DOCUMENT RESUME

ED 203 619 EC 133 304

AUTHOR Rudner, Lawrence M.
TITLE A Realistic Look at the Stanford Achievement Test: What Does it Mean?
INSTITUTION Model Secondary School for the Deaf, Washington, D.C. Office of Research and Evaluation.
SPONS AGENCY Department of Health, Education, and Welfare, Washington, D.C.
PUB DATE Dec 77
NOTE 21p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Evaluation Methods; *Hearing Impairments; Secondary Education; Standardized Tests; Testing; *Test Interpretation; *Test Norms
IDENTIFIERS *Stanford Achievement Tests for Hearing Impaired

ABSTRACT

The use of Stanford Achievement Tests (SATs) with hearing impaired students is considered. Background of the test is given, and the SAT-HI (SAT for the Hearing Impaired) is discussed in terms of its content, item wording, and norms. Among six specific recommendations made are to use the SAT-HI because it minimizes floor and ceiling effects, standardizes administration and provides norms based on hearing impaired students; examine the content of the SAT-HI to determine its relevance to the curriculum; do not interpret the scores of hearing impaired students as precise estimates of ability; and use raw scores, when applicable, or scaled scores to assess student growth. (CL)

1st Friday
A Research Seminar Sponsored by the Division of Research and Evaluation
Model Secondary School for the Deaf, Gallaudet College, Kendall Green, Washington, D.C. 20002

A Realistic Look at the Stanford Achievement Test: What does it mean?

Lawrence M. Rudner

December 2, 1977

This publication was sponsored by the Model Secondary School for the Deaf which is authorized and funded by the U.S. Department of Health, Education, and Welfare, through Public Law 89-694. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgement in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official U.S. Department of Health, Education, and Welfare position or policy.

A Realistic Look at the Stanford Achievement Test: What does it mean?

Like other standardized and normed collections of achievement test batteries, the Stanford Achievement Test (SAT) is often used to evaluate the growth and performance of individual students and groups of students. For students in regular public elementary and junior high school classrooms, the SAT has proven to be quite meritorious. The reliabilities of the various subtests are sufficiently high and the content of the different level examinations adequately parallels the curriculum taught in the corresponding grade levels.

In using SAT or another achievement test designed for regular classroom students with a special population, such as with hearing impaired adolescents, the test user assumes responsibility for the measure being adequate for the tasks and for the proper use of test results. In this paper, several characteristics of achievement tests in general, and the SAT in particular, are examined. Specifically, test content, item wording, and norms are discussed as they relate to evaluating hearing impaired students.

Background

Since the first edition in 1923, the SAT has not been a single test, but rather a collection of achievement test batteries. The 1973 edition contains six carefully developed batteries, each battery consisting of several subtests to assess skills in areas such as reading comprehension, mathematics concepts, mathematics computation, science, and social studies. The Primary Level I Battery is designed for public school students in grades 1.5 to 2.4; Primary Level II, grades 2.5 to 3.4; Primary Level III, grades 3.5

to 4.4; Intermediate Level I, grades 4.5 to 5.4; Intermediate Level II, grades 5.5 to 6.9; and Advanced, grades 7.0 to 9.5.

In the development of the SAT, objectives matching the core curriculum of each grade in elementary schools were developed after a careful survey of existing curricula. These objectives served as a guide in the development of the publishers' textbook series and in writing items for the SAT (see SAT Manual, Part V, p. 12-13).

To insure that the item wording was as appropriate as possible, twice as many items as needed were developed and administered to an item tryout sample. This pre-test sample was selected so as to closely match the United States population in terms of percentages of students by community size, percentage of students by geographic region, median family income, median years of parental schooling, percentage of Black Americans, and on other variables (see Stanford Research Report #3). Based on the item tryout, item statistics were computed and the best items were retained for the final version. In the Primary II battery, for example, 2,565 items were piloted and 1,326 items were retained for the three final forms.

After items were selected for the final forms, norms, such as grade equivalent scores and percentiles, were developed. These norms provide a means for comparing the performance of one student or group of students with that of some particular reference group. While several reference groups are possible, the SAT, like most other standardized achievement test batteries, uses a representative sample of the United States school children population as its benchmark.

In context, these norms can have meaning for most school systems. This is true because the norms describe one reality, i.e., the typical performance of the country's students on these items. However, the context is restrictive. It assumes adequate measurement, relevance of the items, and appropriate interpretation. These limitations are extremely important when a test is used as an assessment device in a program for hearing impaired students.

A 1972 national survey by the Office of Demographic Studies showed that the SAT was the most popular standardized achievement test among educators of the deaf. Of the 29,023 hearing impaired students to receive any standardized achievement test during the 1972-73 school year, approximately 77% or 22,292 students would be taking the SAT (Buchanan, 1973). Because of its popularity, the Office of Demographic Studies decided to facilitate proper use by compiling a special edition of the 1973 SAT, the Stanford Achievement Test-Hearing Impaired version (SAT-HI). While the original items and subtests were retained, the level of the SAT-HI in which the subtests appeared was changed. Thus, for example, the Level II battery of the SAT-HI contains the Vocabulary and Communication Comprehension subtests of the SAT Primary Level I battery and the Mathematics Computation and Spelling subtests of the SAT Primary Level III battery. The other subtests are those which appear in the SAT Primary II battery. This technical modification reduces the number of students scoring at the extreme ends of the subtests (floor and ceiling effects) and, when coupled with standardized administration and the use of special norms for hearing impaired students, provides for the more accurate assessment of student abilities. With these improvements, the SAT-HI is preferred to the SAT for use with hearing impaired students. However, since the item wording and content are unaltered, and since norms developed for hearing students are still used, the SAT-HI is not to be viewed as or interpreted like a test entirely designed, standardized, and normed for a hearing impaired adolescent population.

Content

The content of test items has been identified as one of the most important aspects is selecting and judging tests (Hoepfner, 1977). It is the content of the items that determines which composite skill is being assessed by a particular instrument. Several tests may measure reading ability; however, how reading ability is defined and operationalized can vary greatly from test to test. For example, the second grade level of the Sequential Test of Educational Progress (STEP) contains a large number of items assessing phonetic word attack skills. The SAT and SAT-HI Level I reading tests contain few such items. On the other hand, the low levels of the SAT-HI contain a number of items assessing students' abilities to infer meaning of words from context and the STEP contains none.

As an example of how a particular set of abilities is defined, Table 1 contains a breakdown of the content gauged by the SAT-HI Math Computation subtests at the six different levels. These content classifications help the test user index the subtests' relevancy. If one is teaching high school geometry and a student takes the Level III examination which emphasizes knowledge of the primary facts and the basic addition, subtraction, multiplication and division algorithms; then the students' scores do not reflect their classroom endeavors. Use of such a test to judge mastery of the curriculum would be faulty. However, using it to gauge ability in these basic skills, which may not be covered by the school's curriculum, does provide some useful, although limited, feedback.

In addition to the content covered by a subtest, a test user should consider the proportion of items within each content classification within a subtest. For example, the SAT-HI Math Computation Level I subtest emphasizes

Table 1

Percentage of Items within the Six Levels of the SAT-HI Mathematics Computation Subtests by Content Area

Item Grouping	SAT-HI Level 1	2	3	4	5	6
Addition and subtraction facts	41%					
Mathematical sentences	15%					
Verbal problems	44%					
Knowledge of primary facts		66%	45%	53%	42%	42%
Addition and subtraction alogrithms		17%	22%	13%	11%	11%
Multiplication and division alogrithms		17%	33%	25%	9%	9%
Common fractions				9%	11%	11%
Other operational models					27%	27%

simple verbal problems and addition and subtraction facts. The same subtest at Level III emphasizes all the basic operations (see Table 1). While the subtests have the same name, they really assess different abilities and scores on the two tests do not necessarily represent the same ability. Therefore, caution must be exerted in comparing students taking tests at different levels. A student who obtains a grade equivalent score 2.8 at the Level II exam will not necessarily receive a 2.8 on the Level III exam.

In practice, the curriculum of most school districts, and especially the curriculum in programs for the hearing impaired, is not fully reflected in the content of a standardized achievement test. This does not mean that scores on standardized achievement tests are worthless; only that they must be evaluated in perspective. In a school for the deaf, they provide an index of how well hearing impaired students perform on certain tasks--tasks which are representative of the basic skills taught the standard, albeit normal hearing, school children population at certain elementary grade levels.

Since parts, as opposed to all, of a subtest may be relevant to the curriculum efforts or achievement desires, attention might be given to the results on a particular content classification as opposed to total scores. This is referred to as objective referencing. The SAT was specifically "designed for dual interpretation in the normed-referenced and objective-referenced modes" (SAT Manual, Part V, p. 12), and Item Analysis Reports are available as part of the publisher's scoring service. A school for the deaf may be particularly interested in having its students capable of performing the basic addition, subtraction, multiplication and division operations and not particularly concerned with fraction operations. By determining the school's average on the appropriate items within the Level IV, V, and

VI Mathematics Computation subtest, the test user is in a position to state whether students taking the exam are performing satisfactorily in this area of interest. Those wishing to capitalize on this beneficial analytical method are referred for additional information to the Stanford Research Report #10 and to the SAT Manual Part III: Teacher's Guide for Interpreting.

Item Wording

Items are worded to gauge skills in particular areas. The wording of the item, however, may be such that undesired skills are gauged. For example, an item may be designed to gauge mathematics ability, but because of the wording, it may largely assess reading ability. This is a major problem when tests are used with minority students and has led to much of the work in test and item bias.

Several linguistic structures have been identified as causing undo difficulty for hearing impaired students (Rudner, 1978). These include conditionals (if, then), inferentials (could, should), comparatives (greater than, less than), negations (not, without), and low information pronouns (it, something). The SAT-HI, particularly the vocabulary subtests, contains items that incorporate one or more of these structures. Consequently, the results on various subtests may not always reflect the intended skills and scores can be spuriously low.

Because hearing impaired students taking the SAT-HI tend to be older than hearing students taking the SAT, hearing impaired students are able to draw from a larger repertoire of experience in responding to the items. Thus, other item wordings can favor hearing impaired examinees and spuriously raise their scores.

Effect of Item Wording on Test Accuracy

Perfect test items--items which, for a particular age group, measure only an intended skill--are difficult, if not impossible, to develop. The item tryout procedure will identify the items which are best for a given population, but there still will be errors in measurement. These errors will be increased when special populations are used and increased even more with age differences between the special population and the standardization sample.

Measurement error can be gauged in several ways. The reliability coefficient indexes the expected consistency of test results. The higher the reliability, the greater is the expected consistency. In comparing the SAT subtests used with hearing elementary school children against the SAT-HI subtests used with hearing impaired adolescents, the reliabilities are consistently higher for hearing children. The reliabilities on the Level II reading comprehension subtest, for example, is .95 for hearing examinees and .83 for hearing impaired examinees. This differential reliability means that the scores for hearing impaired students contain more errors than the scores for hearing students.

The reliability coefficient is a useful statistic for evaluating a test; however, it is not directly applicable in interpreting test results. A more useful statistic for this purpose is the Standard Error of Measurement (SEM), which provides an estimate of the variation of the amount of error in the test scores for a given population. The larger the SEM, the larger the interval in which one is confident an examinee's true ability lies.

Table 2 outlines the Standard Errors of Measurement based on hearing impaired examinees for the subtests by level of the SAT-HI, expressed in terms of raw scores. Considering that most subtests contain about 50 items,

Table 2

Raw Score Standard Errors of Measurement for the Subtests of the SAT-HI by Level

	SAT-HI Level					
Subtest Area	1	2	3	4	5	6
Vocabulary	2.8	2.8	2.8	3.0	3.1	3.1
Reading A	2.9	2.7				
Reading B	2.9	3.1				
Reading Comprehension	4.1	4.1	3.6	3.7	3.6	3.6
Word Study Skills	3.4	3.5	3.1	3.0	2.9	
Math Concepts	2.4	2.6	2.5	2.7	2.7	2.6
Math Computation	2.4	2.4	2.6	2.8	2.9	2.8
Math Application		2.3	2.2	2.5	2.6	2.6
Spelling		2.8	2.9	3.3	3.5	3.1
Language			3.2	4.0	4.1	3.9
Social Science		2.3	2.9	3.5	3.3	3.4
Science		2.4	2.9	3.6	3.6	3.5
Communication Comprehension	2.4	2.4	3.2	3.2	32.	

Adapted with permission from Jensema, Trybus & Schildroth (in press).

these SEMs are fairly large. If one were using grade equivalent scores, one SEM would correspond to about three-tenths of a grade equivalent for the median student and more for the higher and lower ability students. Thus, scores on the SAT-HI, like any other test, are not to be taken as precise estimates of ability.

Norms

Norms provide a frame of reference in interpreting test results. Student performance can be compared to national averages by way of grade equivalent scores and percentiles. Student gain over time can be gauged by differences in these norm scores, or more appropriately by differences in scaled scores which are specifically designed for this purpose. While norms based on hearing elementary school students and on hearing impaired students are available, the latter set is most relevant and meaningful to schools for the deaf.

In developing and using norms, careful consideration needs to be made with regard to the reference population (Angoff, 1971). This was done by the test publishers who went to great lengths to insure that the standardization sample closely matched the average United States population (see Stanford Research Report #3). Similarly, the Office of Demographic Studies carefully delineated its sample in developing special norms for hearing impaired students (see Trybus and Karchmer, 1977). Thus the test user need not be particularly concerned with the representativeness of the various norms. However, recognition of their limitations is essential for proper use and interpretation.

Perhaps one of the most misunderstood type of norm is the grade equivalent score (GES). This norm, which is available only with hearing students as the reference population, is intended to provide for a method of defining pupil performance in terms of median public school grade level performance.

Suppose data were collected from representative samples of children at all grade levels in the seventh month of the school year. A table could be developed for the seventh month of each grade which converts raw scores to their percentile equivalents. Similarly, the median score at each grade level could be converted to a grade equivalent score of the form grade level plus .7. The median score for these first graders in the seventh month of the school year would convert to a GES of 1.7, for second graders, 2.7, and so on.

Grade equivalent scores are given by the publishers for points other than 1.7, 2.7, 3.7, etc. These GES's were derived by extrapolating between Fall and Spring norm results. Psychometricians, however, have clearly pointed out that extrapolation of norms is statistically unsound and lends the norms to serious misinterpretation (e.g., Tallmadge, 1977).

One obvious and serious consequence of extrapolating GES is reflected in the anomalies of converting GES's on the 1964 edition of the SAT to GES on the 1973 edition of the same test. Table 3, taken from Research Report #5 of the SAT, shows the corresponding grade equivalent scores on the two Intermediage II (Level #3) tests. For example, a GES of 3.7 on the 1964 Spelling test would correspond to 4.3 on the 1973 Spelling test. That is, by changing from the 1964 to the 1973 version of the SAT, a student would evident .6 GES growth without any corresponding change of ability. Similarly,

Table 3

1964 SAT Intermediate II GES versus the 1973 SAT Intermediate II GES

1973 SAT GRE	Word Mean -- Voc.	Para. Meaning -- Reading Comp.	Spell -- Spell	Arith. Comp. -- Math. Comp.	Arith.[a] Comp. -- Math[b] Concepts
5.5	4.9	4.7	4.5	4.5	5.4
5.4	4.8	4.6	4.4	4.4	
5.3	4.7	4.5			5.3
5.2	4.6	4.4	4.3	4.3	5.2
5.1	4.5	4.3	4.2	4.2	5.1
5.0					5.0
4.9	4.4	4.2	4.1	4.1	4.9
4.8	4.3	4.1	4.0		4.8
4.7	4.2			4.0	4.7
4.7	4.1	4.0	3.9	3.9	4.6
4.5	4.0	3.9	3.8		4.5
4.4	3.9	3.8		3.8	4.4
4.3		3.7	3.7		4.3
4.2	3.8	3.6		3.7	4.1
4.1			3.6	3.6	4.0

in some instances month-to-month gains can be expected without any increase in ability, and the learning curve projected from the GES is unrealistic.

As long as testing is conducted at the seventh month, the percentile scores, unlike the GES, would have clear meaning. A student's performance is defined by the percentile as the percent of students in the reference population scoring less than he on that level of the SAT in the seventh month. It is for this reason that percentiles are given for a specific time of the school year. If testing is conducted at a different time of the year, interpretation of the percentile score is unclear and tenuous at best. It cannot be determined, for example, whether a student taking the SAT in the second month and scoring in the 45th percentile is above or below average with respect to his grade level peers.

While the GES and percentiles can lend themselves to misinterpretation when used with hearing elementary school students, they do provide meaningful anchors when used properly. The test scores of a third grade student in a public school taking the Level II battery can be mapped to a meaningful percentile score and a meaningful GES. Providing that the test was administered at the same time of the school year in which the norming was conducted, the student's percentile score can provide an index of the relative standing of this student with respect to other third graders. For example, had she scored in the 80th percentile, her score can be interpreted as being better than 80 and lower than 20 out of every 100 third grade students taking the test at the same time of the year. The GES can index whether she is performing of other third grade programs in other schools, again providing a gross index of poor, typical, or above average.

Anchored to the populations of grade level public elementary school children, the GES's and percentile scores are of limited value when used to describe the performance of hearing impaired adolescents. The hearing impaired adolescent is older and is not receiving the same curriculum as the hearing student. The feedback provided by a test which says that a high school aged hearing impaired student does better than 40% of the second grade public school children on a test which emphasizes a different curriculum is almost irrelevant. To say that a student had a GES of 3.5 basically defies interpretation. The scale is not relevant to the student or the efforts of the scholastic program. Gross comparison of the performance of hearing impaired adolescents to the public elementary school curriculum is misleading enough; the use of monthly equivalents giving the false impression of greater accuracy, compounds the matter, especially if one considers that there may be at least three months error in that decimal.

A total abandonment of public school grade level percentiles and grade equivalent scores would eliminate much of the misuse of test scores and probably serve to enhance the utility of the SAT-HI. One might argue that a comparison with hearing students is essential since it indexes the potential of deaf students. However, the fact that a test designed for second grade public school children is administered to high school aged hearing impaired students already provides substantial comparative information.

of students in U.S. programs for the hearing impaired, these norms allow for meaningful descriptions of how well a hearing impaired student or group of hearing impaired students performed on the SAT-HI.

Percentiles for hearing impaired students based on test levels are printed on the score reports for those who use the publisher's computer scoring service. These percentiles provide for comparison of a hearing impaired child with other hearing impaired children who were tested at the same difficulty and on the same content. They do not, however, allow for comparisons across levels.

Percentiles for hearing impaired students based on age are available through the Office of Demographic Studies. These percentiles allow one to determine how well an individual student performed in comparison to a national sample of hearing impaired children of the same age, regardless of the level of the SAT-HI. In using these percentiles, one must remember that the content does differ across the levels of the SAT-HI, so the comparison will not always be exact.

Like other percentiles, these percentiles based on hearing impaired examinees only have meaning if testing is conducted at the same time of the year the norming sample took the examination. Thus, if one wants to meaningfully describe student performance, testing should be conducted during the Spring.

One major purpose of achievement tests is to determine whether students have acquired new skills over time. While the SAT-HI may not be amenable describe student performance in terms of relative standing within a group of students. Over time, the group will have progressed, but the student's standing may remain the same. Occasionally, researchers use the GES as a relative index of ability by comparing changes in GES. However, this is not good practice, since the GES is not on an interval scale. The difference between a GES of 3.2 and 3.8 is not the same as the difference between a GES of 6.2 and 6.8. Further, with the error inherent in extrapolating values, the error in the gain scores becomes quite large.

If one is interested in student gains, two psychometrically sound options are available. The first preference is to use raw scores without any transformation. However, this is only possible when the same level of the SAT-HI is given during both administrations. An alternate choice, applicable regardless of the examination level taken, is to use scaled scores.

Scaled scores have the unique advantage of providing approximately equal units on a continuous scale. The scale has two reference points, scaled scores of 132 and 182 correspond to a GES of 3.2 and 8.2, and each unit is intended to represent the average monthly gain over a five year period. However, the absolute meaning of the units is not of interest when assessing gain. By describing location on a continuous, equal-interval scale, scaled scores overcome some of the difficulties of using percentiles and grade equivalent scores to assess gain. A difference of five scaled score points, for example, is the same regardless of where it occurs on the scale.

priately worded and norms which provide meaningful indices of growth and performance are available. The test has been shown to be both valid and reliable for this population.

The content, wording and norms of the SAT were re-examined in this paper in order to clarify the use and interpretation of the SAT with hearing impaired students. Specific recommendations are:

1. Use the Special Edition of the SAT, the SAT-HI, which minimizes floor and ceiling effects, standardizes administration and provides norms based on hearing impaired students.
2. Examine the content of the SAT-HI to determine its relevancy to the curriculum efforts.
3. Use content classification analysis when appropriate.
4. Do not interpret the scores of hearing impaired students as precise estimates of ability.
5. Use the percentiles based on hearing impaired students to access student performance.
6. Use raw scores, when applicable, or scaled scores to assess student growth.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971, Chap. 15.

Buchanan, C. Results of a survey on the use of achievement tests in educational programs for hearing impaired students: United States, 1972-73. In Further studies in achievement testing, hearing impaired students, United States: Spring, 1971. Washington, D.C.: Gallaudet College, Office of Demographic Studies, 1973, Series D, Number 13, 36-45.

Buros, O. K. (Ed.). The seventh mental measurements yearbook. Highland Park, N.J.: The Gryphon Press, 1972.

Hoepfner, R. Achievement test selection for program evaluation. In D. R. Green & M. Wargo (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation New York: McGraw-Hill, 1977.

Jensema, C. J., Trybus, R., & Schildroth, A. A user's handbook for the special edition of the Stanford achievement test for hearing impaired students (SAT-HI). Washington, D.C.: Gallaudet College, Office of Demographic Studies, in press, Series T, Number 2.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Rudner, L. M. The use of standard tests with the hearing impaired: The problem of item bias. Volta Review, 1978, 80, 31-40.

Stanford Achievement Test Manual, Part III: Teacher's guide for interpreting. New York: Harcourt Brace & Jovanovich, Test Department, n.d.

Stanford Achievement Test Manual, Part IV: Technical data report. New York: Harcourt Brace & Jovanovich, Test Department, n.d.

Stanford Research Report #3, Norms development summary. New York: Harcourt Brace & Jovanovich, Test Department, n.d.

Stanford Research Report #4, Stanford norms: Derivation, characteristics, and uses. New York: Harcourt Brace & Jovanovich, Test Department, n.d.

Stanford Research Report #5, Equivalent scores for the 1973 edition of Stanford

Stanford Research Report #10, Interpreting Stanford item analysis reports. New York: Harcourt Brace & Jovanovich, Test Department, n.d.

Tallmadge, G. K. Cautions to educators. In D. R. Green & M. Wargo (Eds.), Achievement testing of disadvantaged and minority students for educational program evaluation. New York: McGraw-Hill, 1977

Trybus, R., & Jensema, C. J. The development, use and interpretation of the 1973 Stanford achievement test, special edition for hearing impaired students. In Report of the proceedings of the 47th meeting of the convention of American instructors of the deaf. Greensboro, N.C., June 1975, 73-78.

Trybus, R., & Karchmer, M. A. School achievement scores of hearing impaired children: National data on achievement status and growth patterns. American Annals of the Deaf, 1977, 122, 62-69.