DOCUMENT RESUME

ED 269 418 TM 860 208

AUTHOR	Bennett, Randy Elliot; And Others
TITLE	The Psychometric Characteristics of the SAT for Nine Handicapped Groups. Studies of Admissions Testing and Handicapped People. Report No. 3.
INSTITUTION	College Entrance Examination Board, New York, N.Y.; Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-85-49
PUB DATE	Nov 85
NOTE	80p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Aptitude Tests; *College Entrance Examinations; *Disabilities; Hearing Impairments; Item Analysis; Learning Disabilities; Physical Disabilities; *Psychometrics; Scores; Secondary Education; Special Programs; Standardized Tests; Testing; Testing Programs; *Test Norms; *Test Reliability; Visual Impairments
IDENTIFIERS	*Scholastic Aptitude Test; Speededness (Tests)

ABSTRACT

This study examined the psychometric characteristics of the Scholastic Aptitude Test (SAT) administered under special conditions for nine handicapped groups. Four psychometric characteristics were studied: level of test performance, test reliability, speededness, and extent of unexpected differential item performance. Psychometric comparisons were made between a non-handicapped sample and each of nine different handicapped classifications. These contrasts were replicated across two forms of the same test, serving to increase confidence in the stability of results and their applicability to other SAT forms. With the exception of performance level, the psychometric characteristics of the SAT were generally comparable for the handicapped and nondisabled groups studied. It is concluded that this result should extend to other forms of the SAT and other disabled students to the extent that these groups and forms, and the conditions under which they are administered, are similar to those employed in this study. That the psychometric characteristics of the test are similar across populations provides evidence necessary to support SAT scores as accurate and fair indicators of the developed scholastic abilities of disabled students. (Author/PN)

RR-85-49

THE PSYCHOMETRIC CHARACTERISTICS OF THE SAT FOR NINE HANDICAPPED GROUPS

Randy Elliot Bennett
Donald A. Rock
and
Bruce A. Kaplan

November 1985

Report No. 3
Studies of Admissions Testing and Handicapped People
A Project Sponsored by

College Entrance Examination Board
Educational Testing Service
Graduate Record Examinations Board

Studies of Admissions Testing and Handicapped People

Most admissions testing programs have long made accommodations for handicapped examinees, though practices have varied across programs and limited research has been undertaken to evaluate such test modifications. Regulations under Section 504 of the Rehabilitation Act of 1973 impose new requirements on institutional users, and indirectly on admissions test sponsors and developers, in order to protect the rights of handicapped persons. The Regulations have not been strictly enforced since many have argued that they conflict with present technical capabilities of test developers. In 1982, a Panel appointed by the National Research Council released a detailed report and recommendations calling for research on the validity and comparability of scores for handicapped persons.

Due to a shared concern for these issues, College Board, Educational Testing Service, and Graduate Record Examinations Board initiated a series of studies in June 1983. The primary objectives are:

To develop an improved base of information concerning the testing of handicapped populations.

To evaluate and improve wherever possible the accuracy of assessment for handicapped persons, especially test scaling and predictive validity.

To evaluate and enhance wherever possible the fairness and comparability of tests for handicapped and nonhandicapped examinees.

This is one of a series of reports on the project, which will continue through 1986. Opinions expressed are those of the authors.

ETS Research Report RR-85-49

The Psychometric Characteristics of the SAT for Nine Handicapped Groups

Randy Elliot Bennett

Donald A. Rock

and

Bruce A. Kaplan

November 1985

Abstract

This study examined the psychometric characteristics of the Scholastic Aptitude Test (SAT) administered under special conditions for nine handicapped groups. Information about test characteristics is central to judging the accuracy and fairness of scores from SAT special administrations.

Four psychometric characteristics were studied: level of test performance, test reliability, speededness, and extent of unexpected differential item performance. Psychometric comparisons were made between a nonhandicapped sample and each of nine different handicapped classifications. These contrasts were carried out twice; that is, they were replicated across two forms of the same test. The use of two samples taking different forms served to increase confidence in the stability of results and their applicability to other forms of the SAT.

Results of the study showed that visually impaired students and those with physical handicaps achieved mean scores generally comparable to students taking the SAT in national administrations. Learning disabled and hearing impaired students scored lower than their nondisabled peers. Differences between Verbal and Mathematical performance were also comparable to those for the nondisabled reference group in all but the hearing impaired-regular type test and visually impaired-braille test samples. Hearing impaired-regular students scored higher on Mathematical than on Verbal relative to their nondisabled peers, while visually

impaired-braille students showed no consistent superiority for Mathematical over Verbal.

Analysis of test reliability revealed no practical differences in measurement precision across groups. Data on test speededness showed no evidence of disadvantage for disabled students; the amount of extended time allotted through special administrations appears to allow roughly equivalent proportions of handicapped and nondisabled examinees to complete the test.

Because of the large number of groups and test items involved, unexpected differential item performance was examined through a two-stage procedure. The first stage centered on the performance of item clusters. Individual items composing clusters showing questionable performance were then examined. This two-stage procedure revealed only a few instances of differential item performance localized to visually impaired students taking the braille test.

It is concluded that, with the exception of performance level, the psychometric characteristics of the SAT are generally comparable for the handicapped and nondisabled groups studied. These results lend support to the contention that scores from special administrations are fair and accurate measures of the developed scholastic abilities of handicapped students. Further studies of these scores--in particular, their factor structure and predictive validity--should provide additional information about their meaning for handicapped students.

Acknowledgements

Appreciation is expressed to Thomas Allen, William Angoff, Henry Braun, Linda Cook, Catherine Nelson, Marjorie Ragosta, Joseph Torgesen, and Warren Willingham for their comments on an earlier version of this manuscript. Gratitude is also extended to Joan Chase, Linda Cook, Chris Kiler, June Morris, Catherine Nelson, Marjorie Ragosta, and Susan Tillett for their content reviews of test items, and to Nancy Wright for providing basic information about the SAT. Finally, appreciation is expressed to the members of the College Board Joint Staff Research and Development Committee and the Graduate Record Examinations Board Research Committee.

Table of Contents

The Scholastic Aptitude Test..............................1

Subjects..6

Results...10

    Level of Performance................................10

    Reliability...14

    Test Speededness....................................16

    Unexpected Differential Performance.................17

        Item clusters...................................20

        Individual items................................28

Summary and Recommendations.............................38

References..43

Tables..47

Figures...68

In 1983, the College Board, Educational Testing Service (ETS), and the Graduate Record Examinations (GRE) Board initiated a joint project, "Studies of Admissions Testing and Handicapped People," in response to a call by a National Academy of Sciences Panel for further research into the use of college and graduate admissions tests for handicapped individuals (Sherman & Robinson, 1982). As part of that joint research effort, this study presents information on the psychometric characteristics of the Scholastic Aptitude Test (SAT) for nine groups of handicapped examinees. The study reports data on the level of performance, test reliability, speededness, and extent of unexpected differential item behavior for these groups. These data, in particular those on reliability and differential performance, are fundamental to evaluating the extent to which the SAT fairly and accurately measures the developed scholastic abilities of handicapped students.

The Scholastic Aptitude Test

The Scholastic Aptitude Test is developed and administered by ETS as part of the Admissions Testing Program of the College Board, an independent, nonprofit membership organization that provides tests and other educational services to students, schools, and colleges (College Board, 1983). The Board's membership is composed of more than 2500 colleges, schools, school systems, and educational associations. Along with other indicators, institutions use the SAT to select students for admission, to monitor changes in the academic capabilities of their applicant and entering-freshmen populations, and to recruit and place students.

The SAT is a multiple-choice examination made up of Verbal and Mathematical sections. The Verbal section of the exam is composed of 85 items falling into four categories: analogies (20 questions), antonyms (25 questions), sentence completion (15 questions), and reading comprehension (25 questions). Analogies items are meant to assess the examinee's ability to detect verbal relationships between pairs of words while antonyms are designed to measure breadth and depth of vocabulary (Dorans, 1982). Together, performance on these item types forms the SAT Vocabulary subscore.

The Reading subscore of the SAT reflects performance on sentence completion and reading comprehension items. Sentence completion tests a student's ability to recognize logical relationships among parts of a sentence. Reading comprehension questions assess a greater variety of abilities including recalling specific details, identifying the main idea, making inferences, analyzing arguments used by the author, detecting the author's tone or attitude, and making generalizations on the basis of presented information (Dorans, 1982). Examples of each SAT-Verbal item type are presented in Figure 1.

Insert Figure 1 about here

The Mathematical section of the SAT contains 60 questions divided among two formats: standard multiple choice (40 questions) and quantitative comparison (20 questions). Quantitative comparisons emphasize the concepts of equality, inequality, and estimation, and generally involve less reading, take less time to answer, and require less computation than standard multiple choice questions (College Board, 1983). The quantitative comparison typically presents two quantities. The test candidate must examine the quantities and select from four options the one that best describes the relationship between the two amounts. Examples of the quantitative-comparison and standard multiple-choice item types are presented in Figure 2.

Insert Figure 2 about here

The content of items in the SAT Mathematical section is divided almost equally among arithmetic, algebra, geometry, and miscellaneous questions designed to measure abilities related to college-level work in the liberal arts, sciences, engineering and other fields requiring mathematics. Miscellaneous questions test logical reasoning, number theory, number systems, or other content that does not readily fit into any of the three basic categories listed above.

When administered, the SAT is divided into five separately timed, 30-minute sections: two verbal, two mathematical, and one experimental section that does not count toward the student's score. The sections are bound together in a test booklet that also contains a 50 question Test of Standard Written English ᵌigned to assist colleges in placing students in freshman English courses. Items of a similar format are typically grouped together within sections, though more than one item format can appear in each section and the same item type can appear in more than one section.

National administrations of the SAT are offered seven times a year. The composition of student groups taking the test at different times of the year varies widely with high school seniors constituting the bulk of examinees during the fall administrations and juniors counting for the larger group during the spring exam period. Differences in average ability are also apparent across administrations, with the more able groups taking the exam during the early fall (seniors) and late spring administrations (juniors).

Special administrations for handicapped students have been offered since 1938, when braille and large-type versions of the test were administered to visually-impaired examinees (Saretsky, 1983). Since that time, special accommodations have been extended to students with physical, hearing, and learning disabilities and extra time and rest periods; cassette, braille and large-type presentations; the use of a reader or scribe; and various combinations of these arrangements have been offered.

Results of SAT administrations are reported for Verbal and Mathematical performance, each on a 200 to 800 standard-score scale with a mean of 500 and standard deviation of 100. The scale is based on the performance of college applicants taking the test in 1941 (Donlon, 1984); the performance of all subsequent groups is statistically equated to that original administration. Hence, the means and standard deviations of groups taking the test have deviated over the years from their original values, but the meaning of scores has stayed the same. Subscores for Vocabulary and Reading are reported on a 20 to 80 scale. Scores are accompanied by the designation, "NON STD," whenever the test was not administered under standard conditions and ETS cannot assume comparability of the scores to those achieved under typical circumstances.

The psychometric characteristics of the SAT have been widely studied in the general population and in some special populations (e.g., black examinees), but not with handicapped students (Bennett, Ragosta, & Stricker, 1984). Median correlation coefficients with college grades based upon 827 predictive validity studies were reported to be .41 for the total test, .37 for Verbal, and .32 for Mathematical (Educational Testing Service, 1980). Median coefficients for high school grade point average (HSGPA) and for the SAT and HSGPA combined were .52 and .58, respectively. As with all averages, these median coefficients mask variation. The predictive validity of the SAT varies as a function of institutional characteristics (selection rules, grading standards, educational program), academic year, student population, and other factors. In some cases, these factors cause the SAT's predictive validity to approach zero, while in many others it is much higher than that attributed to high school grade point aver ge (Breland, 1978).

Subjects

During the period from Fall 1978 through July 1983, the Admission Testing Program's Services for Handicapped Students offered two forms of the SAT, designated as WSA3 and WSA5, to handicapped students requesting special administrations. Because retention of student data from special administrations began in 1980, the only data available for analysis are from March of that year through June 1983, the time that two new forms were put into special service.

During the March 1980 to June 1983 time period, 16,961 students were given special administrations of the SAT. Of these students, 5,213 and 4,236 are known to have taken WSA3 and 5, respectively. Which of the two forms was taken by the remaining students is unknown. During this period, other handicapped students undoubtedly took standard administrations of the SAT on national test dates. Because it is not necessary to reveal the presence of a disability unless a special administration is requested, the number of handicapped students taking standard administrations is unknown.

In this study, data from both WSA3 and 5 are used. By using these two data sets, attention can be focused on those findings that replicate across forms. Because of their co-occurrence, such findings are less likely to be artifacts associated with a single form or particular sample of subjects. They are more probably stable results that will manifest themselves in other samples from the same disability group and on other forms of the SAT.

Students requesting special administrations of the SAT during the study period fell into five major disability groups: visually impaired (VI), physically handicapped (PH), hearing impaired (HI), learning disabled (LD), and multiple handicapped. Types of special administrations offered included braille, large type, cassette, regular type, cassette and large type, braille and cassette, and cassette and regular type. All special administrations included the option of extended time. Tables 1a and 1b show the number of students with each disability taking each type of special administration of WSA3 and WSA5.

Insert Tables 1a and 1b about here

As the tables show, the largest number of special administrations (3552 for WSA3 and 2883 for WSA5) were taken by learning disabled students and the most frequently used format was regular type (3889 for WSA3 and 2924 for WSA5). Visually impaired students represented the second largest disability group (893 for WSA3 and 858 for WSA5) and large type the second most-frequently used format (726 and 676). Of the 35 possible test-format-by-disability-group combinations, the two largest were LD students taking regular-type (2983 for WSA3 and 2316 for WSA5) and visually impaired students taking large-type administrations (486 and 498).

In addition to these two groups, seven other format-by-group combinations have numbers of students (roughly 100 or more on each form) sufficient to support dependable results and justify further study. These groups are, for regular type, visually impaired, hearing impaired, and physically handicapped students; for large type, learning disabled pupils; for braille, visually impaired examinees; and for cassette and cassette and regular type, learning disabled pupils. Table 2 lists the sample sizes and acronyms used to denote these nine groups.

Insert Table 2 about here

To properly evaluate the psychometric characteristics of the SAT for these nine disability groups some reference, or "standard," population is needed. Without such a population, the typical behavior of the test cannot be known and any departures from this behavior by subpopulations cannot be detected. In the present study, several standard groups are used. Among these groups are the 5.1 million high school students taking all forms of the SAT offered during the March 1980 to June 1983 time period. Most comparisons, however, are based on a standard group of high school students who took forms WSA3 and WSA5 under typical testing conditions. WSA3 was administered to 35,424 high school seniors in Texas and California during October 1974; WSA5 was given nationally to 33,161 high school juniors in December of that same year.

Table 3 lists the mean Verbal and Mathematical scores for high school pupils taking WSA3 and WSA5, and for high school students taking other forms of the SAT during the March 1980 to June 1983 period. As the table shows, the high school seniors taking WSA3 perform better than their counterparts taking the SAT during the 1980 to 1983 period on both Verbal and Mathematical, suggesting that the WSA3 group is somewhat more select than the group of seniors typically taking the SAT. On the other hand, the juniors taking WSA5 seem to perform substantially worse than their counterparts taking the test during the 1980-83 period. Hence, students taking the WSA forms may not be broadly representative of those taking the SAT during the 1980-83 period. Still the nonhandicapped group taking the same form under standard conditions, though not ideal, should prove workable where needed as a reference for the nine disability groups used in the study.

Insert Table 3 about here

Results

Results are reported for level of performance, test reliability, speededness, and extent of unexpected differential item performance.

Level of Performance

Table 4 lists scaled score means and standard deviations for the performance of the nine handicapped groups on the Verbal and Mathematical sections of WSA3 and WSA5. Summary statistics for all pupils sitting for the exam during the March 1980 to June 1983 period (designated NHA) are also included. These students have taken test forms other than WSA3 and 5. However, because Verbal and Mathematical scores on the SAT are equated across forms, the scores of this reference population are expressed on the same scale as those of the nonhandicapped students taking WSA3 and 5.

Insert Table 4 about here

To facilitate comparison with students typically taking the SAT, Table 5 presents the difference between the handicapped and nondisabled student means in standard deviation units of the nondisabled group. Review of Table 5 suggests some consistency in the performance of disability groups across SAT forms. On the Verbal section, the mean performance of the three visually impaired groups (VIB, VIR, VIL) and of the physically handicapped group (PHR) is generally better than or just below the nonhandicapped reference group (NHA). The LD (LDR, LDCR, LDC, LDL) and hearing impaired (HIR) groups have substantially lower mean scores than the reference group, usually by at least a half standard deviation. This general pattern appears to hold for the Mathematical section also, with the possible exception of visually impaired students taking the braille format (VIB). These students score relatively close to the nondisabled mean on one form and dramatically below it on the other.

Insert Table 5 about here

In addition to mean performance, the degree of variability evidenced for some groups is also noteworthy (see Table 4). On the Verbal section, restrictions in the range of scores are found on both forms for the LD groups taking cassette (LDC) and cassette and regular tests (LDCR), while an unusually wide range with respect to the reference group is noted for visually impaired-braille (VIB) students. For Mathematical, LD students taking cassette and regular editions (LDCR) show a restricted range on both forms. Consistently widened ranges are found for two visually impaired groups, those taking the regular edition (VIR) and students using the large-type version (VIL).

Aside from mean performance and degree of variability, differences in intra-test scores are of interest. Table 6 shows the extent to which each group scored better on Verbal or Mathematical relative to all students taking the SAT during 1980-83. The tabled indices represent the ratio of the difference between the Verbal and Mathematical scores for a handicapped group divided by the pooled standard deviation for that group to the same quantity calculated for nonhandicapped students. Positive values indicate a difference in the same direction as for the reference group (i.e., Mathematical greater than Verbal), while negative values denote the converse. The magnitude of the index shows the extent to which the standardized difference is as large as the comparable value for the reference group. A value of 1.00 indicates intra-test performance equivalent in magnitude and direction to the reference group.

From the table it can be seen that hearing impaired-regular type students (HIR) show a consistent performance difference in favor of Mathematical about twice as large as for the nonhandicapped reference group. This performance difference is consonant with the documented English language deficiencies of this group (e.g., Meadow, 1980). Visually impaired-braille pupils (VIB) also show consistently different intra-test performance. Unlike the reference group, these students do not evidence uniformly superior performance on Mathematical relative to Verbal. One possible explanation for this finding is that visually impaired-braille students are encountering unusual difficulty with geometry and other math items involving the understanding of figures, tables, or special symbols.

Insert Table 6 about here

A final point of interest relates to differences between each group's performance on the two SAT forms (see Table 7). Examination of Table 7 shows that the scores of some groups differ substantially across forms. Because scores from different forms are equated, variations in performance generally suggest real differences in the abilities of the groups taking one form or another. An alternative explanation is that the equating procedure, which is based on the performance of nonhandicapped students taking standard administrations, operates differently when applied to the scores of disabled pupils taking nonstandard examinations. This latter possibility is not very likely, however, since all the handicapped distributions show considerable overlap with the standard population.

Insert Table 7 about here

Reliability

Reliability refers to the precision or accuracy with which a test measures. Differences in the precision of measurement across groups can negatively impact upon the less accurately-measured group. For example, consider what might happen if an admissions test measured less precisely for deaf than for hearing students. In this situation, the dispersion of the observed scores of deaf students around their true scores (i.e., those scores indicative of their actual abilities), would generally be greater than they would for hearing pupils. The admissions officer's decision to admit or place a deaf student would, therefore, be subject to a greater likelihood of error than for nonhandicapped applicants.

The two indices most often used to assess test reliability are the reliability coefficient and the standard error of measurement (SEM). By definition, the reliability coefficient is affected by the amount of test score dispersion in a group, with smaller variances tending to produce smaller reliability coefficients. Because of this sensitivity to within-group homogeneity, the reliability coefficient is limited as a comparative measure of precision across groups. (It retains utility, however, as an index of the test's ability to separate individuals within a given group.) The standard error of measurement is relatively unaffected by score variance. It, therefore, is better suited to the comparison of measurement accuracy across populations.

Table 8 presents alpha reliability coefficients and standard errors of measurement for handicapped and nonhandicapped students taking WSA3 and 5. For the nonhandicapped students (denoted as NHF), reliability coefficients for the Verbal section are both .92. Coefficients for the disability groups fall within a few points of these values, with the exception of the learning disability-cassette (LDC) and LD-cassette/regular (LDCR) groups, for which the coefficients run between .84-.86. As previously noted, these groups are also among the most restricted in score range.

Insert Table 8 about here

Standard errors of measurement are presented in raw score units. For the high school students taking the 85-item Verbal sections of WSA3 and 5, the raw-score SEMs are 3.73 and 3.75, respectively. Without exception, the SEMs for all handicapped groups are virtually identical to these values, differing by only a few hundredths of an item.

Reliability coefficients for nonhandicapped students taking the Mathematical section range from .91-.92. Again, coefficients for the handicapped groups hover closely about these values, though in this case no group consistently deviates from the nonhandicapped figures. Likewise, the

SEMs for the handicapped samples are virtually indistinguishable from those for the nondisabled group.

The alpha coefficients and SEMs reported above incorporate one primary source of measurement error: that due to differences in the samples of items used to assess scholastic ability. A second major source--error due to differences in the occasions on which ability was assessed--is not included. However, coefficients incorporating both major error sources have been reported for the SAT with similar results for several disability groups (Bennett, Ragosta, & Stricker, 1984), suggesting that consideration of the additional error source does not greatly change the comparability of measurement precision across populations.

Test Speededness

Special administrations of the SAT are commonly given with allowance for extra time and rest periods. However, the amount of extra time afforded may not be enough for the same proportion of disabled students to complete the test as their nondisabled peers, thereby introducing an unfair disadvantage into the testing process.

To check the extent to which the test is speeded for students taking special administrations, two indices, the percent of students completing the section and the percent finishing 75% of the section, were computed and compared to those for high school students taking the WSA forms in standard administrations. Because neither index is a fully satisfactory measure of speededness, they are jointly considered in the evaluation of test timing. (In isolation, the index based on the percent of students completing the section can be particularly misleading because it does not distinguish between students intentionally omitting the last item and those not reaching it. Hence, a closing item that is particularly hard for one group may cause this index to give a spurious indication of speededness.)

Table 9 presents the ratio of each disability group index to its reference-group counterpart. Values of 1.00 indicate equal percentages completing the section or part section for both groups, while those above 1.00 suggest greater completion rates for disabled students. When both speededness indices and both forms are simultaneously considered, it is clear that, with respect to the reference samples, no disability group is consistently disadvantaged by lack of time. On the contrary, several groups, such as hearing impaired-regular type students (HIR) and visually impaired-braille pupils (VIB), may receive more time than necessary on selected SAT sections.

Insert Table 9 about here

Unexpected Differential Performance

The concept of unexpected differential performance is derived from the notion that items on a unidimensional test should measure the same construct for different groups of examinees (Shepard, 1982). Items found to measure different constructs across groups are biased in the sense that, for some groups, they may be assessing factors irrelevant to the purpose of assessment. If found in any number, such items may unfairly lower (or raise) the test scores of a group. In many cases, however, biased items are found, in the aggregate, to affect different groups equally, cancelling any overall advantage or disadvantage that might otherwise be afforded (Berk, 1982; Shepard, 1982). Still, the identification of such items is important, for it alerts test developers to the kinds of questions that should be removed from future test revisions, or at least not disproportionately added lest the balance of questions favoring and disfavoring groups be destroyed.

Most methods of detecting items that operate differently across groups consider an item to be deviant if groups of equal ability perform differently on it. This definition of item bias makes sense only if it can be assumed that the test or subtest under investigation is basically unidimensional (Shepard, 1982). If the measure can be safely considered to be unidimensional, then differences in performance on an item that remain after standing on the dimension has been accounted for must be due to irrelevant sources.

A second common characteristic of item bias methods is that total test score is used as a proxy for ability level (Shepard, 1982). If all items in the test measure the same irrelevant construct for one group, it is possible that no indication of bias will appear; no item will stand out because it measures something different from the others. Item bias methods cannot, therefore, detect pervasive bias in a test because they lack an external criterion. As such, the study of item bias can be only one part of a comprehensive investigation of a test's fairness. At a more macroscopic level, a comprehensive investigation should also consider the test's factor structure, to see if the test actually measures the same general construct across groups, and its relationship with external variables, to ensure that relevant criteria are predicted with equal accuracy.

To detect the possible presence of SAT item types that operate differently across groups, a two-stage method was used. First, items were organized into logical clusters. Cluster structures were based on those characteristics that might prove unusually troublesome for particular groups of handicapped examinees and on groupings typically used in the SAT development process. The performance of these clusters was then investigated. Second, items belonging to deviantly-operating clusters were studied to determine if the cluster itself defined a potentially biased item type or, alternatively, if only a few aberrant items accounted for the unusual cluster performance.

This two-stage approach is somewhat different from the methodology traditionally used in item bias research. In the traditional approach, all items are individually assessed (e.g., see Kulick 1984). Individually assessing all items, however, has significant practical disadvantages in studies when several groups and forms are involved. First, this method necessitates the analysis of a large number of item performances. In the present study, nine disability groups and 145 items per SAT form would generate 2610 performances. Second, even in groups in which bias is known not to exist (e.g., two random samples drawn from the same population), statistical techniques will identify by chance some small proportion of items as biased (Sinnott, 1980). Assuming, for example, a significance level of .05, 2610 contrasts would produce 131 items flagged by chance alone. These items would, of course, be mixed in with other correctly identified questions. Separating the two groups through content analysis would take substantial effort, and in some cases be unsuccessful as the underlying causes for differential operation are frequently unclear (Scheuneman, 1982).

Item clusters. The rationale behind the study of item clusters is generally the same as that used for items: on a test measuring a single construct, a cluster should be of equal difficulty for different groups of examinees of the same ability. If not, the cluster is measuring different abilities in the groups.

To examine the performance of clusters, Verbal section items were divided by type into the four formats used in the test: antonyms (25 items), analogies (20), sentence completion (15), and reading comprehension (25). These item types were assumed to measure a single verbal ability factor, an assumption supported by the results of factor analysis (Rock, Bennett, & Kaplan, in press). The regression of each item-type cluster score on the total Verbal score for nonhandicapped students taking the forms (i.e., the standardization group) was then computed. This regression provided a prediction of performance for the standardization group on each item-type cluster for each Verbal score level. Using the Verbal mean for each disability group in turn, the predicted cluster scores for a nonhandicapped group of the same total ability was calculated. The predicted cluster mean for the nonhandicapped group was then subtracted from the handicapped group's actual cluster mean, yielding a positive residual if the disabled students did better than the reference group and a negative one when performance was worse than predicted. Finally, this residual was divided by the cluster standard deviation for the disability group. A meaningful departure from the expected difficulty of the cluster was said to exist for a group if the standardized residual exceeded an absolute value of .2 standard deviations on both SAT forms. This .2 standard deviation criterion has been previously suggested as a minimum for identifying the presence of meaningful effects in the social sciences (Cohen, 1969).

Previous research and clinical findings raise the possibility that some disability groups experience unusual difficulty on selected Verbal item types. For example, some studies have found items associated with lengthy passages to be differentially difficult for deaf students (Rudner, 1978; Trybus & Buchanan, 1973, in Rudner, 1978). As SAT reading comprehension items are of this type, unusually poor performance on this subtest--that is, with respect to nonhandicapped students achieving the same total score--might be expected. Vocabulary items also are reported to be difficult for these students (Ragosta & Kaplan, in press), as well as for those with learning disabilities. Learning disabled pupils are said to have particular difficulty with antonyms and with the logical relationships required by verbal analogies (Wiig & Semel, 1973, 1974, 1975, in Wiig & Semel, 1976). Finally, analogies have been found to be differentially difficult for other special populations, in particular black examinees (Dorans, 1982; Kulick, 1984).

Table 10 presents standardized residuals for each disability group's performance on the four Verbal item types. As can be seen, most values are below .1 standard deviations in magnitude and no value exceeds .2 standard deviations on both forms. The pair of values that comes closest to the .2 criterion is for hearing impaired students on Sentence Completion, an item type that might prove somewhat differentially difficult because of the syntactic complexity of the constructions occasionally used.

Insert Table 10 about here

For the Mathematical section, standardized residuals for the item clusters were calculated in a way similar to that for Verbal with two exceptions. First, the cluster scores of nonhandicapped examinees were regressed on total Mathematical (instead of total Verbal) score to obtain a prediction of expected cluster performance for the disability groups. Second, several overlapping cluster structures were tested based on the presence of graphical material, content, and reading load. More than one structure was tested because Mathematical items appear, at least on the surface, to require a broader constellation of basic skills for solution, thereby allowing more room for bias. For example, in addition to reasoning ability, some Mathematical items require the visual skills needed to read graphs, tables, or special symbols, or to manipulate figures in space; for visually impaired examinees, these items may be more a measure of visual-spatial than math reasoning skills. Other items, such as word problems, entail reading. The functioning of these items should be considered suspect for poor readers and for pupils with limited language skills, such as learning disabled and deaf examinees.

For the first analysis, four clusters based on the presence or absence of graphics were used. To form clusters, items were first split into standard multiple choice and quantitative comparison item types. Each of these groups was divided into graphics (i.e., items including tables or figures) and nongraphics, or text, items to form the following clusters: text multiple choice, text comparisons, graphics multiple choice, and graphics comparisons. Items were considered to involve graphics only if a graphic was actually presented.

Standardized residuals for these clusters are presented in Table 11. While, most of the standardized residuals fall between -.1 and .1 standard deviations, striking difficulty effects for visually impaired-braille students on the graphics multiple choice cluster are apparent. The results for these students on graphics comparisons are less consistent, with WSA5 showing a large differential difficulty effect and WSA3 evidencing an effect just below the .2 criterion. Investigation of the items suggests that the type of graphics used for this cluster on WSA3 are less complex and diverse than those used on WSA5.

Insert Table 11 about here

The second cluster structure investigated involved eight item groups primarily based on test content. Again, items were split into multiple choice and quantitative comparisons. These divisions were then separated into arithmetic, algebra, geometry, and miscellaneous sets. Because the resulting miscellaneous comparisons set included only 2 items on one form and one item on the other, this cluster was dropped from the analysis.

Standardized residuals for the seven content clusters are presented in Table 12. As inspection of the table bears out, the residuals for this cluster structure generally appear larger than those for the previous one. Still, the .2 criterion on both forms is exceeded only three times. Among those exceeding the criterion, the algebra comparisons cluster appears unexpectedly *easy* for learning disabled-cassette (LDC) and for hearing impaired-regular (HIR) examinees. Similar, though insignificant, effects are also found for the other groups on this cluster. One factor that may be contributing to this finding is that the cluster is disproportionately loaded with late-appearing items, items that those taking extended-time administrations would be more likely to reach. Of the six items on WSA3, two are at the end of the 35 question section (#32 and 34), while two of five items are at the close of WSA5 (#32 and 35).

Insert Table 12 about here

In addition to the significant effect for the algebra comparisons cluster, miscellaneous multiple choice items were found to be unexpectedly difficult for visually impaired-braille students (VIB). Analysis of item content for this cluster suggests that it is composed of a collection of items that may prove unexpectedly difficult for different reasons. Among the potential sources of differential difficulty are items that utilize novel symbols, assess concepts (e.g., probability) often taught using graphics (e.g., Venn diagrams), assume clear translation to braille of and facility with visually-based symbol systems (e.g., the tally system), or require skill in mentally manipulating figures in space.

The final Mathematical cluster structure examined involved three groupings based on reading load: nonreading, minimal reading, and reading. Items were placed in the reading category if they contained more than one line of text in the stem or response options. Minimal reading items were those with approximately one line of text or less, while nonreading items contained no words, only mathematical symbols and numerals. Written directions at the beginning of each of the two Mathematical sections were not included in the analysis as the amount of reading entailed was constant for all items.

Table 13 presents standardized residuals for the reading load clusters. No consistent effects are found, except for hearing impaired-regular students (HIR) who find the nonreading cluster unexpectedly easy. Again, a contributing factor mav be the disproportionate loading of this cluster with late-appearing items. This explanation is consistent with the effect sizes: in the WSA3 cluster, three of eight items are at the end of the test section and an effect of .36 standard deviation units is found, while the WSA5 cluster has fewer late-appearing items (three of 13) and a much smaller effect (.2 standard deviations).

Insert Table 13 about here

A second possible explanation for this effect is that hearing impaired students perform better on this cluster because it is comparatively free of language. This explanation would be supported by the discovery of difficulty effects for this group on the reading cluster, which contains a fair amount of language. Since such effects are not uniformly apparent, the explanation may not be wholely satisfactory.

With the possible exception of deaf students, then significant difficulty effects for reading load are not evident. This finding is encouraging, especially for learning disabled examinees who generally possess reading and language deficits. For such groups, these results imply that, with extended-time and other relevant special modifications (e.g., cassette presentation), the reading load associated with Mathematical items is light enough to avoid interfering with measurement of the underlying mathematical reasoning ability presumedly tapped by the test.

In sum, the analysis of item clusters has identified five consistent effects of a magnitude large enough to warrant closer study. All identified effects are associated with the Mathematical section of the SAT. The negative effects--that is, those indicating unexpected difficulty--are concentrated among visually impaired-braille students and evidence themselves on the graphics multiple choice and miscellaneous multiple choice item clusters. For the former cluster, the effect was hypothesized as being due to the presence of complex graphics, tables, and figures which measured basic visual-spatial skills in addition to mathematical reasoning ability. For the latter cluster, a conglomeration of factors, including unfamiliar symbols and operations requiring visual-spatial skills, were posed as sources of differential performance.

Positive effects--denoting that the associated clusters were unexpectedly easy--were found for the hearing impaired-regular and learning disabled-cassette groups on algebra comparisons, and for the hearing impaired group on the nonreading cluster. In all three cases, the effects were suggested to be the result of a methodological artifact: the disproportionate presence of late-appearing items in a cluster.

Individual items. The identification of item clusters can be considered the first, or screening, stage in a two-tiered procedure for detecting broad item classes that appear to operate differently for handicapped and nondisabled populations. Therefore, after screening the Verbal and Mathematical item clusters and identifying groupings that appeared to operate differently, a second, more focused methodology was applied. This methodology was designed to detect individual items that seemed to contribute significantly to the finding of differential difficulty for the cluster. In so doing, the methodology indicated the extent to which cluster effects were due to a few isolated items or, alternatively, to the preponderance of items composing the type. In addition, the methodology was meant to provide rigorous statistical tests of the implicit assumption that the relationship between total score and the probability of passing a given item is the same in the standard and handicapped populations.

To accomplish these goals, logistic regression was used to analyze performance on those items composing the clusters identified as differentially difficult or easy for a handicapped group. Within each identified cluster, the item performance of the handicapped group was contrasted with the standardization population (i.e., nonhandicapped students taking WSA3 or WSA5) to determine if the expectations of passing a given item (conditioned on total test score) were equivalent across groups. In addition, logistic regression was used to compare the equality of the slopes of an item performance on total test score for handicapped and nondisabled groups. This latter comparison indicated the extent to which an item evidenced differential operation as a function of ability level (e.g., no differential operation for low-scoring handicapped examinees but differential difficulty for high-scoring ones).

More formally, in the standardization population, _c_, the probability, P, of passing item _i_ given a total score X = X' is:

$$P_{ci} = P(x = 1 \mid X = X')$$

where _x_ is a 0 or 1 score obtained on item _i_. Similarly, in any given handicapped group, _h_,:

$$P_{hi} = P(x_i = 1 \mid X = X')$$

The question to be answered is whether:

$$P_{ci} - P_{hi} \neq 0?$$

The logistic regression model first estimates the unknown regression parameters in the following equation:

$$\log (P / (1 - P)) = B_0 + B_1 D + B_2 X \qquad (1)$$

where P is the probability of passing a given item, D is a dummy variable indicating whether an individual is in the standardization or handicapped groups, and X is the total test score.

Given maximum likelihood estimates of the unknown regression parameters B(0), B(1), and B(2), the expected probability of a standardization group student passing item _i_ is:

$$\hat{P}_{ci} = 1 / 1 + e^{-B_0}$$

and the expected probability of a handicapped group student passing item _i_ is:

$$\hat{P}_{hi} = 1 / 1 + e^{-(B_0 + B_1)}$$

Tests of the equivalence of slopes are carried out by adding a cross-product term to equation (1).

Table 14 presents results for the visually impaired-braille (VIB) and standardization (NHF) groups on items belonging to the graphics multiple choice cluster. For each item in the cluster, the probabilities of passing for each group, the difference in those probabilities, and the presence of an interaction effect are listed. Differential operation was said to exist when the logistic regression coefficient (B1) was significantly different from zero or when tests of the equivalence of slopes indicated significant differences. Whether the item was unexpectedly easy or hard is indicated by the presence or absence of a negative sign in the difference column; a negative difference in the probability of passing an item indicates that the item was unexpectedly difficult for the VIB group.

Because of the large sample sizes in the standardization populations, relatively trivial differences in probabilities are often significant. It is, therefore, suggested that differences in probabilities be at least .1 before a statistically significant result is considered practically meaningful. Unfortunately, for interaction effects, no criterion for practical importance can be easily derived. Therefore, when significant, these effects may indicate only the most minimal deviations and, hence, should be interpreted with caution.

Insert Table 14 about here

As the table indicates, six of the ten items in the WSA3 cluster show statistically significant effects: four main effects and two interactions. Of the items showing main effects, one far exceeds the .1 difference criterion and two approach it. In the WSA5 cluster, two of eight main effects are significant, as are two interactions. The two items with main effects approach, but do not reach, the .1 difference level.

The item evidencing the greatest difference in the probabilities of passing on the two forms (II6) requires the examinee to choose from among five options the size of one of several angles resulting from the intersections of a series of lines, given information about the relationship between the lines and the sizes of related angles (see Figure 3a). While the text of this item contains several special symbols (two to label lines and one denoting the parallel relationship of the lines), definitions for all symbols are provided either in the test directions or in the accompanying figure. Further, other items which use similar notations do not evidence difficulty effects. Hence, the specific content responsible for the observed effect is not immediately clear. This failure to identify the likely cause of differential operation is, unfortunately, a common occurrence in item performance studies (Scheuneman, 1982).

Insert Figure 3 about here

Of the items approaching the .1 criterion, one requires the mental rotation of two graduated dials, one embedded within the other; two involve determining the area of a figure; and one computing the length of a line given intermediate distances. The graduated cylinder item, in particular, may require cognitive-spatial skills that are less well-developed in visually impaired examinees.

Table 15 presents results for the performance of visually impaired-braille (VIB) and nonhandicapped students (NHF) on the miscellaneous multiple choice cluster. On WSA3, five of six items show statistically significant main effects, two of which also show interaction effects. Of the five significant items, one far exceeds the practical criterion and one approaches it. For WSA5, three of six differences are significant, with only one item achieving the criterion for practical importance. (One of these three items [I3] was included above as significant in the graphics multiple choice cluster.)

Insert Table 15 about here

The item showing the largest difficulty effect across both forms (I6) is presented in Figure 3b. This item asks the examinee about the tally system. In this system, the number five is represented by four vertical lines crossed by a diagonal, seven is denoted by the symbol for five followed by a group of two vertical lines, 10 is shown by two symbols for five, and so on. One plausible cause for the observed difficulty effect is that this symbol system is less familiar to blind students. A second probable contributing factor is that the versions presented to blind and sighted students were slightly different. Because the print symbol for five (i.e., four lines crossed by a diagonal) could not be represented easily as a raised line drawing within the braille text of the item, it was denoted by a group of five uncrossed braille symbols for the letter "l". To reflect this modification, the text of the item was changed from, "How many uncrossed tallies would be used in the representation of 29 in this system," to the somewhat more complex, "How many tallies not in sets of five would be used in the representation of 29 in this system?" (emphasis in original). The added linguistic complexity of this modification, along with the novelty of the tally system, are likely causes of the observed differential difficulty for blind students.

The other item reaching criterion (I10) is unexpectedly <u>easy</u> for VIB students. This item (see Figure 3c) requires the examinee to choose from among five options the set of travel directions that would produce the same result as a given sequence. The spatial skills required by this task may be similar to those used by blind students in memorizing directions and in forming mental representations of frequently-used physical environments (e.g., paths, rooms, buildings). It is possible that blind individuals have developed such skills to a greater degree than sighted peers of equal math reasoning ability, thus accounting for their unexpectedly high performance on this item.

Presented in Table 16 are results for the performance of hearing impaired-regular (HIR) and nondisabled students (NHF) students on the nonreading cluster. Most effects for this group are positive, a result consistent with the finding that this item grouping was differentially easy for these examinees. On WSA3, three of eight items show significant main effects, with two of these three also evidencing interactions. None of the significant items approaches the .1 practical criterion. On WSA5, seven of 13 items are significant: five items show main effects, one both a main and interaction, and one an interaction effect. None of the main effects comes reasonably close to .1. For WSA3, the significant effects are associated with items appearing at the end of a section, a finding consistent with the hypothesis that this cluster was easier for hearing impaired students because extended time permitted them to reach these items in greater proportions than their nonhandicapped peers. However, though some items at the close of the section on WSA5 also show significant effects, so do several other items placed earlier in the test, suggesting that something other than, or in addition to, timing is responsible for the differential performance of this group.

Insert Table 16 about here

Performance results for hearing impaired-regular (HIR) and nonhandicapped (NHF) students on the algebra comparisons cluster are given in Table 17. Again, as expected, most effects are positive. Three of six items show main effects on WSA3, with one also displaying an interaction. (Two of these three were noted as significant in the discussion of the nonreading cluster.) On WSA5, two of five items (both of which also appear in the nonreading cluster), are significant; one of these items also exhibits an interaction. None of the two significant main effects comes reasonably close to the .1 practical criterion. Again, significant items appear at the end of test sections and in earlier locations, suggesting that extra time alone is not a sufficient explanation for differential operation.

Insert Table 17 about here

Table 18 presents performance results for learning disabled-cassette (LDC) and nonhandicapped (NHF) students on this same subtest. Two of the six items on WSA3 and two of the five on WSA5 show effects, all of which are positive. In addition, one significant interaction appears on each form. The four items showing main effects are the same as those that showed positive effects for hearing impaired-regular examinees. Again, however, none of the effects approximates the .1 criterion and no consistent clustering at the end of test sections is apparent.

Insert Table 18 about here

In summary, the analysis of individual items composing errant clusters has produced several results. First, of the 61 item performances studied, 34 were statistically significant: 22 performances exhibited only main effects, five showed both main effects and interactions, and seven only interaction effects. Of the main effects, only three were of a magnitude large enough to be considered practically meaningful. Two of these three items were differentially difficult and one differentially easy. The deviant operation of all three items was associated with visually impaired students taking the braille version of the SAT.

Second, the small number of unequivocally deviant items discovered for visually impaired students taking braille tests suggests that graphics multiple choice and miscellaneous items are not generally inappropriate for this group; several items in these clusters appeared to operate equivalently for visually impaired and nondisabled students. Rather, selected items falling within these broad classes may be inappropriate because they appear to measure constructs other than mathematical reasoning ability. Such items may present unfamiliar symbol systems (e.g., the tally item), add linguistic complexity as a result of modified translations, or require cognitive-spatial operations that are not easily performed by blind students and which are only tangentially related to mathematics reasoning (e.g, the graduated cylinder item).

Last, the analysis suggests that the .2 criterion used for cluster screening was relatively sensitive. Several clusters exceeding the criterion were found to be composed of items evidencing minimal effects (e.g., WSA3 Algebra Comparisons). Even for those clusters far exceeding the .2 criterion, only a few isolated instances of differential item performance were detected.

Summary and Recommendations

This study has investigated the psychometric characteristics of special administrations of the SAT for nine handicapped groups. Data on these characteristics are central to evaluating the accuracy of scores for measuring the developed scholastic abilities of disabled examinees.

With respect to level of performance, the first characteristic described, handicapped groups varied widely. In general, visually-impaired students and those with physical handicaps achieved mean scores comparable to students taking the SAT in national administrations. In contrast, learning disabled and hearing impaired students performed more poorly than the general SAT-taking population, usually by at least a half standard deviation. In addition, most groups showed differences between Verbal and Mathematical scores that were comparable to the reference population, with the exception of hearing impaired-regular students who performed relatively better on Mathematical than Verbal, and visually impaired-braille students, who showed no consistent superiority for the Mathematical over the Verbal scale.

In contrast to level of performance, the reliability of the SAT was found to be comparable to the reference population for all handicapped groups. This finding suggests that one potential source of unfairness, differences in measurement precision, probably need not be of practical concern.

To ensure that the time extensions allowed in special administrations are enough to permit disabled students to complete the same proportion of the test as their nonhandicapped peers, a third psychometric characteristic, test speededness, was examined. With respect to the reference sample, no disability group was found to be disadvantaged by lack of time, thus suggesting that another possible source of unfairness is probably of little import.

The final psychometric characteristic studied was unexpected differential performance. Investigation of this characteristic was conducted to identify potentially biased item types, types that may not measure the ability assessed by the overall test. Differential performance was evaluated through a two-stage procedure in which the operation of groups of items was first investigated. Five item groupings, or clusters, were identified by this procedure as potentially problematic. The individual items in these clusters were then subjected to a more rigorous analysis to discover whether these broad item classes, or only isolated items, were responsible for cluster effects. This analysis identified only three items, all for visually impaired students taking the braille version, that showed clear evidence of idiosyncratic operation.

The localization of idiosyncratically operating items to visually impaired students taking the braille exam suggests that extra care be taken in the development and translation to braille of SAT forms used by the Admissions Testing Program's Services for Handicapped Students. In addition, the possibility should be considered of pilot testing brailled exams before these tests are put into actual service. The aim of such testing would be to detect any items tapping inappropriate skills, confusing instructions, errors in brailling, or other remaining irrelevant sources of difficulty. Pilot testing need not be carried out with large numbers of examinees. More desirable would be individual or small-group administrations in which examinees could discuss potential difficulties as they arise directly with test development staff. Finally, as an additional check on the success of the test development and brailling processes, periodic analyses of the operation of items on the braille exam should be considered.

In contrast to visually impaired-braille examinees, no items showing practically important indications of differential performance were found for hearing impaired-regular or learning disabled-cassette students. In addition, the large majority of effects that were detected for these two groups were positive, suggesting no negative impact on total score.

In conclusion, with the exception of performance level, the psychometric characteristics of the SAT forms studied appear to be largely comparable for the disabled and nonhandicapped groups taking part in this investigation. This result should extend to other forms of the SAT and other disabled students to the extent that these groups and forms, and the conditions under which they are administered, are similar to those employed in the study. That the psychometric characteristics of the test are similar across populations provides some of the evidence necessary to support SAT scores as accurate and fair indicators of the developed scholastic abilities of disabled students. Further evidence from factor analyses and predictive validity studies should add knowledge about the meaning of these scores for handicapped examinees.

References

Bennett, R. E., Ragosta, M., & Stricker, L. J. (1984). The test performance of handicapped people (RR 84-32). Princeton, NJ: Educational Testing Service.

Berk, R. A. (1982). Introduction. In R. A. Berk (Ed), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Breland, H. M. (1978). Population validity and college entrance measures (RB-78-19). Princeton, NJ: Educational Testing Service.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

College Board. (1983). Taking the SAT. New York: Author.

Cook, L., Petersen, N., & Ervin, N. (1980). College Board Admissions Testing Program statistical summary for academic year 1979-80. Princeton, NJ: Educational Testing Service.

Cook, L., Petersen, N., & Jacob, E. (1981). College Board Admissions Testing Program statistical summary for academic year 1980-81. Princeton, NJ: Educational Testing Service.

Cook, L., Petersen, N., & Flesher, R. (1982). College Board Admissions Testing Program statistical summary for academic year 1981-82. Princeton, NJ: Educational Testing Service.

Cook, L., Petersen, N., & Zicha, M. (1983). College Board Admissions Testing Program statistical summary for academic year 1982-83. Princeton, NJ: Educational Testing Service.

Cook, L., Dorans, N., & Flesher, R. (1984). College Board Admissions Testing Program statistical summary for academic year 1983-84. Princeton, NJ: Educational Testing Service.

Donlon, T. F. (Ed). (1984). The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board.

Dorans, N. J. (1982). Technical review of SAT item fairness studies: 1975-1979 (SR-82-90). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1980). Test use and validity. Princeton, NJ: Author.

Kulick, E. (1984). Assessing unexpected differential item performance of black candidates on SAT Form CSA6 and TSWE Form E33 (SR-84-80). Princeton, NJ: Educational Testing Service.

Meadow, K. P. (1980). Deafness and child development. Berkeley: University of California Press.

Ragosta, M., & Kaplan, B. A. (In press). A survey of handicapped students taking special test administrations of the SAT and GRE. Princeton, NJ: Educational Testing Service.

Rock, D. A., Bennett, R. E., & Kaplan, B. A. (In press). The internal construct validity of the SAT across handicapped and nonhandicapped populations. Princeton, NJ: Educational Testing Service.

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

Saretsky, G. (1983). SATs for the blind offered 45 years ago. Examiner, 13(7), 3.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Sherman, S. W., & Robinson, N. M. (Eds). (1982). Ability testing of handicapped people: Dilemma for government, science, and the public. Washington, D. C.: National Academy Press.

Sinnott, L. T. (1980). Differences in item performance across groups (RR 80-19). Princeton, NJ: Educational Testing Service.

Stern, J. (1978). College Board item bias study of the Scholastic Aptitude Test and the Test of Standard Written English Form XSA2/E4 (SR-78-56). Princeton, NJ: Educational Testing Service.

Trybus, R. J , & Buchanan, C. (1971). Patterns of achievement test performance. In R. J. Trybus, C. Buchanan, & S. DiFrancesca (Eds), Studies in achievement testing hearing impaired students, United States: Spring 1971. Washington, D. C.: Office of Demographi Studies, Gallaudet College.

Wiig, E. H., & Semel, E. M. (1973). Comprehension of linguistic concepts requiring logical operations. Journal of Speech and Hearing Research, 16, 627-36.

Wiig, E. H., & Semel, E. M. (1974). Logico-grammatical sentence comprehension by adolescents with learning disabilities. Perceptual and Motor Skills, 38, 1331-34.

Wiig, E. H., & Semel, E. M. (1975). Productive language abilities in learning disabled adolescents. Journal of Learning Disabilities, 8, 578-86.

Wiig, E. H., & Semel, E. M. (1976). Language disabilities in children and adolescents. Columbus, OH: Charles Merrill.

Table 1a

Numbers of Students Taking Each Type of CAT Special Administration for WSA3

				Group[a]			
Exam Type	VI	PH	HI	LD	Multiple	Unknown	Total
Braille	98	0	1	2	0	1	102
Large-type	486	30	6	185	18	1	726
Cassette	27	2	1	107	3	0	140
Regular	223	346	287	2[illegible]83	27	23	3889
Cassette & large type	29	4	0	23	4	1	61
Braille & cassette	5	1	0	0	0	0	6
Cassette & regular	16	1	1	192	1	0	211
Unknown	9	6	1	60	1	1	78
Total	893	390	297	3552	54	27	5213

[a] VI = visually impaired, PH = physically handicapped, HI = hearing impaired, LD = learning disabled.

Table 1b

Numbers of Students Taking Each Type

of SAT Special Administration for WSA5

Exam Type	Group[a] VI	PH	HI	LD	Multiple	Unknown	Total
Braille	105	1	0	1	0	0	107
Large-type	498	16	5	136	15	6	676
Cassette	11	0	0	113	2	0	126
Regular	175	230	150	2316	29	24	2924
Cassette & large type	27	0	0	25	1	0	53
Braille & cassette	21	1	0	1	0	0	23
Cassette & regular	12	1	0	253	4	1	271
Unknown	9	5	4	38	0	0	56
Total	858	254	159	2883	51	31	4236

[a] VI = visually impaired, PH = physically handicapped, HI = hearing impaired, LD = learning disabled.

Table 2

Sample Sizes and Acronyms Used
to Denote Disability Groups

Acronym	Disability Group	WSA3 Sample Size	WSA5 Sample Size
HIR	Hearing impaired students taking regular-type edition	287	150
LDC	Learning disabled students taking cassette edition	107	113
LDCR	Learning disabled students taking cassette and regular-type editions	192	253
LDL	Learning disabled students taking large-type edition	185	136
LDR	Learning disabled students taking regular-type edition	2983	2316
PHR	Physically handicapped students taking regular-type edition	346	230
VIB	Visually impaired students taking braille edition	98	105
VIL	Visually impaired students taking large-type edition	486	498
VIR	Visually impaired students taking regular-type edition	223	175

Table 3

Performance of Nonhandicapped Students on WSA3 and WSA5 Relative to Students Taking the SAT from 3/80 to 6/83

WSA3

Group	Verbal	Mathematical
Seniors taking form		
Mean	448	493
SD	(108)	(116)
Seniors taking SAT from 3/80-6/83[a]		
Mean	413	454
SD	(104)	(112)

WSA5

Group	Verbal	Mathematical
Juniors taking form		
Mean	424	459
SD	(107)	(113)
Juniors taking SAT from 3/80-6/83[a]		
Mean	442	489
SD	(103)	(112)

[a] Calculated from statistics presented in College Board Admissions Testing Program Statistical Summaries (Cook, Petersen, & Ervin, 1980; Cook, Petersen, & Jacob, 1981; Cook, Petersen, & Flesher, 1982; Cook, Petersen, & Zicha, 1983; Cook, Petersen, & Dorans, 1984).

Table 4

The SAT Performance of Nine Disability Groups

Verbal Scaled Scores

WSA3			WSA5		
Group[a]	Mean	SD	Group	Mean	SD
NHA	424	106	VIR	436	104
PHR	423	112	VIB	434	134
VIB	412	127	VIL	433	111
VIR	401	101	PHR	432	107
VIL	400	110	NHA	424	106
LDR	370	97	LDR	376	96
LDCR	351	81	LDL	366	96
LDC	349	86	LDCR	350	85
LDL	349	91	LDC	328	82
HIR	284	91	HIR	326	103

Mathematical Scaled Scores

WSA3			WSA5		
Group	Mean	SD	Group	Mean	SD
NHA	468	114	VIR	491	133
VIR	456	135	NHA	468	114
PHR	434	131	VIL	468	128
VIL	431	129	PHR	460	116
LDR	411	121	VIB	438	133
LDCR	378	98	LDR	412	111
VIB	376	113	HIR	407	111
LDL	374	105	LDL	391	95
HIR	373	116	LDCR	374	93
LDC	365	101	LDC	360	86

[a] NHA denotes all students taking forms of the SAT administered between 3/80 and 6/83. Scores for this group calculated from statistics presented in College Board Admissions Testing Program Statistical Summaries (Cook, Petersen, & Ervin, 1980; Cook, Petersen, & Jacob, 1981; Cook, Petersen, & Flesher, 1982; Cook, Petersen, & Zicha, 1983; Cook, Petersen, & Dorans, 1984).

Table 5

Disabled Student SAT Performance in
SD Units from the Nonhandicapped Student Mean[a]

Verbal

Group	WSA5 Difference	WSA3 Difference	Weighted Average
PHR	0.08	-0.01	0.02
VIB	0.09	-0.11	-0.01
VIR	0.11	-0.22	-0.07
VIL	0.08	-0.23	-0.07
LDR	-0.45	-0.51	-0.48
LDL	-0.55	-0.71	-0.64
LDCR	-0.70	-0.69	-0.69
LDC	-0.91	-0.71	-0.81
HIR	-0.92	-1.32	-1.18

Mathematical

Group	WSA5 Difference	WSA3 Difference	Weighted Average
VIR	0.20	-0.11	0.03
VIL	0.00	-0.32	-0.16
PHR	-0.07	-0.30	-0.21
LDR	-0.49	-0.50	-0.50
VIB	-0.26	-0.81	-0.53
HIR	-0.54	-0.83	-0.73
LDL	-0.68	-0.82	-0.74
LDCR	-0.82	-0.79	-0.81
LDC	-0.95	-0.90	-0.93

[a] Nondisabled students are all examinees taking the SAT from 3/80 to 6/83. Differences are expressed in SD units of the nonhandicapped group.

Table 6

Difference Between SAT Verbal and Mathematical Scores for Handicapped Students[a]

Group	WSA3 Difference Index	WSA5 Difference Index
HIR	2.14	1.89
VIR	1.15	1.15
LDR	0.94	0.87
LDCR	0.75	0.67
VIL	0.65	0.73
LDL	0.64	0.65
LDC	0.43	0.95
PHR	0.23	0.63
VIB	-0.75	0.07

[a] Difference index is the ratio of the difference between Verbal and Mathematical mean scaled scores for a handicapped group divided by the pooled standard deviation for those scores to the same quant.-y calculated for all students taking the SAT betwee 3/80 and 6/83. A difference index of +1 indicates intra-test performance equivalent in magnitude and direction to the reference group.

Table 7

Differences in SD Units Between Scaled Score Means of Disabled Student Groups taking WSA3 and WSA5[a]

Group	Verbal Difference	Math Difference
HIR	.44 ***	.30 **
VIR	.34 ***	.26 ***
VIL	.30 ***	.29 ***
LDL	.18	.17
VIB	.17	.50 *
PHR	.08	.21 *
LDR	.06 *	.01
LDCR	-.01	-.04
LDC	-.25	-.05

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

[a] Differences are calculated by subtracting the WSA3 mean from the WSA5 mean for a handicapped group and dividing by the pooled Verbal or Mathematical standard deviation for that group. Significance of differences was tested using the two-tailed t-test.

Table 8

SAT Reliability for Disability Groups

Verbal Section

Group[a]	Alpha Reliability WSA3	Alpha Reliability WSA5	SE Measurement[b] WSA3	SE Measurement[b] WSA5
VIB	.93	.95	3.87	3.74
NHF	.92	.92	3.73	3.75
VIL	.91	.92	3.82	3.74
PHR	.91	.91	3.83	3.75
VIR	.90	.91	3.79	3.77
LDR	.89	.90	3.80	3.77
HIR	.88	.91	3.81	3.79
LDL	.87	.89	3.84	3.80
LDC	.86	.84	3.76	3.74
LDCR	.85	.86	3.82	3.81

Mathematical Section

Group[a]	Alpha Reliability WSA3	Alpha Reliability WSA5	SE Measurement[b] WSA3	SE Measurement[b] WSA5
VIR	.94	.94	3.11	3.07
VIB	.93	.94	3.08	3.07
VIL	.93	.93	3.15	3.11
PHR	.93	.92	3.13	3.17
LDR	.93	.92	3.11	3.14
HIR	.92	.91	3.15	3.16
NHF	.92	.91	3.09	3.15
LDL	.91	.89	3.11	3.14
LDCR	.91	.89	3.11	3.13
LDC	.90	.86	3.08	3.12

[a] NHF = nonhandicapped students taking WSA3 or WSA5.

[b] Standard errors of measurement are in raw score units.

Table 9

SAT Speededness for Disability Groups Compared with

Nonhandicapped Students Taking the Same Test Form

Verbal Section I

Group	Ratio of Percent Completing Section[a] WSA3	Ratio of Percent Completing Section[a] WSA5	Ratio of Percent Completing 75% of Section[a] WSA3	Ratio of Percent Completing 75% of Section[a] WSA5
LDC	1.23	1.34	1.00	1.01
PHR	1.23	1.29	1.00	1.01
VIR	1.22	1.34	1.00	1.01
VIL	1.22	1.34	1.00	1.01
LDR	1.19	1.32	1.00	1.01
LDL	1.19	1.32	1.00	1.01
LDCR	1.16	1.26	1.00	1.01
VIB	1.15	1.26	1.00	1.01
HIR	1.14	1.33	1.00	1.01

Verbal Section II

Group	Ratio of Percent Completing Section WSA3	Ratio of Percent Completing Section WSA5	Ratio of Percent Completing 75% of Section WSA3	Ratio of Percent Completing 75% of Section WSA5
LDC	1.09	1.23	1.03	1.03
PHR	1.23	1.08	1.03	1.03
VIR	1.04	1.12	1.03	1.03
VIL	1.19	1.09	1.03	1.03
LDR	.98	.95	1.03	1.03
LDL	1.00	1.03	1.03	1.03
LDCR	1.05	1.03	1.03	1.02
VIB	1.44	1.14	1.02	1.03
HIR	1.25	1.21	1.03	1.02

[a] Ratio is the percentage of disabled students divided by the equivalent value for nondisabled students. Values above 1.00 indicate a higher percentage of disabled than nondisabled students completing the section or part section.

Table 9 (con't)

SAT Speededness for Disability Groups Compared with Nonhandicapped Students Taking the Same Test Form

Mathematical Section I

Group	Ratio of Percent Completing Section[a] WSA3	Ratio of Percent Completing Section[a] WSA5	Ratio of Percent Completing 75% of Section[a] WSA3	Ratio of Percent Completing 75% of Section[a] WSA5
VIL	1.00	1.13	1.00	1.08
HIR	.99	1.20	.99	1.08
LDC	.97	.96	.99	1.03
VIR	.96	1.23	1.00	1.07
LDR	.91	1.01	.99	1.04
LDCR	.91	.99	.98	1.01
PHR	.89	1.19	1.00	1.05
LDL	.88	.91	1.00	1.07
VIB	.88	1.00	.99	1.03

Mathematical Section II

Group	Ratio of Percent Completing Section WSA3	Ratio of Percent Completing Section WSA5	Ratio of Percent Completing 75% of Section WSA3	Ratio of Percent Completing 75% of Section WSA5
VIL	1.[illegible]2	1.22	1.02	1.03
HIR	1.88	1.20	1.02	1.02
LDC	1.67	1.04	1.02	.94
VIR	1.94	1.20	1.02	1.04
LDR	1.73	1.16	1.02	1.01
LDCR	1.69	1.09	1.02	.98
PHR	1.84	1.21	1.02	1.02
LDL	1.71	1.13	1.02	.99
VIB	1.76	1.17	1.02	1.00

[a] Ratio is the percentage of disabled students divided by the equivalent value for nondisabled students. Values above 1.00 indicate a higher percentage of disabled than nondisabled students completing the section or part section.

Table 10

Extent of Unexpected Differential Performance in SD Units on SAT Verbal Item Clusters[a]

	Antonyms (n = 25)		Analogies (n = 20)	
Group	WSA3	WSA5	WSA3	WSA5
VIR	.00	.00	.00	.00
VIL	.01	.01	-.03	-.04
VIB	.05	.06	-.05	-.17
PHR	.10	.01	-.04	-.06
LDR	-.04	-.06	-.01	-.08
LDL	.02	.04	.02	-.13
LDCR	-.01	-.02	-.12	-.02
LDC	-.02	-.08	.00	-.07
HIR	-.07	-.04	.10	.05

	Sentence Completion (n = 15)		Reading Comprehension (n = 25)	
Group	WSA3	WSA5	WSA3	WSA5
VIR	.00	.00	.00	.00
VIL	.05	.07	.05	-.01
VIB	-.05	.07	.12	-.03
PHR	.05	.06	.04	-.02
LDR	.02	-.01	-.05	-.01
LDL	.05	.04	-.15	.01
LDCR	.11	.03	-.01	-.01
LDC	.20	.06	-.10	.11
HIR	-.17	-.16	.11	.10

[a] Tabled values represent the difference between the actual and predicted mean cluster raw scores for each handicapped group divided by that group's cluster standard deviation. Positive values indicate better performance than expected while negative values denote the converse. An absolute value in excess of .2 on both forms is considered practically important.

Table 11

Extent of Unexpected Differential Performance in SD Units on SAT Mathematical Graphics-Load Clusters[a]

Group	Text Multiple Choice WSA3 (n=30)	Text Multiple Choice WSA5 (n=32)	Text Comparisons WSA3 (n=14)	Text Comparisons WSA5 (n=13)
VIR	.00	.00	.00	.00
VIL	.06	.02	-.03	.02
VIB	.07	.06	.00	.05
PHR	.03	.01	.00	-.02
LDR	-.05	-.07	-.07	-.05
LDL	-.04	-.05	-.07	-.08
LDCR	-.10	-.12	-.08	-.06
LDC	-.08	-.11	-.13	-.03
HIR	-.01	-.05	.07	-.07

Group	Graphics Multiple Choice WSA3 (n=10)	Graphics Multiple Choice WSA5 (n=8)	Graphics Comparisons WSA3 (n=5)	Graphics Comparisons WSA5 (n=7)
VIR	.00	.00	.00	.00
VIL	.02	-.06	.06	-.07
VIB	-.31	-.46	-.17	-.49
PHR	-.02	-.15	.11	.04
LDR	.05	.01	.02	.03
LDL	-.02	-.02	-.03	-.05
LDCR	.04	.13	.01	.05
LDC	.07	.14	.15	-.05
HIR	.18	-.01	-.03	.25

[a] Tabled values represent the difference between the actual and predicted mean cluster raw scores for each handicapped group divided by that group's cluster standard deviation. Positive values indicate better performance than expected while negative values denote the converse. An absolute value in excess of .2 on both forms is considered practically important.

Table 12

Extent of Unexpected Differential Performance in SD Units on SAT Mathematical Item-Content Clusters[a]

Group	Arithmetic Multiple Choice WSA3 (n=11)	Arithmetic Multiple Choice WSA5 (n=12)	Algebra Multiple Choice WSA3 (n=12)	Algebra Multiple Choice WSA5 (n=11)
VIR	-.07	-.05	-.13	-.08
VIL	.01	-.02	-.01	.02
VIB	.15	.08	.04	.04
PHR	-.04	-.03	.01	-.01
LDR	-.15	-.14	-.10	-.14
LDL	-.21	-.10	-.04	-.14
LDCR	-.18	-.14	-.14	-.22
LDC	-.22	-.18	-.07	-.24
HIR	-.14	-.03	-.06	-.08

Group	Geometry Multiple Choice WSA3 (n=11)	Geometry Multiple Choice WSA5 (n=11)	Miscellaneous Multiple Choice WSA3 (n=6)	Miscellaneous Multiple Choice WSA5 (n=6)
VIR	-.01	.09	-.10	-.02
VIL	-.01	-.02	-.17	-.02
VIB	-.23	-.17	-.65	-.20
PHR	-.07	-.12	-.22	.04
LDR	.04	.02	-.23	.01
LDL	.05	.00	-.36	-.01
LDCR	.06	.12	-.35	-.08
LDC	.08	.17	-.29	.03
HIR	.18	.02	-.18	-.14

[a] Tabled values represent the difference between the actual and predicted mean cluster raw scores for each handicapped group divided by that group's cluster standard deviation. Positive values indicate better performance than expected while negative values denote the converse. An absolute value in excess of .2 on both forms is considered practically important.

Table 12 (con't)

Extent of Unexpected Differential Performance in SD Units on SAT Mathematical Item-Content Clusters[a]

Group	Arithmetic Comparisons WSA3 (n=6)	Arithmetic Comparisons WSA5 (n=7)	Algebra Comparisons WSA3 (n=6)	Algebra Comparisons WSA5 (n=5)
VIR	.20	.11	.20	.11
VIL	.26	.0[illegible]	.16	.21
VIB	.33	.03	.23	.15
PHR	.28	.02	.22	.18
LDR	.14	.03	.20	.10
LDL	.17	.05	.30	.18
LDCR	.12	.01	.21	.13
LDC	.12	-.06	.27	.29
HIR	.25	-.06	.26	.26

Group	Geometry Comparisons WSA3 (n=6)	Geometry Comparisons WSA5 (n=6)
VIR	.09	-.04
VIL	.10	-.06
VIB	-.09	-.27
PHR	.11	.03
LDR	.03	.00
LDL	-.13	-.11
LDCR	.03	.09
LDC	.03	-.04
HIR	.03	.16

[a] Tabled values represent the difference between the actual and predicted mean cluster raw scores for each handicapped group divided by that group's cluster standard deviation. Positive values indicate better performance than expected while negative values denote the converse. An absolute value in excess of .2 on both forms is considered practically important.

Table 13

Extent of Unexpected Differential Performance

in SD Units on SAT Mathematical Reading Load Clusters[a]

Group	Nonreading WSA3 (n=8)	Nonreading WSA5 (n=13)	Minimal Reading WSA3 (n=22)	Minimal Reading WSA5 (n=13)
VIR	.17	.03	-.05	.07
VIL	.17	.06	-.02	.05
VIB	.23	.04	-.18	-.20
PHR	.19	.06	-.06	.07
LDR	.15	-.01	-.12	.06
IDL	.17	.02	-.23	-.10
LDCR	.15	-.07	-.21	.05
LDC	.11	.02	-.24	-.05
HIR	.36	.20	.01	.23

Group	Reading WSA3 (n=29)	Reading WSA5 (n=34)
VIR	.01	-.01
VIL	.06	-.01
VIB	.00	.00
PHR	.06	-.04
LDR	.01	-.09
LDL	.05	-.06
LDCR	.02	-.07
LDC	12	-.08
HIR	-.03	-.18

[a] Tabled values represent the difference between the actual and predicted mean cluster raw scores for each handicapped group divided by that group's cluster standard deviation. Positive values indicate better performance than expected while negative values denote the converse. An absolute value in excess of .2 on both forms is considered practically important.

Table 14

Performance of Visually Impaired-Braille (VIB) and Nonhandicapped Students (NHF) on Graphics Multiple Choice Items[a]

WSA3

Item	VIB Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
I 3	.12	.16	-.04	
4	.03	.10	-.07 ***	
17	.02	.04	-.02	x **
18	.03	.02	.01	
22	.01	.01	.00	
24	.03	.01	.01	x **
II 5	.08	.05	.03 *	
6	.11	.45	-.34 ***	
7	.06	.14	-.08 **	
12	.01	.01	.00	

WSA5

Item	VIB Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
I 3	.06	.14	-.08 **	
7	.08	.11	-.03	
14	.01	.01	.00	
20	.03	.10	-.07 ***	
24	.02	.02	.00	x ***
II 9	.01	.02	-.01	
12	.03	.04	-.01	
13	.01	.01	.00	x *

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

[a] Performance data are for nonhandicapped students taking forms WSA3 and WSA5.

[b] Differences may not reflect the computed difference between the handicapped and nonhandicapped columns due to rounding error.

Table 15

Performance of Visually Impaired-Braille (VIB) and Nonhandicapped Students (NHF) on Miscellaneous Multiple Choice Items[a]

WSA3

Item	VIB Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
I 3	.12	.16	-.04	
6	.11	.45	-.34 ***	
7	.06	.14	-.08 **	
20	.03	.06	-.03 **	x **
21	.01	.04	-.03 ***	x *
23	.04	.02	.02 **	

WSA5

Item	VIB Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
I 3	.06	.14	-.08 **	
10	.35	.23	.12 *	
15	.04	.03	.01	
23	.01	.02	.00	
25	.01	.02	-.01	
IV 5	.08	.15	-.07 **	

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

[a] Performance data are for nonhandicapped students taking forms WSA3 and WSA5.

[b] Differences may not reflect the computed difference between the handicapped and nonhandicapped columns due to rounding error.

Table 16

Performance of Hearing Impaired-Regular (HIR) and Nonhandicapped Students (NHF) on Nonreading Items[a]

WSA3

Item	HIR Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
II16	.09	.10	-.01	
18	.93	.95	-.02	
20	.05	.04	.01	
22	.97	.96	.01	
25	.08	.06	.02	
30	.07	.02	.05 ***	x **
32	.02	.01	.01 ***	
34	.07	.02	.05 ***	x ***

WSA5

Item	HIR Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
I 2	.12	.13	-.01	
17	.06	.05	.01	
18	.14	.10	.05 *	
21	.03	.01	.02 ***	
II18	.05	.08	-.03 *	
21	.04	.05	-.01	x ***
23	.08	.08	.01	
24	.13	.07	.06 ***	
26	.00	.01	.00	
28	.06	.07	-.01	
31	.11	.07	.05 **	
32	.10	.04	.06 ***	x *
35	.01	.01	.00	

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

[a] Performance data are for nonhandicapped students taking forms WSA3 and WSA5.

[b] Differences may not reflect the computed difference between the handicapped and nonhandicapped columns due to rounding error.

Table 17

Performance of Hearing Impaired-Regular (HIR) and Nonhandicapped Students (NHF) on Algebra Comparisons Items[a]

WSA3

Item	HIR Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
I 18	.93	.95	-.02	
21	.04	.08	-.03 ***	
22	.97	.96	.01	
25	.08	.06	.02	
32	.02	.01	.01 ***	
34	.07	.02	.05 ***	x ***

WSA5

Item	HIR Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
II23	.08	.08	.01	
24	.13	.07	.06 ***	
26	.00	.01	.00	
32	.10	.04	.06 ***	x *
35	.01	.01	.00	

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

[a] Performance data are for nonhandicapped students taking forms WSA3 and WSA5.

[b] Differences may not reflect the computed difference between the handicapped and nonhandicapped columns due to rounding error.

Table 18

Performance of Learning Disabled-Cassette (LDC) and Nonhandicapped Students (NHF) on Algebra Comparisons Items[a]

WSA3

Item	LDC Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
II16	.95	.95	.00	
21	.10	.08	.02	
22	.95	.96	-.01	x *
25	.07	.06	.01	
32	.04	.01	.03 ***	
34	.04	.02	.02 **	

WSA5

Item	LDC Probability of Passing	NHF Probability of Passing	Difference[b]	Interaction
II23	.06	.08	-.01	
24	.10	.07	.04 **	
26	.01	.01	.00	
32	.10	.04	.06 ***	
35	.01	.01	.00	x *

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

[a] Performance data are for nonhandicapped students taking forms WSA3 and WSA5.

[b] Differences may not reflect the computed difference between the handicapped and nonhandicapped columns due to rounding error.

Figure 1

SAT Verbal Item Types

Analogies

Each question below consists of a related pair of words or phrases, followed by five lettered pairs of words or phrases. Select the lettered pair that <u>best</u> expresses a relationship similar to that expressed in the original pair.

Example:

> YAWN:BOREDOM :: (A) dream:sleep (B) anger:madness (C) smile:amusement (D) face:expression (E) impatience:rebellion
>
> Ⓐ Ⓑ ● Ⓓ Ⓔ

36. COW:BARN :: (A) pig:mud (B) chicken:coop (C) camel:water (D) cat:tree (E) horse:racetrack

Antonyms

Each question below consists of a word in capital letters, followed by five lettered words or phrases. Choose the word or phrase that is most nearly <u>opposite</u> in meaning to the word in capital letters. Since some of the questions require you to distinguish fine shades of meaning, consider all the choices before deciding which is best.

Example:

> GOOD: (A) sour (B) bad (C) red (D) hot (E) ugly
>
> Ⓐ ● Ⓒ Ⓓ Ⓔ

1. VERSATILE: (A) unadaptable (B) mediocre (C) impatient (D) egocentric (E) vicious

2. FRAUDULENT: (A) ...her pleasing (B) extremely beneficial (C) courteous (D) authentic (E) simplified

Sentence Completion

Each sentence below has one or two blanks, each blank indicating that something has been omitted. Beneath the sentence are five lettered words or sets of words. Choose the word or set of words that <u>best</u> fits the meaning of the sentence as a whole.

Example:

> Although its publicity has been —, the film itself is intelligent, well-acted, handsomely produced, and altogether —.
>
> (A) tasteless..respectable (B) extensive..moderate (C) sophisticated..amateur (D) risqué..crude (E) perfect..spectacular
>
> ● Ⓑ Ⓒ Ⓓ Ⓔ

16. He claimed that the document was — because it merely listed endangered species and did not specify penalties for harming them.

    (A) indispensable (B) inadequate (C) punitive (D) aggressive (E) essential

Figure 1 (cont'd)

SAT Verbal Item Types

Reading Comprehension

Each passage below is followed by questions based on its content. Answer all questions following a passage on the basis of what is stated or implied in that passage.

Mars revolves around the Sun in 687 Earth days, which is equivalent to 23 Earth months. The axis of Mars's rotation is tipped at a 25° angle from the plane of its orbit, nearly the same as the Earth's tilt of about 23° Because the tilt causes the seasons, we know that Mars goes through a year with four seasons just as the Earth does.

From the Earth, we have long watched the effect of the seasons on Mars. In the Martian winter, in a given hemisphere, there is a polar ice cap. As the Martian spring comes to the Northern Hemisphere, for example, the north polar cap shrinks and material in the planet's more temperate zones darkens. The surface of Mars is always mainly reddish, with darker gray areas that, from the Earth, appear blue green. In the spring, the darker regions spread. Half a Martian year later, the same process happens in the Southern Hemisphere.

One possible explanation for these changes is biological: Martian vegetation could be blooming or spreading in the spring. There are other explanations, however. The theory that presently seems most reasonable is that each year during the Northern Hemisphere springtime, a dust storm starts, with winds that reach velocities as high as hundreds of kilometers per hour. Fine, light-colored dust is blown from slopes, exposing dark areas underneath. If the dust were composed of certain kinds of materials, such as limonite, the reddish color would be explained.

29. It can be inferred that one characteristic of limonite is its
    (A) reddish color
    (B) blue green color
    (C) ability to change colors
    (D) ability to support rich vegetation
    (E) tendency to concentrate into a hard surface

30. According to the author, seasonal variations on Mars are a direct result of the
    (A) proximity of the planet to the Sun
    (B) proximity of the planet to the Earth
    (C) presence of ice caps at the poles of the planet
    (D) tilt of the planet's rotational axis
    (E) length of time required by the planet to revolve around the Sun

31. It can be inferred that, as spring arrives in the Southern Hemisphere of Mars, which of the following is also occurring?
    (A) The northern polar cap is increasing in size.
    (B) The axis of rotation is tipping at a greater angle.
    (C) A dust storm is ending in the Southern Hemisphere.
    (D) The material in the northern temperate zones is darkening.
    (E) Vegetation in the southern temperate zones is decaying.

Source: College Board (1983). Taking the SAT. New York: Author

BEST COPY AVAILABLE

Figure 2

SAT Mathematical Item Types

Multiple Choice

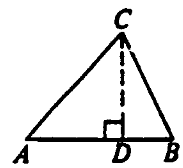
In this section solve each problem, using any available space on the page for scratchwork. Then decide which is the best of the choices given and blacken the corresponding space on the answer sheet.

The following information is for your reference in solving some of the problems.

Circle of radius r: Area $= \pi r^2$; Circumference $= 2\pi r$
The number of degrees of arc in a circle is 360.
The measure in degrees of a straight angle is 180.

Definitions of symbols:

$=$	is equal to	$\leq$	is less than or equal to
$\neq$	is unequal to	$\geq$	is greater than or equal to
$<$	is less than	$\parallel$	is parallel to
$>$	is greater than	$\perp$	is perpendicular to



Triangle: The sum of the measures in degrees of the angles of a triangle is 180.

If $\angle CDA$ is a right angle, then

(1) area of $\Delta ABC = \frac{AB \times CD}{2}$

(2) $AC^2 = AD^2 + DC^2$

Note: Figures which accompany problems in this test are intended to provide information useful in solving the problems. They are drawn as accurately as possible EXCEPT when it is stated in a specific problem that its figure is not drawn to scale. All figures lie in a plane unless otherwise indicated. All numbers used are real numbers.

1. If $\frac{9}{3} + \frac{x}{5} = 2$, then $x =$

   (A) 0 (B) 1 (C) 2 (D) 3 (E) 4

2. A triangle with sides of lengths 4, 8, and 9 has the same perimeter as an equilateral triangle with side of length

   (A) $5\frac{1}{2}$ (B) 6 (C) $6\frac{1}{2}$ (D) 7 (E) $7\frac{1}{2}$

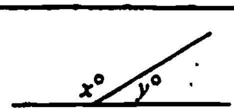
Quantitative Comparison

Questions 8-27 each consist of two quantities, one in Column A and one in Column B. You are to compare the two quantities and on the answer sheet blacken space

- A if the quantity in Column A is greater;
- B if the quantity in Column B is greater;
- C if the two quantities are equal;
- D if the relationship cannot be determined from the information given.

Notes:
1. In certain questions, information concerning one or both of the quantities to be compared is centered above the two columns.
2. In a given question, a symbol that appears in both columns represents the same thing in Column A as it does in Column B.
3. Letters such as x, n, and k stand for real numbers.

EXAMPLES

	Column A	Column B	Answers
E1.	2×6	$2 + 6$	● Ⓑ Ⓒ Ⓓ
			
E2.	$180 - x$	y	Ⓐ Ⓑ ● Ⓓ
E3.	$p - q$	$q - p$	Ⓐ Ⓑ Ⓒ ●

	Column A	Column B
8.	0	0×2
9.	$a + 25$	$a - 5$

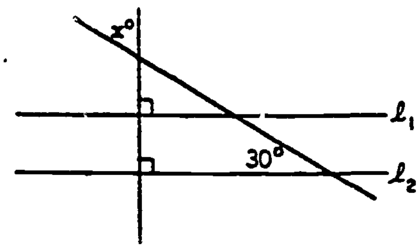
Source: College Board (1983). Taking the SAT. New York: Author

Figure 3

Items Showing Main Effects Exceeding

the .1 Criterion

(a) WSA3



II 6. In the figure above, where $\ell_1 \parallel \ell_2$, x =

(A) 20 (B) 30 (C) 45 (D) 50 (E) 60

(b) WSA3

I 6. In a certain tally system, 12 is represented by HHT HHT II and 15 is represented by HHT HHT HHT. How many uncrossed tallies would be used in the representation of 29 in this system?

(A) None (B) One (C) Two

(D) Three (E) Four

(c) WSA5

N—travel 1 mile north
E—travel 2 miles east
S—travel 3 miles south
W—travel 4 miles west

I 10. If N, E, S, and W are defined as shown above and if a combination of the letters means to perform the instructions in the order given, which of the following yields the same result as NWS ?

(A) W (B) E (C) SEN (D) EWN (E) WSN