Are Aptitude Tests Valid for the Highly Able?

Predictions of achievement based on test-score differences among high-scoring persons are reviewed.

Henry Chauncey and Thomas L. Hilton

It is frequently said that even though aptitude tests may be valid for students of average ability, they are not valid for students of high ability—that is, they do not accurately separate the able, the abler, and the ablest from each other. Or it is said that intellectual skills are important up to a point, but beyond that point other qualities take over as determinants of quality of performance (*1*). Or it is said, especially of achievement (rather than aptitude) tests, that objective tests not only fail to distinguish but actually discriminate against the most able students, by penalizing them for their ability to see imperfections in keyed answers which average students accept without qualms as correct (*2*).

These criticisms do not represent the same position. Some readers would subscribe to one but not the others. But they are concerned with a common subject: the validity of aptitude tests for the more able individuals. We shall here examine a number of studies pertinent to this question. It is our conviction that aptitude tests *are* useful in detecting differences in the upper range of intellectual ability, and that both their advantages and their limitations for this purpose need to be better understood.

Our major concern will be with students, although the first studies to be summarized concern mature scientists. Most of these studies are based on groups who rank in ability in approximately the top 1 percent of the general population. Where the data are derived from a preselected population, as for example college students, the "high" ability sample is a larger percentage of the subpopulation in question.

Since the individuals under consideration are highly able, there is one sense in which some aptitude measures, by virtue of their design, clearly may not be valid for them. When the number and difficulty of the items, or the method of timing the test or of reporting the scores, is such that all or almost all of the highly apt students receive the highest possible score, the test will be valid in the sense of discriminating the highly apt from the average or low-scoring students, but not in the sense of discriminating among those of high aptitude. Some tests simply are not designed to provide distinctions at high levels, and when they appear invalid, the fault lies in inappropriate usage rather than in any inherent limitations in objective tests.

When suitable aptitude tests are used, the ceiling effect is not serious. For example, accumulated data on the Scholastic Aptitude Test (SAT), which was designed for college-bound high school students, indicate that only about one candidate in 16,000 achieves the maximum score of 800 on the verbal scale, and one in 1100 achieves it on the mathematics scale. These data are based on over 2 million candidates who have taken the test since January 1958 (*3*). The U.S. colleges which are the most highly selective (as judged from distributions of test scores for their entering classes) typically have less than ten students who have the maximum score on either the verbal or math parts of the SAT. A similar state of affairs holds for the Graduate Record Examinations (GRE). Out of approximately 22,000 students who took the GRE aptitude tests in the spring of 1961, 14 received scores of 800 and above (the highest score interval) on the verbal part and 263 scored in the comparable interval of the quantitative tests (*4*).

Validity for Scientific Personnel

Turning now to the evidence in regard to validity, we will first summarize several studies of able individuals in the sciences from whom aptitude-test data were obtained earlier in their lives.

The continuing research on graduate fellowship applicants conducted by the Office of Scientific Personnel of the National Research Council includes a number of studies of the test scores of the graduate students and scientists in their samples—clearly a population of high intellectual stature. Harmon (*5*) examined the high school backgrounds of 2853 graduates awarded the Ph.D. in the science fields in 1958. They constituted approximately 70 percent of recipients of science doctorates in the U.S. that year. (The 30 percent not included had attended high schools outside the U.S. or high schools which could not be identified, or did not respond to the questionnaires, or did not have the required data.) Harmon converted the scores from the different intelligence tests they had taken to a common scale—that of the Army General Classification Test (AGCT)—and then counted the number of these Ph.D.'s found at each interval of the scale. Combining these data with the known distributions of intelligence scores in the general population, he arrived at an estimate of the ratio of science Ph.D.'s to the population at each level of ability. The results are shown in Table 1. In the IQ intervals below 100, there is a total of less than one Ph.D. per 1000. In the 100-to-110 level the number is about one per thousand, and it increases in each successive level of intelligence until it reaches 189 per 1000 among individuals with IQ's of 170 and above. For our present purposes, the important observation is that even between very high IQ levels as measured in high school, the proportion of Ph.D.'s differed considerably.

In 1959 Harmon (*6*) identified 355 men who had been candidates for fel-

Dr. Chauncey is president of Educational Testing Service, Princeton, New Jersey. Dr. Hilton is a research psychologist at the same institution.

Table 1. Distribution of intelligence-test scores for total doctorate population. Adapted from Harmon (*5*, p. 682).

IQ interval (AGCT units)	Approx. general population at age 32 in 1958	Observed number of Ph.D.'s in all fields	Number of Ph.D.'s per 1000*
170 and higher	530	46	189
160–169	2,670	101	83
150–159	12,150	337	61
140–149	39,250	530	30
130–139	108,000	826	17
120–129	218,200	806	8
110–119	361,800	520	3
100–109	457,400	298	1
Below 100	1,200,000	103	0.2

* Adjusted to include estimated number of Ph.D.'s at each level among missing cases.

lowships awarded by the Atomic Energy Commission in 1948 and had taken the GRE at that time. By means of questionnaires mailed to the supervisors of the men, he obtained confidential ratings of their scientific achievement based on their scientific or technical contributions and on-the-job performance. For the 136 men not awarded fellowships none of the GRE scores was significantly related to the ratings, but for the award recipients the Quantitative Ability and Advanced Test scores were so related (Table 2). The correlations were small, accounting for only about 5 percent of the variance in the ratings. But several factors precluded high correlations. The reliability of the ratings was limited. Eleven years elapsed between the test-taking and the collection of the criterion data. Additional personal and situational factors influenced performance, and many of the recipients were preselected by means of the GRE. Many graduate schools use it for admission decisions, and it was also used as one of the bases for awarding fellowships; as a result the range of the scores and, thereby, the observed correlations were curtailed. Lastly, the Graduate Record Examinations were designed to predict graduate school performance and not necessarily on-the-job performance. In view of these factors, it is not surprising that the correlations are not higher. The results nevertheless demonstrate that measures of quantitative aptitude and achievement obtained early in the graduate school years can provide statistically significant predictions of the quality of work, approximately 10 years later, of a group of scientists of exceptional ability and promise.

D. W. Taylor (*7*) examined the relation of certain tests, including aptitude tests, to the rated performance of research scientists. Terman's Concept Mastery Test (Form B), which is primarily a test of verbal intelligence, was not significantly correlated with supervisors' ratings of creativity and productivity. This is consistent with the absence of correlations between the GRE verbal-ability score and similar ratings. A test of high-level mechanical knowledge and visualization, the Owens-Bennett Mechanical Comprehension Test (Form CC), did, however, have significant positive correlations with the criteria. The reliability of the ratings was limited (one estimate placed it at .57) and, although not reported, the range of the test scores was, no doubt, restricted.

Table 2. Validity coefficients of Graduate Record Examinations against ratings of scientific accomplishment. Adapted from Harmon (*6*, p. 8). (N = 219 AEC fellowship awardees and 136 non-awardees.)

Group	r*
Verbal ability	
Awardees	.11
Non-awardees	−.09
Total	.08
Quantitative ability	
Awardees	.22†
Non-awardees	.08
Total	.21†
Advanced achievement test	
Awardees	.28†
Non-awardees	.02
Total	.21†

* The correlation coefficients (r) here are weighted means of the r's for five groups in Harmon's study based on field: mathematics, physics, chemistry, engineering, and biology. In the original report, means and standard deviations of the test scores and the ratings are given only for the individual groups, and they are given in deciles. The means for the awardees are consistently higher than the means for the non-awardees; the standard deviations are quite uniform. Other reports by Harmon and his associates indicate that the S.D.'s for both groups are somewhat curtailed, usually ranging from 70 to 90. The test is designed to have a S.D. of 100. † Significant at the 1-percent level.

The studies of American scientists by Anne Roe (*8*, *9*) are well known. One (*9*) summarizes the test scores of 64 outstanding scientists who were selected from a sample of physical, behavioral, and social scientists nominated by panels of scientists who themselves were outstanding in the respective fields. The nominated scientists were individually given a high-level scholastic aptitude test, the Verbal-Spatial-Mathematical (VSM) test prepared by Educational Testing Service. The IQ equivalents of the scores were obtained in a separate study. These IQ's are summarized in Table 3. The study does not yield information about discrimination within the group, but it does show how the scientists stood relative to the general population. Roe's conclusion in regard to these data was:

> It is clear that the average ability of the scientists is very great. This is not surprising. On the other hand, it is surprising, and a matter of very considerable importance, that there are among the scientists a number who are not facile at the types of tasks presented by the VSM, but who have been able to make contributions of great value to society (*9*, p. 30).

Whether a score equivalent to an IQ of 121 is accurately interpreted as indicating the person is "not facile" is, in our opinion, subject to question. Even this score, which is the lowest reported by Roe, is well above the average of the general population. It also should be kept in mind that Roe's tests were individually administered, very possibly under less than ideal conditions in some cases. Scores were not obtained for five subjects, and two subjects declined to take the spatial and math tests, which suggests the possibility that others may have been reluctant, indifferent subjects.

College Scores of Successful Individuals

A study by Kallop (*10*), which concerns success in general, was of individuals who had taken the SAT between 1926 and 1939 and who later were included in *Who's Who* or *American Men of Science* or both. The mean SAT score for 232 men and women in *Who's Who* was 564 with a standard deviation of 93; the comparable figures for 49 men in

American Men of Science were 575 and 103. Precisely how high these scores are above present-day scores is impossible to say with any certainty, since the scales are not equated to current scales. (The practice of equating the scales from one form of the test to the next was not started until 1941, which was after Kallop's experiments.) Also, an unknown sampling bias is present, since test scores were found for only a fraction of the individuals cited in the volumes. We can, however, compare Kallop's subjects with other college students who took the SAT in an appropriate period—the years 1926 to 1939. When these results are analyzed by the method Harmon used, a geometric progression similar to his is obtained (*11*). As shown in Table 4, the estimated proportion of the college population represented in *Who's Who* steadily progressed with increasing test scores. The percentage with scores of 696 or higher is almost four times as great in the eminent group as in the total SAT population; the percentage with scores below 450 is less than one-third that in the total SAT population.

Data for another outstanding group of individuals—the 1963 Rhodes Scholars—were recently obtained by Pearson (*12*). SAT scores and scores on College Entrance Examination Board (CEEB) achievement tests were found for 23 of the 32 Scholars, the missing scores being largely for those from Western states. Most of the Scholars had taken the test four to five years earlier when they had applied for admission to college. The mean verbal, math, and achievement scores of 666, 708, and 674 (see Table 5) indicate that this group, who were chosen for their "literary and scholastic attainment, . . . fondness of and success in manly outdoor sports . . . , qualities of manhood . . . , and moral force of character and instincts to lead" were characterized by well-above-average test performance prior to entrance to college. The approximate percentile ranks for the mean scores when compared with the CEEB scores of high school seniors in a recent year are 95 for both the verbal and math tests. Over half of the Scholars had either verbal or math scores (or both) in the 99th percentile.

In addition, the SAT scores of 10 of the 12 students who in June 1961 were selected by *Time* magazine correspondents as being among the "top graduates of the top schools" were located in the files of Educational Testing Service. Their median verbal and math scores were 681 and 680, which are at approximately the 95th percentile when compared with the reference group described in the preceding paragraph.

Table 3. Range and median of test scores (IQ equivalents) of outstanding scientists studied by Roe (*8*, pp. 164–169).

Subtest	Score			*N**
	Highest	Median	Lowest	
Verbal	177	166	121	59
Spatial	164	137	123	57
Mathematical	194	154	128	39

* The scores of five of the scientists were not reported, presumably because they were not obtained. Two anthropologists declined to take the spatial and mathematical tests. Because the mathematics test was not sufficiently difficult for the physicists in the sample, Roe omitted them from the summary.

Table 4. Aptitude scores and proportion of eminent individuals at each level of SAT college population (*10*, p. 16).

SAT score interval	% expected (*E*) in each interval*	% in *Who's Who* group (*O*)	*O/E*
696 and above	2.6	10	3.8
629–695	7.3	15	2.1
563–628	18.9	25	1.3
500–562	21.2	25	1.2
450–499	19.1	15	.8
Below 450	30.9	10	.3

* Assuming the eminent individuals were evenly distributed throughout the SAT college population and that the distribution of scores was normal.

Table 5. Means and standard deviations of SAT and CEEB Achievement Test scores of 1963 Rhodes Scholars. Computed from Pearson (*12*).

Test	*N**	M	S.D.	Interquartile range
SAT verbal	23	665.9	74.7	595–740
SAT mathematical	23	707.8	77.2	638–761
Achievement tests	52	674.3	99.9	523–800

*Scores were located for 23 Scholars, most of whom took two or more achievement tests, which accounts for the fact that 52 achievement-test scores were included.

Table 6. Concept-Mastery Test scores of Terman's gifted subjects according to educational level achieved (*15*, p. 58).

Educational level	*N*	CMT scores	
		Mean	S.D.
Ph.D.	51	159.0	19.3
M.D.	35	143.6	23.2
LL.B.	73	149.4	20.7
Master's or equivalent	151	144.3	25.4
Graduate study, 1 or more years	122	143.0	26.9
Bachelor's degree only	263	135.7	26.6
College 1–4 years	163	128.7	29.7
No college	146	118.4	28.5

Follow-up of Gifted Children

The well-known studies by Terman and his associates (*13-15*) afford an unusual opportunity to examine whether discernible differences in intellect exist within a group of adults who as children were identified as intellectually gifted. When first given the Stanford-Binet in 1921 they stood in approximately the top 1 percent of all California school children of their age (an average of 9.7 years). In a follow-up of the gifted children in 1940, the Concept Mastery Test (Form A) was given to 954 of the original group, and in 1950 the Concept Mastery Test (Form T) was given to 1004 of the original group. Seven hundred and sixty-eight of the gifted subjects participated in both the 1940 and the 1950 follow-ups. The correlation between their scores on the two occasions was .87, even though the testing was 10 years apart and the two forms of the Concept Mastery Test (CMT) were not precisely parallel. This is evidence of remarkable reliability in the test performance of a large group of individuals of high ability.

Between the original Binet scores and the 1950 CMT scores a correlation of .29 was observed, despite the high curtailment of the distribution of Binet scores. The differences in mean CMT scores among subgroups grouped in accordance with their Binet IQ's were highly significant as tested by the F ratio ($p < .001$). In addition, equally significant differences were found between the mean scores for subjects grouped by educational level (Table 6).

In an earlier follow-up study Terman compared the characteristics of gifted subjects who had been especially successful with a less successful segment of the group. Ratings of "success" were made by three judges, largely on the basis of academic performance and professional recognition (inclusion in *Who's Who*, for example). As shown in Table 7, the mean score

of the group rated "high" in success exceeded that of the "low" group by only 5 points on the Binet IQ scale given in 1921, and by 18 points on the CMT scale given in 1940. This prompted Terman and Oden (*14*, p. 324) to conclude that differences in intelligence did not account for the contrast in accomplishment between the groups, and that "where all are so intelligent, it follows necessarily that differences in success must be due largely to nonintellectual factors." Four traits which discriminated between the groups were identified: "persistence in the accomplishment of ends," "integration towards goals," "self-confidence," and "freedom from inferiority feelings." No attempt was made to adjust for or partial out any covariance of intelligence with the traits, nor did the authors attempt to estimate statistically the contribution of the traits relative to the contribution of measured intelligence.

Table 7. Mean intelligence of two subgroups of Terman's gifted subjects (*14*, p. 323).

	Rating of "success"		Critical ratio
	High	Low	
	Binet IQ, 1922		
*N**	96	92	
Mean	155.0	150.0	3.00
S.D.	13.3	9.1	
	Concept Mastery Test, 1940		
*N**	79	116	
Mean†	112.4	94.1	4.20
S.D.	28.4	31.3	

* The *N*'s differ because it was not possible to administer tests to some subjects living in distant locations. † The Concept Mastery Test scores are not on the same scale as the Binet IQ.

High Test Scores and College Performance

There have been thousands of studies of academic prediction, but few that have focused on students with high aptitude. These few are summarized here.

The first is the extensive study of the National Merit Scholars conducted by Holland and his associates (*16*). The Merit Scholars are a group of very superior ability, having a mean Stanford-Binet IQ estimated in one study to be about 150, with a minimum of 130 (*17*). The range of their aptitude scores is very narrow; Holland and Astin (*16*) report standard deviations of approximately 52 for the SAT Verbal Test and 69 for the SAT Math Test. As far as predictive validity of the SAT scores is concerned, Holland (*18*) concluded that there is no relation between aptitude and academic performance for students of high ability. This conclusion was based on a correlation of only .09 and .11 between the verbal and math scores and the first-year grades of the men, and even lower correlations for the women. Subsequent studies of Merit Scholars have produced similar results.

There is, however, an alternative interpretation of these low correlations. The Scholars attended some 70 colleges and universities. For his criterion of academic performance, Holland pooled the first-year grade averages without adjusting for differences in standards of evaluation; he gave no more weight to an A at a highly selective and competitive college than to an A at any other college. Since it is likely that the more able Scholars attended the more competitive colleges and, on the average, received grades which were no higher, if not lower, than those of the less able Scholars who went to the less competitive colleges, it is not surprising that a high positive correlation was not obtained. In fact, a negative correlation might well have been obtained.

Support for this interpretation is given by the correlations reported for individual colleges, which, as shown in Table 8, are generally higher than the correlations for the pooled sample. This is true even though the standard deviations for the individual college samples are reported by Holland (*18*) to be generally smaller than those for the total samples. This indicates that the standard deviations for the individual colleges are approximately 40 for the Verbal Test and 60 for the Math Test, meaning that for these individual schools the Scholars represent a highly preselected group with a greatly curtailed distribution of scores.

Table 8. Correlation of grades of Holland's subjects with Scholastic Aptitude Test scores within colleges. Adapted from Holland (*18*, Table 4, p. 139).

College	*N*	*r*	
		SAT verbal	SAT math
	Boys		
California Institute of Technology	26	.09	.16
Harvard	81	.15	.07
Massachusetts Institute of Technology	73	.22	.49†
Princeton	44	.36*	.31†
Stanford	27	.18	.23
Yale	49	.29*	.14
	Girls		
Radcliffe	24	.49*	.31
Wellesley	24	.40*	.29
	Weighted average	.25	.25

* Significant at the .05 level. † Significant at the .01 level.

An additional effect resulting from stringent selection concerns the relations among the predictors rather than their lack of range as such. When selection decisions are based on the sum of the applicant's scores on several measures, it is possible to select a group in which the individual measures have a negligible positive intercorrelation or, in the extreme case, a negative correlation. Assume, for example, that the admission decision of a school is based primarily on the applicant's previous academic grades and his test scores, and that if his grades are relatively low he will be admitted provided that his test scores are high, and vice versa. From the point of view of admissions, this "multiple-selector" procedure is quite sensible. It reflects the well-founded belief that diligence may compensate for limitations in intellectual aptitudes, and that high intellectual aptitudes that have not been well employed in high school may find better expression in college. But the system inevitably lowers the correlation of each of the separate measures with the criterion—college grades—and both the separate correlations and the correlation between measures are further lowered by the exclusion of students with both low test scores and low high school grades, who would have made college records closely correlated with both measures.

Despite a severe restriction of range and probably some reduction in validity because of the multiple-selector effect just described, Whitla (*19*) showed that in recent entering classes at Harvard, SAT scores were as discriminating in the upper as in the lower ranges in predicting freshman grades. When the SAT scores and freshman grades of four classes at Harvard are pooled and lines are drawn connecting the means of the scores in each interval, the slopes of the lines are constant, even at the high end of the scale. Contrary to what might have been predicted, Whitla's regression lines, shown in Fig. 1, do not "taper off" for high scores.

Similar results were obtained from data which were part of an analysis of test scores completed at Educa-

tional Testing Service. In this study, by French (*20*), test scores and college grades were obtained from the 1960–61 freshman classes of eight colleges with especially highly selected student populations. The regression of grades on test scores for three of these schools is shown in Fig. 2. The mean first-year grades of the students in each 40-point scale-score interval and the percentage of students whose grades were in the top 20 percent of the sample from each school are given in Table 9. In general, the mean grades steadily increase with increasing test scores. A notable exception is the case of the SAT verbal scores for the freshmen at the school of science and engineering. These scores appear to be unrelated to first-year grades. This lack of correlation between verbal aptitude and grades in science, which is frequently observed, results from the fact that it is not verbal aptitude but mathematical aptitude that is important for achieving high levels of performance in science. In the women's college the distributions of grades within the high-score intervals reveal a definite skewness downward, as if they rested against a ceiling. This piling up of the grades results in the deceleration in the slope of the lines at the high end.

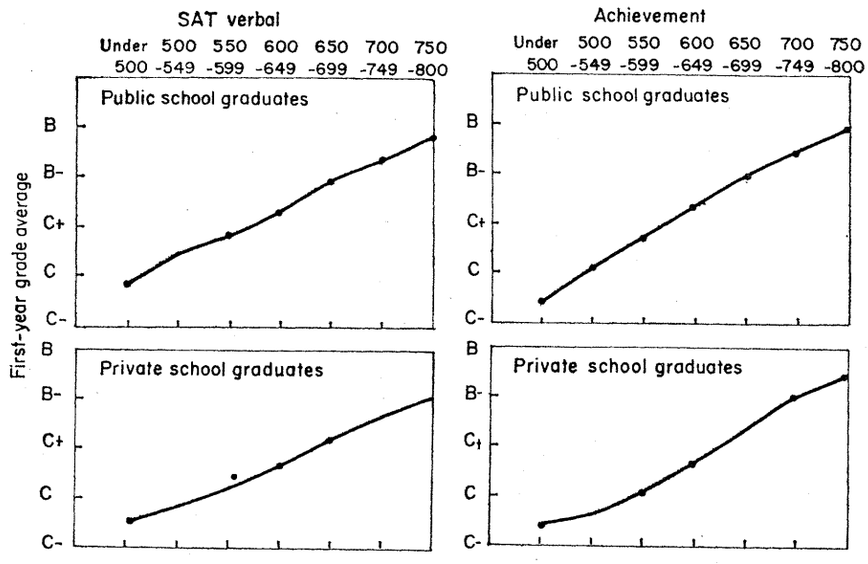

Fig. 1. Mean first-year grades of Harvard students grouped by test scores—classes of 1959, 1960, 1961, 1964. [From Whitla (*19*)]

Prediction of Performance in Graduate School

Studies of the academic performance of graduate students provide numerous examples of predictive validity, at that level, of objective measures of intellectual skill and achievement. A few of these which involve students of especially high ability will be mentioned.

In a study of the validity of the Advanced Chemistry Test of the GRE, 59 graduate students were rated on the basis of their course work and their research. Of the 14 students with scores above 790 (above the 99th percentile for college seniors) 9 were rated A, 4 were rated B, and 1 was rated C. Of those scoring between 700 and 790, none was rated A, 19 were rated B, and 4 were rated C. For those scoring below 700 the comparable frequencies were 2, 7, and 13. It is interesting that two students with scores below 700 were assigned A ratings. For our present purposes, however, the more important fact is that *in toto* there were clear differences between the

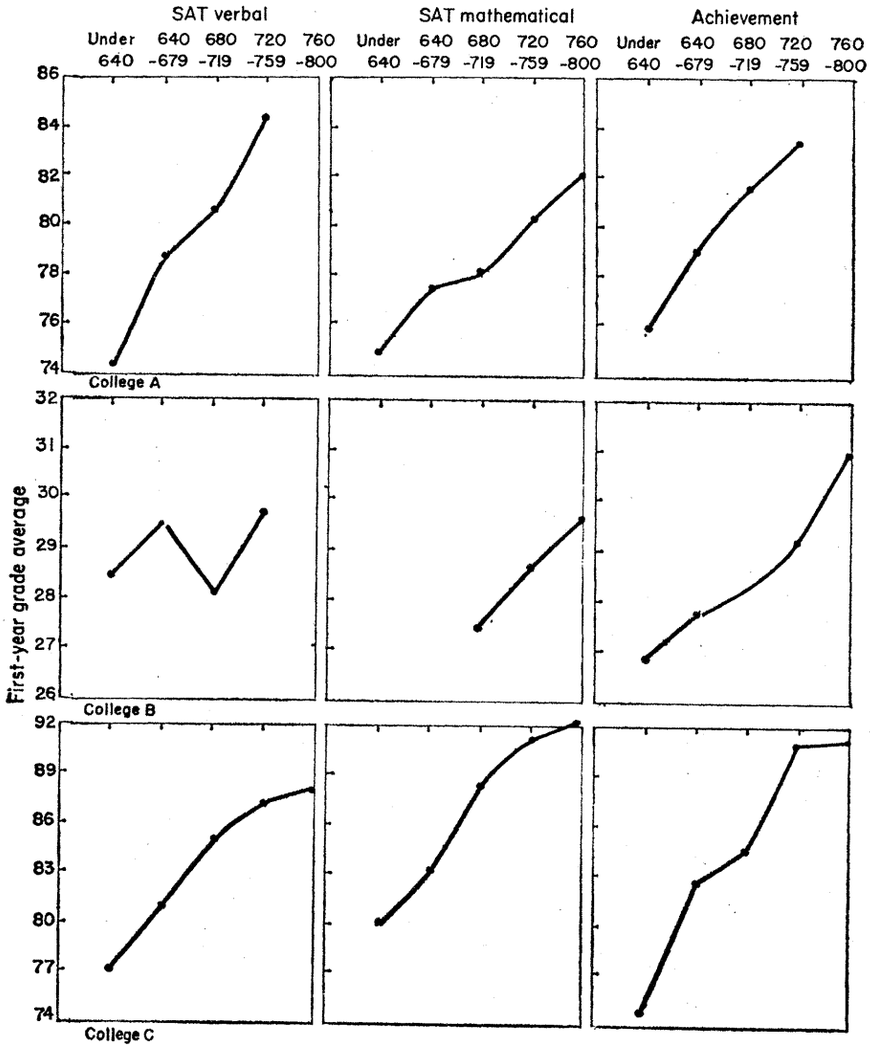

Fig. 2. Mean first-year grades of college students grouped by test scores (see Table 9). College A—male, liberal arts. College B—male, engineering and science. College C—female, liberal arts.

groups defined in terms of the test scores (*21*, p. 7).

Similarly, a study of 68 graduate students in physics at the University of Chicago showed that 92 percent of those with Advanced Physics Test scores of 600 or above received Ph.D.'s, whereas only 53 percent of those with scores below 600 were successful (*21*, p. 8).

A follow-up study of 44 graduate students admitted to the English Department of Princeton University from 1950 to 1955 showed that 80 percent of those with GRE Verbal Ability scores of 700 or higher were rated average or above average in academic performance, whereas 38 percent of those with scores below 700 were so rated. Of the 17 students with scores below 690 on either the Verbal Ability Test or the Advanced Literature Test, only five received the Ph.D., and these were each rated below average on their final oral examinations. The investigator concluded that "the predictive ability of the GRE alone is about as good as a combination of all the other information regularly supplied, including academic records and letters of recommendation even after adjustment in the light of knowledge of the institution and the recommenders" (*22*). Significantly, the positive results were obtained only from the verbal aptitude and achievement measures, not from the measure of quantitative aptitude, no doubt because of the nature of the graduate work involved, which demonstrates again the importance of having relevant aptitude measures.

Other studies, of graduate students at Florida State University, New York University, Stanford University, State University of Iowa, and Syracuse University, provide further evidence of the GRE's power to discriminate among high-level students. The obtained correlations ranged from .15 to .65, the size appearing to depend primarily on the degree of curtailment of the sample of students in question (*21*, *23*).

In a study of 119 Research Foundation Fellows in 15 different departments at Purdue University, King and Besco (*24*), using faculty ratings of overall performance as criteria, found a steady increase in the percentage rated above the median performance level as the verbal scores on the GRE Aptitude Test increased. Of those with a verbal score of at least 530, 59 percent were rated above the median; of those with at least 570, 64 percent were above the median, and of those with at least 640, 70 percent were above.

A study done for the National Research Council by Creager (*25*) showed that within a group of 2196 graduate students who applied for NSF Graduate Fellowship Awards—which represents a highly able group of individuals—the likelihood of completing graduate work and receiving the Ph.D. varied directly with their ability as measured by the GRE. The applicants were divided into five ability groups on the basis of the unweighted sum of the GRE Aptitude and Advanced Test scores. Thirty-five percent of the men in the lowest ability grouping eventually received the Ph.D., 56 percent in the next lowest, 70 percent of the next, 80 percent in the next, and 88 percent of those in the highest ability group.

The accumulated evidence from a number of studies in which the Miller Analogies Test (MAT) was used as a predictor of graduate school performance indicates that that instrument also has substantial validity for high-level students. In a study by Finch (*26*) on 112 graduate students selected for a doctoral program (and thus a curtailed sample) a biserial correlation of .56 for a pass-fail criterion was obtained. The median MAT score of the 59 successful students was 68 (out of a possible 100) and of the 53 unsuccessful candidates, 59. Similar results were obtained for the MAT at the University of Pittsburgh and the University of Tennessee, as well as other universities.

In their well-known study Kelly and Fiske (*27*) observed that MAT scores significantly predicted the rated performance of clinical psychology trainees. In addition, a preliminary follow-up showed the highest mean MAT scores for those who by that time had received the Ph.D., and then descending means for the group still in training, the group who voluntarily resigned, and the group dismissed, in that order. By the time of the first full-scale follow-up of the trainees, conducted approximately 10 years after the first data collection (*28*), the distinct differences in the means disappeared, possibly because of the varying standards of the many different programs from which the members of the original sample were graduated. If the average MAT scores were higher for the graduate schools with the higher standards and, in addition, with higher student-withdrawal rates, then a negative correlation between pooled MAT scores and graduate school completion could result, in the same way that Holland might have obtained negative correlations.

Perhaps the most direct effort to answer the questions raised in this article was by Angoff and Huddleston (*29*). In contrast to the foregoing studies, they approached the question of test reliability and validity for students of high ability by determining whether a scholastic aptitude test especially tailored for such students would show greater reliability and validity than a conventional test. They devised a test with a narrow dispersion of item difficulties, the average level being high, and gave it along with a conventional test to 429 college students of high ability. When the results from the two tests were compared, it was found that the narrow-range test showed reliability coefficients .03 or .04 higher than the broad-range (conventional) test, and that for the prediction of academic grades the validity coefficients were .01 or .02 higher. These differences were judged to be not sufficient to justify the use of separate SAT's for students at different levels of ability.

Discussion and Conclusion

From the studies that have been summarized, it is clear that if the charge that aptitude tests do not discriminate for high-ability students is valid at all, it is valid only in some special sense.

The first possibility considered was that objective tests do not discriminate among highly able students because all such students receive the maximum score. But in fact the ceilings of the tests can be sufficiently high to provide room for all to demonstrate their ability. In principle, there is no limit to the ceiling; by including enough very difficult items in a test, the ceiling can be raised above the level of all who take it. In practice, however, this is both uneconomical and unnecessary.

The evidence is that substantial validity for high-level subjects is possible when (i) the instrument is appropriate for the sample and the predictive task at hand, (ii) a relevant criterion measure is available, and (iii) the correlation is not reduced by the effects of (a) a low ceiling on the criterion, (b) a very narrow range of ability in the group tested, or (c) what was referred to as the "multiple-selector" system. In summary, there is no evidence that

Table 9. Mean first-year grades, and percentages in top fifth of class, of college students grouped by test scores.

Score interval	SAT verbal			SAT mathematical			Achievement		
	N*	Mean grade	Percentage in top fifth	N*	Mean grade	Percentage in top fifth	N*	Mean grade	Percentage in top fifth
				College A (male, liberal arts)					
760–800	0			17	82.2	47.0	0		
720–759	18	84.5	72.2	43	80.2	41.9	20	83.6	60.0
680–719	52	80.6	32.7	52	78.0	17.3	26	81.6	42.3
640–679	60	78.9	16.7	40	77.8	12.5	44	79.5	27.3
639 and below	92	74.3	4.4	70	75.1	5.7	132	75.6	6.8
				College B (male, engineering and scientific)					
760–800	0			69	29.7	21.7	30	31.1	36.7
720–759	20	29.8	10.0	50	28.7	22.0	51	29.1	19.6
680–719	37	28.1	18.9	28	27.4	10.7	30	28.2	16.7
640–679	47	29.6	29.8	0			21	27.8	9.5
639 and below	43	28.5	13.9	0			15	26.8	6.7
				College C (female, liberal arts)					
760–800	14	87.9	42.8	14	91.9	50.0	24	91.0	41.7
720–759	64	87.1	29.7	19	91.6	36.8	48	91.0	31.2
680–719	90	85.2	20.0	55	88.4	27.3	62	84.5	17.7
640–679	45	80.6	6.7	53	82.4	15.1	58	82.8	13.8
639 and below	25	76.9	8.0	97	80.0	10.3	46	74.6	8.7

* Whenever the end intervals contained ten cases or fewer, those cases were combined with those in the adjacent interval.

aptitude tests are less valid for individuals of high ability than for individuals of average ability.

The research of Harmon, Roe, Kallop, and others showed that various aptitude measures administered early in the lives of the individuals involved significantly predicted certain aspects of their later careers. In addition, research on students of high ability demonstrated that quality of performance increases with increasing test scores even at very high levels. Holland's work (*16, 18*) would seem to be a demonstration to the contrary, but his negative findings can be accounted for, as Holland himself points out, by the way in which the data were pooled.

Holland's findings also illustrate in a particularly vivid fashion the problem of curtailment of range. When scholastic aptitude tests are used in two successive screenings of a group of candidates, the resulting group is unusually homogeneous in this respect and, thereby, difficult to discriminate among. It is not surprising that Holland had to turn to nonintellectual measures in order to find differences among the students in his samples.

Whether there is evidence in regard to the criticism that objective tests discriminate *against* highly able students is not answered. If there is such discrimination and it is extreme, then the studies that have been examined are irrelevant: the very students who would have provided pertinent data would have been excluded from consideration, since most of the studies reported focused on students who had scored high enough on one objective test or another to qualify for inclusion in some sample, for example, a college entering class. If the discrimination is not so extreme (which seems likely), there is still the possibility that only a small group of exceedingly able students is discriminated against and that the lack of validity for these is not detected when large samples are observed. In none of the studies were perfect correlations reported. The possibility that some of the departures from prediction resulted from the alleged discrimination cannot be completely discounted.

Several of the studies described demonstrated that the predictor aptitude must be relevant to the academic performance in question—a truism, but one which is frequently lost sight of. A measure of verbal aptitude may not be an appropriate predictor of the quality of work in science of students high in mathematical aptitude (even though it may be a useful predictor for students with low aptitude, since whether such students have enough ability to communicate effectively may be at stake, a question which does not ordinarily exist for students high in quantitative aptitude).

A remaining question concerns how much of the variance in the intellectual performance of highly able students is explained by aptitude scores. Do the scores account for all or nearly all of the differences in performance? Although we know of few rigorous efforts to answer this question, it is clear that the answer is no. Most of the correlations reported are within a range of .40 to .70, which means that typically less than half of the variance in the criteria is accounted for by the various predictors. Also, we know that numerous studies have shown other measures to be related to intellectual performance. In the Taylor study (*7*) of research scientists, for example, a special test of how to plan and conduct research was more highly correlated with performance ratings than the aptitude measures were. Other studies have reported noncognitive correlates of high-level performance. Persistence, need for achievement, ego involvement, originality, high self-evaluation, intellectuality, and socialization are only a few such correlates. Unfortunately, in most of the studies there have been no attempts to control or partial out the contribution of differences in intelligence to differences in performance, but even if such steps were taken, it is highly likely that significant noncognitive effects would remain. Otherwise, the whole structure of our conception of human functioning would need drastic reformulation.

In conclusion, there is ample evidence that aptitude tests can discriminate reliably among students high in ability, and also can validly predict relevant characteristics of the performance of such individuals. This is not to say that they account for all or nearly all of the variance in high-level performance. But from the available research it appears likely that aptitude measures account for fully as much variance as do other single measures.

References and Notes

1. D. C. McClelland, A. L. Baldwin, U. Bronfenbrenner, F. L. Strodtbeck, *Talent and Society* (Van Nostrand, Princeton, N.J., 1958), p. 13.
2. B. Hoffmann, *The Tyranny of Testing* (Crowell-Collier, New York, 1962), pp. 99–101.
3. W. E. Coffman, "The Scholastic Aptitude Test—a forward look, 1963," unpublished.
4. Unpublished test analysis conducted at ETS.
5. L. R. Harmon, *Science* **133**, 679 (1961).
6. ———, "Validation of Fellowship Selection Instruments against a Provisional Criterion of Scientific Accomplishment," *Res. Fellowship Selection Techniques, Office Sci. Personnel, Natl. Acad. Sci., Natl. Res. Council Tech. Rept. no. 15* (1959).
7. D. W. Taylor, "Variables related to creativity and productivity among men in two research laboratories," in *Scientific Creativity*, C. W. Taylor and F. Barron, Eds. (Wiley, New York, 1963), pp. 228–250.
8. A. Roe, *The Making of a Scientist* (Dodd, Mead, New York, 1952).
9. ———, "A Psychological Study of Eminent Psychologists and Anthropologists, and a Comparison with Biological and Physical Scientists," *Psychol. Monographs: Gen. Appl.* **67**, 1–55 (1953).
10. J. W. Kallop, "A Study of Scholastic Aptitude Test and Eminence," thesis, Princeton Univ., 1951.
11. It should be kept in mind that Harmon's results were based on the general population and Kallop's were on a selected college population. To obtain a more direct comparison of the two studies we can estimate the AGCT score which is comparable to Kallop's 75th percentile SAT score. On the basis of a table provided by C. C. Brigham [*A Study of Error* (College Entrance Examination Board, New York, 1932), p. 336] and a second table provided by the Staff, Personnel Research Section, Adjutant General's Office [*J. Ed. Psychol.* **38**, 385 (1947)], the SAT score of 629 is estimated to be comparable to an AGCT score of 150. Kallop observed that 25 percent of his cases fell above this score; Harmon observed 14 percent. In addition, by means of the same estimates, we can say that a person in the top 1 percent of the general population has approximately 25 times as much chance of being in *Who's Who* as has the population in general.
12. R. Pearson, "On the Use of Multiple-Choice Tests in College Admission" (College Entrance Examination Board, New York, 1963).
13. L. M. Terman and M. A. Merrill, *Measuring Intelligence* (Houghton Mifflin, Boston, Mass., 1937).
14. L. M. Terman and M. H. Oden, *The Gifted Child Grows Up* (Stanford Univ. Press, Stanford, Calif., 1947).
15. ———, "The Gifted Group at Mid-life," *Genetic Studies of Genius, V* (Stanford Univ. Press, Stanford, Calif., 1959).
16. J. L. Holland and A. W. Astin, *J. Educ. Psychol.* **53**, 132 (1962).
17. J. R. Warren and P. A. Heist, *Science* **132**, 330 (1960).
18. J. L. Holland, *J. Educ. Psychol.* **50**, 135 (1959).
19. D. Whitla, "PRL (Predicted Rank List) Revisions," unpublished Harvard College Study, 1962.
20. J. W. French, "The Validity of New Tests for the Performance of College Students with High-Level Aptitude," *Res. Bull. 63-7* (Educational Testing Service, Princeton, N.J., 1963).
21. G. V. Lannholm, "Abstracts of Selected Studies on Relationship between Scores on the Graduate Record Examinations and Graduate School Performance," *Graduate Record Examinations Spec. Rept. 60–3* (Educational Testing Service, Princeton, N.J., 1960).
22. J. Thorpe, unpublished report, Princeton University Graduate School, 1959.
23. G. V. Lannholm and W. B. Schrader, *Predicting Graduate School Success* (Educational Testing Service, Princeton, N.J., 1951).
24. D. G. King and R. O. Besco, "The Graduate Record Examination as a Selection Device for Purdue Research Foundation Graduate Research Fellows," unpublished report, Graduate School, Purdue University, 1960.
25. J. A. Creager, "A Study of Graduate Fellowship Applicants in Terms of Ph.D. Attainment," *Res. Fellowship Selection Techniques, Office Sci. Personnel, Natl. Acad. Sci., Natl. Res. Council Tech. Rept. No. 18* (1961).
26. See W. S. Miller, *Manual, Miller Analogies Test* (Psychological Corp., New York, 1952), p. 9.
27. E. L. Kelly and D. W. Fiske, *The Prediction of Performance in Clinical Psychology* (Univ. of Michigan Press, Ann Arbor, 1951).
28. E. L. Kelly and L. R. Goldberg, "Correlates of Later Performance and Specialization in Psychology," *Psychol. Monographs: Gen. Appl.* **73**, no. 12 (1959).
29. W. H. Angoff and E. M. Huddleston, "The Multi-level Experiment," *Statist. Rept. 58-21* (Educational Testing Service, Princeton, N.J., 1958).

30. We are indebted to John W. French and William B. Schrader for their valuable criticisms of a draft of the manuscript.

News and Comment

Antarctica: Congressional Urge for Tidy Research Administration Manifests Itself in New Proposal

One of the most persistent themes in government relations with science is the Congress's inclination to tidy up the administrative structure of research and the executive's desire to protect what Jerome B. Wiesner once referred to as the "anarchy" of research.

Thus, Congress has from time to time toyed with proposals for a Cabinet-level Department of Science, to encompass most or all of the federal government's research activities. Strenuous opposition from the executive's science advisers has helped prevent these proposals from acquiring the necessary votes. But now and then a less ambitious plan for administrative tidiness manages to develop significant support. Such was the case, for example, in 1962, when, on the initiative of the Senate Commerce Committee, Congress passed a bill giving the White House Office of Science and Technology (OST) responsibility for coordinating the oceanographic research of the 24 federal agencies operating in that field. Folklore says that government offices inexorably quest for greater power, but OST didn't want to take on oceanography or any other operational responsibilities. President Kennedy pocket-vetoed the bill—one of the nine vetoes of public bills during his presidency—and the coordination of oceanography remained the responsibility of an interagency committee. Congress obviously didn't agree, but the administration felt that the interagency committee offered the virtues of coordination and decentralization. No effort was made to override the veto.

The latest example of congressional interest in tidiness of research administration concerns another interagency effort, the Antarctic research program, in which the Defense Department, through the Navy, handles logistics, the National Science Foundation is responsible for research, and the State Department provides coordination and guidance under the 14-nation Antarctic Treaty. Some Navy officials have complained about what they consider to be poorly defined lines of authority in this three-agency arrangement, but there seems to be fairly general satisfaction, among researchers and Defense Department officials, with the way things have worked out. Nevertheless, an effort is now under way in the House to place the Antarctic program under what would be called the Richard E. Byrd Antarctic Commission. This would consist of a director, two deputy directors, and an 11-member consulting board of governors, all of which, as things go in the federal government, is a lot of brass for a program that is budgeted for about $27 million a year.

In an apparent effort to take some steam out of this proposal, the administration recently set up a three-member Antarctic Policy Group, consisting of