facts. *Science, 222,* 1029–1031.

Brody, L. E., & Benbow, C. P. (1990). Effects of high school coursework and time on SAT scores. *Journal of Educational Psychology, 82,* 866–875.

Carroll, J. B., & Horn, J. L. (1981). On the scientific basis of ability testing. *American Psychologist, 36,* 1012–1020.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15–41.

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

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36,* 1067–1077.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30,* 1–14.

Cronbach, L. J. (1989). Lee J. Cronbach. In G. Lindzey (Ed.), *A history of psychology in autobiography* (Vol. 3, pp. 62–93). Stanford, CA: Stanford University Press.

Davis, J. A. (1965). Undergraduate career decisions: Correlates of academic choice. *NORC Monographs in Social Research,* No. 2.

Gottfredson, L. S., & Crouse, J. (1986). Validity versus utility of mental tests: Example of the SAT. *Journal of Vocational Behavior, 29,* 363–378.

Green, K. C. (1989). A profile of undergraduates in the sciences. *American Scientist, 77,* 475–480.

Haney, W. (1981). Validity, vaudeville, and values: A short history of social concerns over standardized testing. *American Psychologist, 36,* 1021–1034.

Jensen, A. R. (1980). Prêcis of *Bias in Mental Testing. Behavioral and Brain Sciences, 3,* 325–371.

Keating, D. P., & Stanley, J. C. (1972). Extreme measures for the exceptionally gifted in mathematics and science. *Educational Researcher, 1*(9), 3–7.

Kuhn, T. S. (1962). The structure of scientific revolutions. In *International encyclopedia of unified science* (pp. 53–272). Chicago: University of Chicago Press.

Linn, R. L. (1982a). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Widger & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies, Part III* (pp. 335-388). Washington DC: National Academy Press.

Linn, R. L. (1982b). Admissions testing on trial. *American Psychologist, 37,* 279–291.

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

Minor, L. L., & Benbow, C. P. (1986, April). Construct validity of the SAT-M: A comparative study of high school students and gifted seventh graders. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Olszewski-Kubilius, P. (1990, November). Growth in SAT scores during the high school years among gifted students. Paper presented at the Annual Convention of the National Association for Gifted Children, Little Rock, AK.

Phye, J., & Benbow, C. P. (1991). *Predictors of academic underachievement among highly gifted students.* Unpublished manuscript.

Renzulli, J. S. (1986). The three-ring conception of giftedness: A developmental model for creative productivity. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (pp. 53–92). Cambridge, England: Cambridge University Press.

Sanders, C., Benbow, C. P., & Albright, P. (1991). *Gender differences in career goals among mathematically talented students.* Unpublished manuscript.

Scarr, S. (1981). Testing *for* children: Assessment and the many determinants of intellectual competence. *American Psychologist, 36,* 1159–1166.

Stanley, J. C., & Benbow, C. P. (1986). Youths who reason exceptionally well mathematically. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (pp. 361–387). Cambridge, England: Cambridge University Press.

Walberg, H. J., Strykowski, B. F., Rovai, E., & Hung, S. S. (1984). Exceptional performance. *Review of Educational Research, 54,* 84–112.

Werts, C. E. (1967). Career changes in college. *Sociology of Education, 40*(1), 90–95.

Received December 26, 1990
Revision received July 19, 1991
Accepted August 15, 1991 ■

Instructions to Authors

Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (3rd ed.). All manuscripts must include an abstract containing a maximum of 960 characters and spaces (which is approximately 120 words) typed on a separate sheet of paper. Typing instructions (all copy must be double-spaced) and instructions on preparing tables, figures, references, metrics, and abstracts appear in the *Publication Manual.* All manuscripts are subject to editing for sexist language. Manuscript length per se is not an issue, although length should be related to the manuscript's "information value." For further details about appropriate manuscript length and content, authors are referred to the Editorials in the March 1991 issue (Vol. 83, No. 1, pp. 5–7) and the March 1992 issue (Vol. 84, No. 1, pp. 3–5) of the *Journal.*

APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy also prohibits duplicate publication, that is, publication of a manuscript that has already been published in whole or in substantial part elsewhere. Authors have an obligation to consult journal editors if there is any chance or question that the paper might not be suitable for publication in an APA journal. Authors who do not inform the editor of the prior publication history of previously published material will be engaging in unethical behavior. Also, authors of manuscripts submitted to APA journals are expected to have available their raw data throughout the editorial review process and for at least 5 years after the date of publication. Authors will be required to state in writing that they have complied with APA ethical standards in the treatment of their sample, human or animal, or to describe the details of treatment. (A copy of the APA Ethical Principles may be obtained from the Ethics Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242).

Because the *Journal* has a masked review policy, authors submitting manuscripts are requested to include with each copy of the manuscript a cover sheet that shows the title of the manuscript, the authors' names and institutional affiliations, the date the manuscript is submitted, and footnotes identifying the authors or their affiliations. The first page of the manuscript should omit the authors' names and affiliations but should include the title of the manuscript and the date it is submitted. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities.

Authors should submit manuscripts in quadruplicate. All copies should be clear, readable, and on paper of good quality. A dot matrix or unusual typeface is acceptable only if it is clear and legible. Manuscripts not meeting readability and APA *Publication Manual* specifications will be returned for repair before being reviewed. In addition to addresses and phone numbers, authors should supply electronic mail addresses and fax numbers, if available, for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. Mail manuscripts to the Editor, Joel R. Levin, Department of Educational Psychology, University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706-1796.