DOCUMENT RESUME

ED 268 154 TM 860 175

AUTHOR Rock, Donald A.; And Others
TITLE The Internal Construct Validity of the SAT across Handicapped and Nonhandicapped Populations.
INSTITUTION Educational Testing Service, Princeton, NJ. Graduate Record Examination Board Program.
SPONS AGENCY College Entrance Examination Board, Princeton, N.J.; Tennessee Univ., Knoxville. Dept. of Distributive Education.
REPORT NO ETS-RR-85-50
PUB DATE Nov 85
NOTE 43p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Disabilities; *Factor Structure; Goodness of Fit; Hearing Impairments; Higher Education; High Schools; High School Students; Learning Disabilities; *Mathematics Tests; Models; Physical Disabilities; Scores; Test Interpretation; Test Use; *Test Validity; *Verbal Tests; Visual Impairments
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

The comparability of Scholastic Aptitude Test (SAT) Verbal and Mathematical scores was investigated for one nonhandicapped and nine handicapped populations. The handicapped populations included hearing impaired, visually impaired, learning disabled, and physically disabled students. Special methods of test administration included Braille and large type tests, and administration by a cassette tape. A simple two-factor model based on Verbal and Mathematical item parcels was tested with respect to: the number and intercorrelation of factors; the pattern of factor loadings; and the equality of scale units. The model provided a reasonable fit in all populations, with the mathematical reasoning factor generally showing a better fit. Compared with the nonhandicapped population, these factors were less correlated in most of the handicapped groups. This somewhat greater specificity implied the increased likelihood of achievement growth in one area independent of the other, suggesting that the two scores be interpreted separately rather than as an SAT composite. The presence of two verbal factors was suggested, particularly for the learning disabled. There was evidence that multiple-choice mathematical items led to different observed score scale units for the learning disabled students taking a cassette administration, suggesting that mathematical scores underestimate the reasoning ability of these students. (GDC)

RR-85-50

THE INTERNAL CONSTRUCT VALIDITY OF THE SAT ACROSS HANDICAPPED AND NONHANDICAPPED POPULATIONS

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

November 1985

Report No. 4
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-85-50

The Internal Construct Validity of the SAT Across Handicapped and Nonhandicapped Populations

Donald A. Rock

Randy Elliot Bennett

and

Bruce A. Kaplan

November 1985

Abstract

This study investigated the comparability of SAT Verbal and Mathematical scores for one nonhandicapped and nine handicapped populations. A simple two-factor model based on Verbal and Mathematical item parcels was posed and tested for invariance across populations with respect to: (1) the number and intercorrelation of factors, (2) the pattern of factor loadings, and (3) the equality of scale units.

The two common-factor model provided a reasonable fit in all populations, with the mathematical reasoning factor generally showing a better fit to the population data than the verbal reasoning factor. Compared with the nonhandicapped population, these factors tended to be less correlated in most of the handicapped groups. This somewhat greater specificity implies the increased likelihood of achievement growth in one area independent of the other, suggesting that the two scores be interpreted separately rather than as an SAT composite.

With respect to the pattern of factor loadings, some indication of the presence of two verbal method factors wa discovered. Antonym items and, to a lesser extent, reading comprehension items formed such factors. While present to a certain extent for all groups, the method variance seemed to be somewhat greater for the learning disabled populations.

Finally, there was evidence that multiple-choice Mathematical items led to different observed score scale units for the learning disabled students taking a cassette

administration. Because of the small sample sizes on which this finding is based, it should be treated with caution. However, if dependable, the finding suggests that Mathematical scores may underestimate the reasoning ability of these students.

Acknowledgements

Appreciation is expreesed to Neil Dorans, Gordon Hale, Kim Reid, Joseph Torgesen, and Warren Willingham for their helpful comments on an earlier draft of this report. Gratitude is also expressed to the members of the College Board Joint Staff Research and Development Committee.

This study is primarily concerned with the comparability of Scholastic Aptitude Test (SAT) scores obtained under differing administration procedures. Of particular concern is whether the Verbal and Mathematical scores are comparable across individuals tested under standard administration procedures versus those requesting special administrations. If it can be shown empirically that the interrelationships among homogeneous subsets of items remain relatively constant across groups regardless of whether they received a nonstandard administration, then one can be reasonably confident that the total scores have the same meaning and can be interpreted in the same manner across groups.

In addition to its implications for score interpretation, the results of this study provide evidence supporting the appropriateness of the assumptions used in a companion item-bias study (Bennett, Rock, & Kaplan, in press). This latter study contrasted the differential item performance of each of the handicapped groups with that of the nonhandicapped group after matching on total Verbal or total Mathematical scores. This comparison of groups matched on total test score assumes that the total test score has the same meaning across groups. The expectation of similar performance on a given item for handicapped and nonhandicapped individuals would only be reasonable if groups were being matched on test scores that indeed represented the same ability. Otherwise, the finding of

differential item performance may simply reflect matching on different abilities.

The investigation of what a test measures and whether or not what it measures is invariant across different populations is a problem in construct validity. Empirical verifications of the construct validity of a given test usually rely on the application of factor analytic methodology to individual test items. Rather than factor individual items this study scored "parcels," or homogeneous groups of items, and then used maximum likelihood confirmatory factor analysis to verify the stability of the correlational structure of these parcels across populations. Scores based on item parcels rather than individual items were factor analyzed because:

- Relationships between dichotomously-scored items tend to be non-linear and as a result yield more factors than are present in the data. These artifactual factors are sometimes referred to as difficulty factors since items with similar difficulty indices tend to form factors regardless of whether they are measuring the same constructs. Item parcel scores, however, provide continuous scores and tend to exhibit linear relationships among themselves.

- The computation of multiple factor solutions based on numerous items across as many as nine populations can lead to complex interpretational problems.

- The use of scores on parcels which in turn are more likely to yield multivariate normal distributions allows the use of the more powerful hypothesis testing factor analytic techniques.

- Single items tend to be too unreliable to provide stable markers of factors across populations.

Of particular interest in this study, is whether the two major factors, verbal and mathematical, explain similar

proportions of the reliable test score variance across all populations. Given that it can be demonstrated that the proportion of reliable variance explained by the two major factors is relatively invariant across populations, then questions concerning the invariance of the factor intercorrelations, the pattern of loadings on these factors, and the equality of scale units can be investigated.

It should be noted that a relatively simple two-factor model is being posed and fitted in this study. More complicated models could be, and have been, fitted using the SAT. Rock and Werts (1979) and more recently, Cook, Dorans, Eignor, and Petersen (1985) have fitted more complicated models involving item-type or method factors on quite large samples. Unlike the current study, the Cook et al. study was not attempting to examine the invariance of a factor structure across many potentially disparate populations, but was looking for evidence of violations of the unidimensionality assumption used in IRT (item response theory) equating.

Given the many populations and the relatively small sample sizes involved in the present investigation, there could easily be a tendency to over-fit (i.e., identify spurious factors). We have taken the approach of first defining a well-overdetermined model based on the major factors and then of inspecting the residual variances and covariances in order to identify any remaining method factors. The factor model will be well-overdetermined since

both the verbal and mathematical factors will be identified or "marked" by a number of item parcels. Since we are using a statistically-based factor model, we can examine the residuals and determine if they are significantly different from zero. If the residuals are significantly different from zero and these differences can be considered stable, then we can make a case for the presence of additional method factors without fear of over-fitting.

Comparability as Convergent and Divergent Construct Validity

Baltes and Nesselroade (1973), in a discussion of human development, refer to two specific types of change: structural and quantitative. They define structural change as differences in the numbers of factors, the pattern of factor loadings, and/or the factor intercorrelations. They refer to quantitative change as being related to differences in magnitude, such as the finding of differences in the size of the raw-score factor loadings (i.e., differences in scale units). Baltes and Nesselroade's concept of structural change will be defined here in the context of internal convergent and divergent construct validity.

If it can be shown that the data are reasonably supportive of the invariance of the pattern of factor loadings, then additional questions can be asked about the internal convergent and divergent validity of the test across populations. For example, assume that our factor model is based on item parcels constructed from within the

Verbal and Mathematical sections of the SAT to be "marker" variables for verbal and mathematical reasoning factors. If we find that the resulting verbal and mathematical factors are more highly correlated in certain populations than in others, we then would have to conclude hat the divergent validity of the Verbal and Mathematical scales varies across handicapped populations.

It is worth mentioning here that such a finding does not mean that the test is measuring different things for different populations. It simply means that verbal and mathematical reasoning abilities are less differentiated in some populations than in others. If, however, the relationship in certain groups is sufficiently high as to question the presence of separate verbal and mathematical constructs, then there is a threat to the assumption of the invariance of the test's divergent validity.

Questions concerning the invariance of the internal convergent validity of the test are related to notions of factor reliability, equivalence of scale units, and the broader notion of construct validity. Internal convergent validity refers to how internally consistent, for example, the item parcels that mark the verbal factor are in one handicapped group versus another. If we express factor loadings in standardized rather than raw-score units, we are dealing with the question of whether the factor reliabilities are invariant across populations. Once again this is both a question of reliability and construct

validity. In this case, we are referring to the possibility of both group differences in the overall factor reliability as well as possible differences in the internal consistency reliability of the individual item parcels given a single underlying common factor. If, for example, the reading comprehension item parcel loadings on the verbal factor are smaller for one handicapped group than for all others, the convergent validity of the reading comprehension items has to be questioned with respect to their being equally good measures of verbal ability for that group.

When we express factor loadings in raw-score terms, we are referring to possible group differences in the observed score scale units across populations. This concept of quantitative differences as reflected in differing scale units across groups suggests that the observed score scale units may not have the same meaning. The raw-score factor loadings can be interpreted as the regression of the observed scores on the "true" or factor scores. For example, if the raw-score factor loading of a sentence completion parcel on the verbal factor is 1.0 in one group and 2.0 in a second group, this suggests that a unit increase in "true" verbal ability is reflected as a two point increase in the parcel score in the second group but only a one point increase in the first group. In such a case, one might conclude that the observed score units of this parcel may not be comparable across these groups. In fact, the observed Verbal scores tend to underestimate the

"true" verbal reasoning ability of members of the first group compared to those in the second.

One, however, should be cautious about rejecting the hypothesis of equivalent scale units if: (1) the groups being compared are centered in different parts of the score distribution, and (2) the groups being compared are quite small in terms of sample size. The first condition refers to the possibility that the groups may be at quite different ability levels. If that is the case, the finding of equivalency of scale units across groups quite disparate in ability suggests that one raw-score point at the low end of the score distribution is equivalent to one raw-score point at the middle or upper end of the distribution. One would not expect many test score metrics to meet this assumption. (It is of interest to note, however, that computerized adaptive testing likely would lead to equal scale units throughout the score distribution.) The second condition refers to the possibility of obtaining unstable estimates in small samples. Classical statistical tests do not always protect from over-interpretation of unstable differences if one of the groups is quite large, thereby increasing the power to reject the null hypothesis of no differences.

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 throughout the United States. 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 each of the remaining students is unknown. During this period, other handicapped students undoubtedly took standard administrations of the SAT. Because it is not necessary to reveal the presence of a disability unless a special administration is requested, the number of handicapped students taking a standard administration is unknown.

Data from both WSA3 and 5 were used in this study. By using both data sets attention can be focused on those findings that are replicated across forms. Such repeated occurrences are less likely to be artifacts associated with a particular test form or sample of subjects.

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, and cassette and regular type. All special

administrations included extended time. Tables 1 and 2 show the number of students with each disability taking each type of special administration of WSA3 and 5.

Insert Tables 1 and 2 about here

The tables show that the largest number of special administrations was taken by learning disabled students and the most frequently used format was regular type. Visually impaired students represented the second largest disability group and large type the second most frequently used format. Of the 35 possible test-by-format-by-disability group combinations, the two largest were LD students taking regular type and visually impaired students taking large type administrations.

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 justify further study. These groups are, for regular type, visually impaired, hearing impaired, and physically handicapped students; for large type, learning disabled students; for braille, visually impaired students; and for cassette and cassette and regular type, learning disabled students.

The reference group used in this study consists of high school seniors taking WSA3 and WSA5. The sample taking WSA3 contains 35,424 students taking the test in Texas and California during October 1974. The sample taking WSA5

includes 33,161 examinees taking the test in national administrations in December of that same year.

Table 3 shows the sample sizes and acronyms used to denote each of the study groups.

Insert Table 3 about here

Factor Model and Method

Figure 1 shows the factor model assumed to underlie the SAT Verbal and Mathematical sections. Each of the factors composing the model is marked by the item types that comprise the two scales of the SAT. For the Verbal scale, these types are antonyms, sentence completion, analogies, and reading comprehension. The Mathematical scale is made up of quantitative comparisons and regular multiple-choice items. Each of these item types is further subdivided into three parcels, with parcels within an item type balanced on difficulty.

Insert Figure 1 about here

The asterisks in Figure 1 indicate that a factor loading is to be estimated. Conversely a "0" denotes that the indicator variable will have a zero loading on that particular factor. For example, the Verbal antonyms-A parcel is expected to have a non-zero loading on factor one (the verbal factor) and a zero loading on factor two (the

mathematical factor). The maximum likelihood factor estimation procedure (Joreskog & Sonbom, 1984) will be used to estimate the unknown factor loadings (i.e., the asterisks) subject to the patterns of "zero" constraints and assuming that the factors are allowed to be intercorrelated.

The question posed by the study is how well the above two-factor "simple structure" model fits the data within each handicapped population. By simple structure it is meant that a parcel is constrained to load only on its assumed underlying factor. Various measures of overall goodness of fit will be computed within each population which, in turn, will provide a measure of how well the model fits in each respective group. A reliability discrepancy procedure along with the analysis of residual covariances will be used to identify what parts of the factor model do not fit in any one population.

There are three primary ways in which the above model may not fit the data in one or more of the populations. First, there is always the possibility that a single factor model will fit cognitive data as well as a multi-factor model. This possibility is generally reflected in a confirmatory factor solution by excessively high intercorrelations among factors.

Given that the two-factor model can be fitted to the data, a second possibility for poor fit would be a finding that the pattern of constrained and unconstrained loadings differs by population. Such a situation would be indicated

by relatively low overall goodness of fit indices in those populations where the constrained factor pattern is inappropriate. In addition, the factor model estimates of item parcel reliabilities would be discrepant from coefficient Alpha estimates of reliability.

The third reason for lack of fit is that in one or more populations there may be more than two common factors. This type of lack of fit would lead to poor overall goodness of fit indices including factor and Alpha reliability discrepancies, as well as relatively large residual covariances. Significant covariance among the residuals indicates that part of the original covariance cannot be explained by the two-factor constrained solution. If there are more than two common factors in a particular population, then one should be able to point to systematic non-zero patterns of covariance among the residuals. For example, if the verbal factor "broke down" into two factors--one consisting of sentence completion and reading comprehension parcels and the other of antonyms and analogies--one would expect to observe non-zero covariances between corresponding parcels within these subgroupings.

Results

Results are discussed in terms of the number and intercorrelation of factors, the pattern of factor loadings, and the equality of scale units.

Number and Intercorrelation of Factors

Tables 4 and 5 present goodness of fit indices and factor intercorrelations by test form across population for the two-factor (verbal and mathematical) model. As indicated, this analysis allows the loadings to be freely estimated within each population subject to the constraints defined by the hypothesized verbal and mathematical factor pattern. This analysis helps to determine if the hypothesized number of factors and their associated pattern of loadings is a reasonable underlying structure for the test within each population. Inspection of Tables 4 and 5 suggests that the factor intercorrelations are sufficiently low to infer that a two-factor verbal and mathematical model demonstrates sufficient divergent validity within all populations. However, with certain minor exceptions, there appears to be a somewhat higher relationship between verbal and mathematical performance for the nonhandicapped population than for the handicapped groups. This somewhat greater factor divergence within the nonhandicapped population is consistent across both forms.

Insert Tables 4 and 5 about here

The relatively greater independence of factors in the handicapped samples may be the result of several influences. First, some groups of handicapped students, in particular those with learning disabilities, are identified as

handicapped because they exhibit an uneven pattern of abilities. This selection factor would understandably result in the greater independence of abilities noted in these handicapped groups. Second, in some populations the presence of more ...ct abilities may be secondary to the handicapping condition. So, for instance, because of their inability to hear, prelingually deaf students may never develop advanced verbal skills, though their development of mathematical reasoning abilities may progress more normally. Finally handicapped students and their teachers may tend in their educational programs to emphasize pupil strengths. In the secondary school years, students with math-based learning disabilities may opt to take fewer math courses than their nonhandicapped peers while at the same time pursuing advanced English, history, and other courses likely to strengthen their verbal abilities.

The two goodness of fit indices, goodness of fit ratio (GFR) and root mean square residual (RMSR), are measures of the overall fit of the hypothesized two-factor model withi each population. Both these measures are estimated independently of sample size and thus can be compared across groups of different sizes. Generally GFRs in the middle to the high nineties are considered very good fits while GFRs in the middle eighties to the low nineties are considered reasonably good fits. Indices in the low eighties or lower may be considered somewhat questionable evidence with respect to supporting the hypothesized factor model. The

root mean square residual can be considered as the average covariance among the parcels that is left over after the hypothesized two-factor model has been fitted.

Because of the relatively small sample sizes in many of the handicapped populations, it will be our practice to identify "poor fits" only if: (1) both the GFR and the RMSR are in agreement, (2) this agreement extends across both test forms, and, more importantly, (3) the index of reliability discrepancy is large. (This latter index will be defined in detail further on.) We take this approach because the hypothesized factor model is fitted in each population separately and the stability of the estimates are a function of sample size. Unless replicated, a finding of a "poor fit" in a population based on 100 or so cases may have little meaning.

Inspection of the fit indices presented in Tables 4 and 5 suggests that on the whole the hypothesized two-factor model fits the handicapped groups as well as the nonhandicapped sample. Two possible exceptions are the fit in the learning disabled-cassette group and, to a lesser extent, the fit in the visual-braille sample. It should be noted, however, that while the GFRs are in the low eighties for these groups, their RMSR index is only slightly higher compared to the other groups.

With the possible exception of the learning disabled-cassette group, inspection of the residuals indicates very little in the way of systematic patterns that might lead one

to hypothesize additional common factors. The residual matrix for this group (not shown) is characterized by a cluster of positive residuals among the antonym parcels. Since these residuals appeared only among the antonym parcels, it would seem that this finding reflects the presence of a method factor rather than an additional common factor. This phenomenon will be discussed further in the section below which deals with the use of the reliability discrepancy procedure to identify specific item parcels contributing to the overall lack of fit as measured by the GFR and RMSR.

Pattern of Factor Loadings

Reliabilities of the item types can be estimated from the factor model which, in turn, assumes (depending on the item type) that the item types are all indicators of either the verbal or mathematical factors. Furthermore, the factor model estimates of the item-type reliabilities will be underestimates of coefficient Alpha reliabilities to the extent that the item type does not share the same single underlying factor as the other item types. The factor model estimates of the reliabilities of the Verbal item parcels assume single factoredness holds across all Verbal item types. If, for example, the factor estimates of the reliability of the reading parcels are found to be smaller than their coefficient Alpha estimates (which only assume single factoredness within item type), then one can assume that they are a less reliable indicator of the underlying

common factor in that population. Such parcels could be expected to contribute significantly to the overall lack of fit as measured by the GFR and/or RMSR.

Tables 6 and 7 present the factor reliabilities and the coefficient Alpha reliabilities by item type within form across the various populations. Inspection of the reliabilities of the Verbal item types in Table 6 shows that the largest discrepancy between the factor reliabilities and the Alpha estimates is for the antonym item type within the learning disabled-cassette group. This discrepancy occurs across both forms for this group. This is the group which also shows the worst GFR with respect to the hypothesized two-factor model. Apparently part of the lack of fit of the factor model in the learning disabled-cassette population is due to the fact that the antonym item type has somewhat less internal convergent validity with respect to the verbal factor for this handicapped group. Further inspection of Table 6 indicates that, compared to the other Verbal item types, there is a tendency for antonyms to show greater factor versus Alpha reliability discrepancies and, thus, less convergent validity for almost all groups. These discrepancies are typically largest, however, for the various learning disabled groups.

Insert Tables 6 and 7 about here

The learning disabled groups also tend to show larger reliability discrepancies for the reading comprehension items, suggesting that the reading parcels have comparatively less convergent validity with respect to the verbal factor for these groups. Inspection of the residuals for these groups indicates that both forms tend to show positive clusters of residuals among the reading comprehension parcels. This phenomenon is particularly present for the learning disabled large type group.

For the visual-braille group--the one other handicapped population that showed some tendency for a poorer factor model fit--there is no indication that the convergent validity (as measured by reliability discrepancies) for any of the Verbal item types is abnormally low. That is, according to the reliability discrepancy procedure, the two-factor model seems to be explaining virtually all the reliable item-type variance within this population. In addition, an examination of the residual matrix shows no systematic patterns among the residual covariances. It may be that the GFR and the RMSR indices of fit are overly sensitive to slight variations in fit compared to the reliability discrepancy procedure.

An overall index of the goodness of fit of the single verbal common factor is the average proportion of reliable variance in the item-type parcels explained by the factor model. On average, approximately 97-98% of the reliable variance in the sentence completion and analogies parcels is

explained by the single verbal factor. Similarly, 90 and 95% of the reliable variance in the antonym and reading comprehension parcels, respectively, is explained by the single verbal factor.

Inspection of the reliability discrepancies for the item types used to "mark" the mathematical factor suggests that a single mathematical common factor fits comparatively well in all handicapped populations. When averaged across both groups and forms, approximately 98% of the reliable Mathematical parcel variance is explained by the single mathematical common factor.

While the single mathematical factor seems to fit in all populations, there are some differences in the Mathematical item-type reliabilities across the various populations. The reliabilities of both the quantitative comparison item parcels as well as the multiple-choice math parcels are proportionately lower for the learning disabled-cassette group. Whether this difference is "real" or simply reflects differences in population homogeneity, is addressed in the following analysis which tests the assumption of identical raw-score factor loadings (scale units) across populations.

Equality of Scale Units

The question arises, "What would happen if we should 'force' the factor structure and scale units of the nonhandicapped population on each of the handicapped groups in turn?" That is, what happens when one takes the best

fitting two-factor solution in the nonhandicapped population and applies it within each handicapped group to identify those groups where the item-parcel scale units seem to be most different from those of the nonhandicapped group? By equality of scale units, we mean that if we fix one of the raw-score factor loadings at unity, we would expect the remaining loadings to have the same proportionality ratios across groups. That is, if we set the loading associated with the first indicator of the verbal factor to 1.0, all other verbal factor loadings should maintain the same proportionality ratios across populations, where the proportionality ratios are based on the nonhandicapped population. For those populations where this nonhandicapped factor-scaling model fits, it can be said that the factor pattern is not only the same but that the raw-score scale units also appear to be the same. That is, the strength of the relationship between the item parcel raw scores and the factor is invariant across populations. Since we are interested in raw-score scale units, this analysis was done (as was the previous analysis) on the variance-covariance matrices.

This approach reduces the possible group comparisons to nine. It is also consistent with the item-bias study mentioned earlier (Bennett et al., in press), which used the nonhandicapped group as the standard for comparison. This approach would have been less feasible if there was clear

evidence from the above analysis of differing numbers of common factors in the various populations.

Table 8 presents the goodness of fit indices by population within test form when the nonhandicapped two-factor pattern and raw-score loadings are fitted to the nonhandicapped data in each handicapped group. Inspection of the GFRs in Table 8 suggests that the scale units appear to be most different from the nonhandicapped population in the visual-braille and learning disabled-cassette groups. This difference is consistent across both forms. We prefer to lean more heavily on the GFR indices than the RMSR since the GFR takes into consideration the additional constraints imposed in this more restrictive model. In addition, the RMSR can be misleading if the populations differ considerably in their heterogeneity.

Insert Table 8 about here

Table 9 summarizes differences in the average raw-score loadings by item parcel for the nonhandicapped, the visual-braille, and the learning disabled-cassette groups. A comparison of the average raw-score factor loadings between the nonhandicapped group and the visually impaired group shows little in the way of systematic differences that are stable across both forms. However, a similar comparison between the nonhandicapped and learning disabled-cassette groups suggests that the multiple-choice Mathematical

have consistently smaller raw-score factor loadings on the mathematical reasoning factor in the latter group.

Insert Table 9 about here

It would appear that, when administered with a cassette, some Mathematical items may yield differences in scale units for the learning disabled group compared to the nonhandicapped group. The fact that the raw-score parcel loadings were smaller in the learning disabled group compared to the nonhandicapped sample suggests the observed Mathematical scores may underestimate (compared to the nonhandicapped) the "true" mathematical reasoning ability of members of the learning disabled group. This conclusion is further supported by the finding in the item-bias companion study (Bennett et al., in press), that when controlling on observed Mathematical score, members of the learning disabled-cassette group performed better than expected on certain Mathematical items.

Summary and Conclusions

A two-factor verbal and mathematical model was fit to the variance-covariance matrix of item parcels in each of nine handicapped groups and one nonhandicapped population. As a result of the analysis, several conclusions were reached. First, the two-factor model was found to fit the data reasonably well in each population. In general, the mathematical factor tended to show a better fit to the

population data than the verbal factor. When averaged across both groups and forms, approximately 98% of the reliable Mathematical parcel variance was explained by the single mathematical common factor. For verbal, approximately 97-98% of the reliable variance in sentence completion and analogies parcels, and 90% and 95% of the reliable variance in reading comprehension and antonym parcels was explained, respectively, by a single verbal factor.

Second, the verbal and mathematical factors tended to be less correlated in most of the handicapped populations compared to the nonhandicapped population. This somewhat greater specificity in the handicapped populations suggests the increased likelihood of achievement growth in one area independent of the other. This phenomena may be due to a variety of factors including selection bias, conditions secondary to the handicap itself, and the focus of special education programs.

Third, there was some indication that the antonym item type was measuring something in addition to general verbal reasoning ability. This finding tended to apply somewhat to all groups, but was particularly true for the learning disabled populations, especially the learning disabled-cassette group. A similar situation, though less pronounced, was found to exist for the reading comprehension items within the learning disabled groups.

Finally, there was evidence that multiple-choice Mathematical items led to different observed score units for the learning disabled-cassette group compared to those of the nonhandicapped group. This finding, however, should be treated with caution due the small sample sizes of this group.

In summary, the factor structure of the SAT appears, for the most part, the same for handicapped and nonhandicapped groups. It would appear that, from an internal construct validity perspective, Verbal and Mathematical scores can be interpreted in much the same fashion across groups with two possible exceptions. First, for the learning disabled-cassette population, either the uniqueness of the population and/or the cassette administration appears to yield Mathematical scores that may be on a different observed score scale than those of the nonhandicapped population. This implies that observed scores may underestimate "true" mathematical ability for learning disabled-cassette group members. Second, because there appears to be more specificity between the measured verbal and mathematical abilities in the handicapped populations, the two SAT total scores might be best interpreted separately rather than as a single composite.

References

Baltes, P. B., & Nesselroade, J. R. (1973). The developmental analysis of individual differences on multiple measures. In J. R. Nesselroade and H. W. Reese (Eds), Life-span developmental psychology: Methodological issues. New York: Academic Press.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (In press). The psychometric characteristics of the SAT for nine handicapped groups. Princeton, NJ: Educational Testing Service.

Cook, L. L., Dorans, N. L., Eignor, D. R., & Petersen, N. S. (1985). An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating (RR-85-30). Princeton, NJ: Educational Testing Service.

Joreskog, K., & Sonbom, D. (1984). LISREL 6: An analysis of linear structural relationships by the method of maximum likelihood. Mooresville, IN: Scientific Software, Inc.

Rock, D. A., & Werts, C. E. (1979). Construct validity of the SAT across populations: An empirical confirmatory study (CEEB RR-79-2). Princeton, NJ: Educational Testing Service.

TABLE 1. NUMBER OF STUDENTS TAKING EACH TYPE OF SAT SPECIAL ADMINISTRATION FOR WSAS

	GROUP[a]						
EXAM TYPE	VI	PH	HI	LD	MULTIPLE	UN-KNOWN	TOTAL
BRAILLE	98	0	1	2	0	1	102
LARGE TYPE	486	30	6	185	18	1	726
CASSETTE	27	2	-	107	3	0	140
REGULAR	223	346	287	2983	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 2. NUMBER OF STUDENTS TAKING EACH TYPE OF SAT SPECIAL ADMINISTRATION FOR WSAS

	GROUP[a]						
EXAM TYPE	VI	PH	HI	LD	MULTIPLE	UN-KNOWN	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 3. SAMPLE SIZES AND ACRONYMS USED TO DENOTE STUDY GROUPS

ACRONYM	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	2,983	2,316
N	High school seniors taking regular-type edition in standard, timed administrations	35,424	33,161
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 4. GOODNESS OF FIT INDICES AND FACTOR INTERCORRELATIONS FOR FORM WSA3 ASSUMING THE SAME PATTERN OF LOADINGS

GROUP	FACTOR INTERCORRELATIONS	GOODNESS-OF-FIT RATIO	ROOT MSR
NONHANDICAPPED	0.755	0.950	.114
VISUAL-BRAILLE	0.674	0.838	.180
VISUAL-LARGE TYPE	0.697	0.947	.110
VISUAL-REGULAR	0.[illegible]3	0.906	.171
LD-REGULAR	0.668	0.906	.084
LD-CASSETTE	0.691	0.830	.170
LD-CASSETTE-REGULAR	0.753	0.909	.115
LD-LARGE TYPE	0.642	0.890	.143
HEARING-REGULAR	0.698	0.918	.156
PHYSICAL-REGULAR	0.673	0.922	.159

TABLE 5. GOODNESS OF FIT INDICES AND FACTOR INTERCORRELATIONS FOR FORM WSA5 ASSUMING THE SAME PATTERN OF LOADINGS

GROUP	FACTOR INTERCORRELATIONS	GOODNESS-OF-FIT RATIO	ROOT MSR
NONHANDICAPPED	0.760	0.892	.155
VISUAL-BRAILLE	0.798	0.826	.160
VISUAL-LARGE TYPE	0.755	0.950	.114
VISUAL-REGULAR	0.760	0.892	.155
LD-REGULAR	0.661	0.961	.091
LD-CASSETTE	0.587	0.835	.177
LD-CASSETTE-REGULAR	0.656	0.906	.136
LD-LARGE TYPE	0.674	0.874	.151
HEARING-REGULAR	0.710	0.880	.151
PHYSICAL-REGULAR	0.697	0.895	.13

TABLE 6. FACTOR VS. ALPHA RELIABILITIES FOR SAT-V ITEM TYPES

	WSA3 Alpha Rel (1)	WSA5 Alpha Rel (2)	WSA3 Factor Rel (3)	WSA5 Factor Rel (4)	WSA3 Diff (1)-(3)	WSA5 Diff (2)-(4)
READING COMPREHENSION						
vir	0.74	0.78	0.70	0.76	0.04	0.02
vil	0.78	0.80	0.75	0.76	0.03	0.04
vib	0.80	0.86	0.78	0.86	0.02	.00
phr	0.78	0.78	0.71	0.74	0.07	0.04
n	0.78	0.78	0.76	0.76	0.02	0.02
ldr	0.76	0.76	0.73	0.71	0.03	0.05
ldl	0.67	0.74	0.59	0.66	0.08	0.08
ldcr	0.71	0.72	0.66	0.66	0.05	0.06
ldc	0.72	0.74	0.67	0.72	0.05	0.02
hir	0.70	0.78	0.69	0.75	0.01	0.03
ANTONYMS						
vir	0.73	0.73	0.66	0.62	0.07	0.11
vil	0.75	0.74	0.71	0.69	0.04	0.05
vib	0.82	0.81	0.81	0.76	0.01	0.05
phr	0.77	0.74	0.73	0.66	0.04	0.08
n	0.74	0.69	0.69	0.62	0.05	0.07
ldr	0.71	0.70	0.63	0.61	0.08	0.09
ldl	0.70	0.72	0.64	0.63	0.06	0.09
ldcr	0.54	0.64	0.45	0.55	0.09	0.09
ldc	0.65	0.59	0.55	0.43	0.10	0.16
hir	0.72	0.69	0.69	0.59	0.03	0.10
SENTENCE COMPLETION						
vir	0.66	0.65	0.65	0.64	0.01	0.01
vil	0.70	0.65	0.70	0.64	.00	0.01
vib	0.74	0.79	0.74	0.79	.00	.00
phr	0.64	0.66	0.65	0.69	-0.01	-0.03
n	0.65	0.67	0.64	0.64	0.02	0.03
ldr	0.64	0.64	0.63	0.64	.01	.00
ldl	0.61	0.59	0.52	0.66	0.09	-0.07
ldcr	0.51	0.53	0.55	0.57	-0.04	-0.04
ldc	0.57	0.51	0.55	0.52	0.03	-0.01
hir	0.61	0.74	0.62	0.73	-0.01	0.01
ANALOGIES						
vir	0.73	0.75	0.70	0.75	0.03	0.00
vil	0.75	0.80	0.71	0.79	0.04	0.01
vib	0.78	0.82	0.76	0.87	0.02	-0.05
phr	0.75	0.75	0.73	0.76	0.02	-0.01
n	0.77	0.78	0.79	0.75	-0.02	0.03
ldr	0.69	0.71	0.64	0.70	0.05	0.01
ldl	0.72	0.73	0.68	0.68	0.04	0.05
ldcr	0.62	0.65	0.61	0.63	0.01	0.02
ldc	0.68	0.56	0.63	0.53	0.05	0.03
hir	0.61	0.73	0.63	0.72	-0.02	0.01

TABLE 7. FACTOR VS. ALPHA RELIABILITIES FOR SAT-M ITEM TYPES

	WSA3 Alpha Rel (1)	WSA5 Alpha Rel (2)	WSA3 Factor Rel (3)	WSA5 Factor Rel (4)	WSA3 Diff (1)-(3)	WSA5 Diff (2)-(4)
			QUANTITATIVE COMPARISONS			
vil	0.83	0.84	0.82	0.84	0.01	.00
vir	0.85	0.82	0.82	0.83	0.03	-0.01
vib	0.81	0.81	0.80	0.81	0.01	.00
phr	0.82	0.80	0.82	0.80	.00	.00
n	0.84	0.82	0.84	0.83	.00	-0.01
ldr	0.83	0.81	0.82	0.79	0.01	0.02
ldl	0.82	0.75	0.81	0.72	0.01	0.03
ldcr	0.80	0.79	0.78	0.75	0.02	0.04
ldc	0.79	0.78	0.79	0.74	.00	0.04
hir	0.80	0.78	0.80	0.75	.00	0.04
			MULTIPLE CHOICE			
vil	0.91	0.88	0.91	0.88	.00	.00
vir	0.92	0.90	0.91	0.90	.00	.00
vib	0.90	0.93	0.91	0.93	.00	.00
phr	0.90	0.87	0.90	0.88	.00	.00
n	0.88	0.90	0.88	0.91	.00	-0.01
ldr	0.90	0.87	0.90	0.87	.00	.00
ldl	0.85	0.82	0.84	0.82	0.01	.00
ldcr	0.88	0.83	0.88	0.82	.00	0.01
ldc	0.81	0.76	0.80	0.71	.00	0.04
hir	0.87	0.88	0.87	0.88	.00	.00

TABLE 8. GOODNESS OF FIT INDICES BY FORM WHEN THE NON-HANDICAPPED FACTOR LOADINGS ARE FITTED TO THE DATA IN EACH HANDICAPPED GROUP

	FORMS			
	WSA3		WSA5	
	GFR	RMSR	GFR	RMSR
NONHANDICAPPED	.950	.114	.892	.155
VISUAL-BRAILLE	.793	.453	.808	.352
VISUAL-LARGE TYPE	.910	.305	.936	.204
VISUAL-REGULAR	.866	.306	.892	.155
LD-REGULAR	.930	.279	.952	.156
LD-CASSETTE	.810	.275	.795	.287
LD-CASSETTE-REGULAR	.887	.235	.892	.204
LD-LARGE TYPE	.856	.290	.857	.246
HEARING-REGULAR	.874	.299	.856	.307
PHYSICAL-REGULAR	.886	.327	.885	.212

TABLE 9. AVERAGE RAW SCORE FACTOR LOADINGS BY ITEM TYPE FOR THOSE GROUPS WHO WERE MOST DISPARATE FROM THE NONHANDICAPPED GROUP

	GROUPS					
ITEM TYPES	NONHANDICAPPED		VIS BRAILLE		L.D. CASSETTE	
	WSA3	WSA5	WSA3	WSA5	WSA3	WSA5
ANTONYMS	1.001	.974	.796	1.000	.965	.966
SENTENCE COMP.	.639	.730	.490	.702	.725	.761
ANALOGY	1.121	1.067	.657	.996	1.010	.934
READ	1.334	1.392	.804	1.324	1.266	1.863
MATH Q-C	1.106	1.206	.93	.97[a]	.824	1.000
MATH MC	1.829	2.217	1.916	1.977	1.304	1.357

FIGURE 1. FACTOR MODEL

MARKER VARIABLES	VERBAL FACTOR	MATHEMATICAL FACTOR
Antonyms-A	*	0
Antonyms-B	*	0
Antonyms-C	*	0
Sentence completion-A	*	0
Sentence completion-B	*	0
Sentence completion-C	*	0
Analogies-A	*	0
Analogies-B	*	0
Analogies-C	*	0
Reading comprehension-A	*	0
Reading comprehension-B	*	0
Reading comprehension-C	*	0
Quantitative comparison-A	0	
Quantitative comparison-B	0	*
Quantitative comparison-C	0	*
Regular math-A	0	*
Regular math-B	0	*
Regular math-C	0	*