DOCUMENT RESUME

ED 075 971 EC 051 776

AUTHOR Keating, Daniel P.; Stanley, Julian C.
TITLE Discovering Quantitative Precocity.
PUB DATE [73]
NOTE 11p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Exceptional Child Education; *Gifted; *Identification; Junior High School Students; *Mathematics; Standardized Tests; Testing; *Test Interpretation

ABSTRACT

Differentiation among gifted junior high students who score at the 98th or 99th percentile on in-grade achievement tests of quantitative abilities can be accomplished by administering a test normally given to older students such as the College Entrance Examination Board's Scholastic Aptitude Test-Mathematical (SAT-M). SAT-M scores of 396 7th, 8th, and accelerated 9th grade students show a wide range of abilities among students scoring at the ceiling of in-grade tests. The rationale for discrimination among gifted students should be individualized educational planning which may include college courses and early college admission for the gifted junior high student who also scores high on the SAT-M. Because younger students may have to make greater use of reasoning abilities to solve problems to which older students apply learned formulas, this reasoning ability can be predicitive of success in advanced courses of new material. Gifted junior high students have been placed in college courses with unbroken success. (See EC 051 774 for a related document). (DB)

Discovering Quantitative Precocity

Daniel P. Keating and Julian C. Stanley

The Johns Hopkins University

Identification of the gifted students in a school population is logically the first step in any educational program for such a group. The criterion which is set for inclusion in a "gifted" group depends on the aims and goals of the particular program. Among the criteria which have been used in the past are: above some specific IQ score, as in the Terman Gifted Study (1925, et seq.), which set a 145 IQ minimum; the top 10% on measures of general scholastic aptitude; or, more recently, above some criterion score on a test of "creativity" (more correctly, "ideational fluency," according to Wallach [1971]).

In the Study of Mathematically and Scientifically Precocious Youth (SM&SPY) at The Johns Hopkins University, the initial goal is to select the ablest mathematical reasoners of junior high school age from among an already able group. Two mathematics competitions have been held for seventh, eighth, and accelerated ninth graders, the first in March, 1972, and the second in January, 1973. No official screening was done for the first competition, but it was recommended that the student have at least a 95th percentile (national norms) on a standardized test of arithmetic reasoning. The results of the first competition led to this restriction for the following year: a 98th or 99th percentile score (national norms) on a standardized test of arithmetic reasoning.

The primary test for both competitions was the College Entrance Examination Board's (CEEB) Scholastic Aptitude Test-Mathematical (SAT-M), which is normally given to high school juniors or seniors seeking college admission.

This test would be extremely difficult for most 12-14 year olds, of course, but it was chosen mainly for that reason. If one has a highly able group to begin with, further in-grade testing is unlikely to separate them efficiently or reliably. Before proceeding on to the detailed rationale for the use of such high-level tests for these groups, let us examine briefly the results of the testing competitions.

In the 1972 competition, 396 students took SAT-M. Detailed results have been reported elsewhere (Keating & Stanley, 1972a, b; Keating 1972), but several major findings bear repetition. A score of 540 on SAT-M is about the 78th percentile of male high school seniors; 89 contestants, all 7th, 8th, or accelerated 9th graders, scored that high or higher. A score of 620 is the 91st percentile, and 41 contestants scored at least that high. From the top of the distribution down to a score of 660, the scores and frequencies were: 790 - 1; 780 - 1; 740 - 2; 730 - 1; 710 - 2; 690 - 2; 680 - 5; 670 - 3; and 660 - 5. These are scores which the great majority of high school seniors never achieve. For example, the average freshman at The Johns Hopkins University scored 657. The complete grouped frequency distribution of scores for the first competition are shown in Table 1.

- - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - -

The full results of the second competiton are not yet available, but partial results indicate that the level of talent found this year exceeds that discovered last year. One boy, an accelerated 9th grader who was 13 years 0 months old at the time of the testing, earned an extrapolated score of 807 on SAT-M (i.e., one raw score point above the minimum required for a scaled score of 800). Another accelerated 9th grade boy scored 770 on SAT-M and 710 on SAT-Verbal.

Table 1: Grouped frequency distribution by grade and sex on SAT-M and M-I of 396 students participating in mathematics contest

Score	7th Grade				8th Grade				9th Grade[1]			
	SAT-M		M-I		SAT-M		M-I		SAT-M		M-I	
	B[2]	G[3]	B	G	B	G	B	G	B	G	B	G
760-800	0	0	0	0	1	0	0	0	1	0	1	0
710-750	2	0	1	0	2	0	2	0	1	0	0	0
660-700	2	0	1	0	13	0	1	0	0	0	0	0
610-650	3	0	0	0	17	0	6	0	1	0	1	0
560-600	8	3	0	0	19	11	8	0	1	0	1	0
510-550	11	8	6	3	21	23	15	9	0	1	1	0
460-500	20	16	9	5	19	20	27	17	0	0	0	1
410-450	14	17	27	22	17	14	33	34	0	0	0	0
360-400	18	22	31	36	13	14	32	29	0	0	0	0
310-350	7	7	11	10	6	7	4	6	0	0	0	0
260-300	3	2	4	1	1	3	1	0	0	0	0	0
210-250	2	2	0	0	0	3	0	0	0	0	0	0
N	90	77	90	77	129	95	129	95	4	1	4	1
Median	457	420	394	388	534	470	442	421	690	510	590	500
Mean	460	423	408	396	523	457	458	426	690	510	618	500
S.d.	104	75	71	49	105	88	81	50	88	---	110	---

[1] Accelerated 9th graders were eligible, i.e., those not yet 14 years old at time of testing (March 4, 1972).

[2] Boys

[3] Girls

The level of mathematical reasoning ability evidenced by such scores is surprising to individuals accustomed to in-grade comparisons of gifted youngsters, even exceptionally gifted ones. When one is used to dealing with the full range of scholastic aptitude (as measured by standardized tests); from quite low to quite high, as is true of most school personnel, the feeling that anything above the 98th or 99th percentile "doesn't really make any difference" is understandable but unjustifiable. If the goal of the schools, like that of SM&SPY, is to provide the best possible educational alternatives to each individual, rather than to specifiable sub-groups, then the distinctions even at this high level can be as important as in-grade testing.

There are several reasons for the use of high-level tests with gifted youngsters in educational as well as research settings. The first of these is that the 98th or 99th percentile students on in-grade tests are likely to be as different from each other as the group is different from 75th percentile youngsters. It is not intuitively obvious, but there is (theoretically) more range above the 99th percentile than there is from the 1st to the 99th percentile. A conversion to z-scores illustrates the point. In a normal distribution, the 1st percentile has a z-score equivalent of approximately -2.33. The 99th percentile is of course +2.33. Thus, the total z-score range is $2.33-(-2.33) = 4.66$. The region beyond the 99th percentile, however, extends to $+\infty$. An example of this is available from another study of sixth graders nominated as gifted who took the Academic Promise Tests (APT) for 6th-9th grade. A 99%ile score on the numerical subscale for 6th graders was 40 or greater on a 60-item test. One student got 58 right, another 40. Both were 99th percentile on in-grade norms, but the same raw score difference of 18 between 40 and 22 was the difference between a 99th and a 65th percentile score.

Also, the 99.9th %ile score is as different from the 99.0th %ile score in standard score terms as the latter is from the 94th %ile score.

Tests in-grade, however, rarely make these distinctions. In some cases the distinctions could be made on the basis of standardized tests in-grade, but more often such tests lack sufficient ceiling to separate these individuals accurately and reliably. The problem is especially acute if, as in some cases, the 98th and 99th %ile cutoff levels are only a few points below the ceiling of the test. In such a case errors of measurement are likely to lead to misclassification of a sizable number of high scorers.

Consider a (hypothetical) 60-item test of mathematical reasoning which is standardized for a large group whose mean score on the test is 30. The reliability coefficient for the test is .81 and the standard error of measurement is 5.00. For an individual who scores at the ceiling, i.e., 60 out of 60 right, the 95% confidence interval around his estimated "true score" is 62.24 to 46.36 (Stanley, 1971). The "true score" of such an individual may, however, be far above that, but this test, because of lack of ceiling, can make no "true score" estimate higher than 62.24, even at a 95% confidence level. (The upper limit of the confidence interval at the 99% level is still only 63.74). If the reliability coefficient of the test were higher and the standard error of measurement lower, the confidence intervals would of course be even smaller. In the extreme case of perfect reliability the one and only "true score" estimate would 60. The point of the example is that for an individual whose "true score" lies above the ceiling of a particular test, that test is both practically and theoretically incapable of estimating his actual ability.

Thus, within the confines of classical test theory, there are two major problems with using most in-grade tests with exceptionally gifted students, both connected to lack of ceiling: 1) the tests can give no indication of how such students are different from each other; 2) the tests can not give an accurate estimate of an individual's ability if his true level is above the ceiling of the test.

Both of these objections can be overcome by using adequately difficult tests. But before arguing this in detail, the question of the purpose of such accurate measurement should be addressed. It is especially important in that some teachers and parents have suggested that using such difficult tests may, in fact, be harmful in some way to young children. It may be traumatic for the average child, but these exceptionally able students seem to relish the challenge.

There would be no purpose to using these tests if we were to collect the scores, note them with some curiosity, and pass on to other endeavors. But this is not the case, at least in SM&SPY. The purpose is precisely to assist the student in planning his education best, and these plans will be quite different for the student who scores 800 on SAT-M when 13 years 0 months old, and for another student the same age who scores 540, even though both are at the 99th %ile on in-grade tests. Educational facilitation for the first student will certainly include released time and summer college courses while in high school, and probably early admission into college as well (Fox, 1972).

This raises the second major reason for using high level tests with such students, and it is a primarily empirical rather than theoretical one: the more difficult tests have predictive validity for the kinds of challenging experiences that facilitating their education will present.

Testifying to the fact that these tests have good predictive validity is a series of (thus far) unbroken successes in placing young students found through these competitions in college courses on released time or during the summer. None of the students has received a grade less than B, and the majority have received A's. Prior to the official start of SM&SPY, two 8th grade students were admitted to Johns Hopkins as full-time students before the age of 14. They had scores on CEEB aptitude and achievement tests superior to those of most entering freshmen, and their success at Hopkins has been discussed in case studies elsewhere (Keating & Stanley, 1972a; Stanley, 1972). A third student (who was not found through the testing competitions) was admitted in 1972 at age 16 after the 10th grade on the basis of excellent scores on college level tests and SM&SPY's recommendation. He compiled a 4.0 (straight A) average in his first semester.

Related to these considerations is the fact that we are interested in quantitative precocity, not just quantitative aptitude. Precocity, as the term is used here, means arriving at some stage of development earlier than expected, such that the individual's current state of development is more like someone much older. In this context, "quantitative precocity" means having attained a stage of cognitive development in the quantitative area more like the developmental stage of someone several years older than the norm for age-mates.

The simplest way to discover this is to assess it directly. Thus, to find out which of a given group of able 12-14 year olds has attained a level of quantitative reasoning ability more like able high school seniors, one need only give them the same test of mathematical reasoning one would give to a group of high school seniors. The excellent and frequently used test for this purpose is SAT-M.

This is not to imply that a score of 680 for a 12-14 year old on SAT-M necessarily means the same thing in terms of quantitative reasoning ability as the 680 earned by the high school senior. In fact, the younger student is probably being required to use more of his reasoning ability, since some of the formulas and identities which the older student has learned in high school are not available to the younger student, and he consequently must figure them out by using a higher level process.

Before we elaborate this distinction, it should be noted that this element probably "biases" the predictive validity positively if the criterion is "success in learning new material in introductory courses" (as reflected by the grade in the course), for almost certainly the reasoning element will be more important in such a situation than "amount of knowledge previously accumulated" will be.

The different meaning and interpretation of the test score depending on who the test-taker is raises an important point. As Anastasi (1972) has suggested, the test item is not an unchanging and objectively determinable "stimulus" across all groups. The sample of behavior which each item seeks to evaluate is a complex interaction of the item and the individual, including his background and experience, and the way in which he reacts to the particular item.

Two examples serve to illustrate this point with regard to mathematics items. One such item on a college level test involved the division of one fraction by another. For the college population the test was constructed for, the item was appropriately placed in the mathematical computation section of the test. When the item was presented to an 11 year 9 month old boy, however, it was a "different" item. He had not then learned the rule for division by a fraction (i.e., invert and multiply), but got the item right nonetheless. It was clear from his explanation afterward that he had used excellent mathematical reasoning to complete the item correctly.

The second example is the manner in which a 9 year old boy went about solving a problem involving the area of a triangle. He had not learned the formula (1/2 x base x height) in school, of course, being only a 4th grader. But, as he explained later, he knew that the area of a rectangle was "base x height." He recognized the hypoteneuse of the triangle as being the diagonal of a(n imaginary) rectangle. Thus, he reasoned, if he calculated the area of the rectangle and took half of that area, he would have the area of the triangle (i.e., base x height x 1/2).

These "clinical" item analyses need to be investigated by more conventional statistical methods. A full item analysis of the SAT of the two testing competition groups is planned, and the results are to be compared to item analyses of the SAT conducted by Educational Testing Service (ETS) in its regular administrations. The comparisons, if the above discussion is accurate in its conclusions, should prove most enlightening.

The techniques described in this paper for the discovery of quantitative precocity are of course applicable in other areas. The general finding that tests which are adequately scaled for a population defined by age and grade may not be the most useful for those near the top of the scale. This difficulty can be overcome by administering a higher level test to a select sub-population on the basis of the first (in-grade) test. (It should not be administered to the whole population, of course, because it would be both useless and discouraging for all but the top few). This simple procedure, which is, as we have shown, both theoretically and practically justifiable is one often overlooked, even by those who most need the information to assist and counsel the student.

References

Anastasi, A. Discovering and nurturing precocious talent in mathematics and physical sciences: Discussion. Paper presented at the Annual Meeting of the American Association for the Advancement of Science (AAAS), Washington, D.C., December, 1972.

Fox, L. H. Facilitating educational development of mathematically precocious youth. Paper presented at AAAS Annual Meeting, Washington, D.C., December, 1972.

Keating, D. P. The study of mathematically precocious youth. Paper presented at AAAS Annual Meeting, Washington, D.C., December, 1972.

Keating, D. P., & Stanley, J. C. From eighth grade to selective college in one jump: Two case studies in radical acceleration. Paper presented at AERA Annual Meeting, Chicago, March, 1972 [a].

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

Stanley, J. C. Intellectual precocity. Paper presented at AAAS Annual Meeting, Washington, D.C., December, 1972.

Terman, L. M., et al. Genetic studies of genius. Vols. I-V. Stanford, Calif.: Stanford Univ. Press, 1925-1957.

Wallach, M. A. The intelligence / creativity distinction. New York: General Learning Press, 1971.

Stanley, J. C. Reliability. In Thorndike, R. L. (ed.), Educational Measurement. Washington, D. C., American Council on Education, 1971.