DOCUMENT RESUME

ED 219 443 TM 820 467

AUTHOR Forehand, Garlie A.
TITLE Testing the Handicapped: Validation and Test Interpretation.
PUB DATE Mar 82
NOTE 12p.; Paper presented at the Annual Meeting of the National Council of Measurement in Education (New York, NY, March 20-22, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Ability; *Disabilities; *Evaluation Criteria; Measures (Individuals); Methods Research; Models; *Predictive Validity; Test Construction; Testing Problems; *Test Interpretation; *Test Validity

ABSTRACT

Problems in validating ability tests for handicapped students and research approaches to predictive validity are discussed. Validity for handicapped persons tested under regular conditions; for applicants to special programs, and for tests taken under special administrative conditions are considered. Item analysis and the construction of new scales designed to improve validity are discussed. Construct validity studies to reveal the extent to which a test measures the same variables in handicapped and non-handicapped populations are suggested. Limitations in validation as a model for evaluating tests for the handicapped are discussed with various decision-making mechanisms and alternative perspectives of validation. (Author/CM)

Testing the Handicapped: Validation and Test Interpretation*

Garlie A. Forehand
Educational Testing Service

The Panel on Testing Handicapped People has recently released the executive summary of its report: Ability Testing of Handicapped People: Dilemma for Government, Science and the Public. The panel's key recommendations are that test users, test publishers, and researchers

- Develop modified tests to meet the needs of individuals with sensory and motor handicaps.
- Perform predictive validation studies on these tests, to be reported within four years.
- Undertake research to contribute to greater understanding of tests for handicapped persons, their validity, their modification, their supplantability by other measures, and their role in decision making.

How should the educational research community respond to the panel's recommendations? What will it take to accomplish their objectives? Where are we now and in what directions do we need to move? These are questions to be considered in this presentation.

It is appropriate that the Panel subtitled its report a dilemma. Dilemmas are unavoidable when one attempts to use and interpret the test scores of handicapped people. It is obvious that handicaps often produce conditions that give rise to both issues of interpretation and issues of use. Simple but realistic examples are sufficient to demonstrate the existence of dilemmas. A blind student solving a problem with a diagram must rely on a tactile representation and verbal description of the diagram. Is it not likely that this mathematics performance

---

*Presented as part of a symposium on Admissions Testing for the Handicapped at the 1982 Annual Meeting of the National Council of Measurement in Education, New York, March 21, 1982.

is more saturated with verbal ability than is that of a sighted student using a visual aid? A deaf candidate for occupational certification is given a verbal test. Might not the necessary occupational skills be assessed more fairly with a non-verbal test?

It is reasonable to look to research for clarification of problems of test use and test interpretation. For the researcher, the problem is not only that there are few answers in the literature, but also that the proper research questions are elusive and innumerable. The closer one comes to specifying the problems of use and interpretation, the more the research problems proliferate. The purpose of this presentation is to begin to identify and classify the research problems posed by problems of test use and interpretation, and thus to generate some guides for the design and conduct of research. Two conclusions may be stated at the outset. First, a productive approach to research on handicapped students is going to require a new look at handicapped students. Virtually all approaches to testing the handicapped are based on procedures developed for the non-handicapped. Thus, we talk about adapted testing procedures and exceptional admission rules. Research based on this assumption is necessary but it will not answer questions about potential approaches designed for the handicapped. The second conclusion is that the problems of test use and validity raised in connection with handicapped students are in fact basic problems of psychometrics, problems that we have often been able to finesse by assuming populations that are homogeneous with regard to sensory, motor, and expressive abilities. Thus, the research questions posed are pertinent to the evaluation of test interpretations and test uses beyond the specific application to handicapped students.

Predictive Validity for Selection Applications

In selection there is always some probability of being rejected and therefore some probability of unsatisfactory consequences: rejection of a qualified applicant or acceptance of an unqualified one. When the number of spaces is large relative to the applicants, negative consequences can be minimized by a policy of leniency. In

effect, handicapped students may be given a chance to try a program and their initial success becomes a part of the selection process. When applicants compete for a few positions, so that selection of one applicant implies rejection of another, a lenient policy produces its own inequities. There are a few important cases in which handicapped applicants compete with other handicapped applicants: special college programs for the deaf or for the learning disabled for example. In these instances evaluation of selection tests becomes a critical issue.

Predictive validity of such measures falls into several cases:

1. Validity for handicapped students tested under regular conditions. ACT has reported a number of studies of this sort and generally finds the test scores to be as valid for handicapped as for non-handicapped students. However, these were by definition handicapped students who were able to cope with regular testing conditions. It seems likely either that their handicaps influenced their performance relatively little or that they had learned skills to overcome the difficulties imposed by their handicaps. These factors may equally influence criterion performance.

2. Validity for applicants to special programs. Douglas Jones and Margery Ragosta of ETS have conducted validity studies for two groups of handicapped students enrolled in colleges with special programs: deaf students at California State University, Northridge, and students with learning disabilities at Curry College. These institutions, because they have special programs, have substantial numbers of students with comparable criterion data. Because they offer many special services, their results cannot be generalized to institutions without special programs.

In both studies, valid regression estimates of college grades for handicapped students were found. The validity coefficients were not significantly different from those for nonhandicapped students in the same institutions. The equations that yielded the significant validity coefficients, however, were markedly different for the two groups. For the deaf sample, high school grade-point average overpredicts freshman college grades, while SAT scores, especially SAT-V underpredicts college performance. For the learning disabled group, applying

a regression estimate developed on nonhandicapped students to the LD group would consistently overestimate college performance. The differences in prediction, though statistically significant, were relatively small in relation to the standard error of estimate. In both cases, intercorrelations of the measures suggest that test scores are measuring different variables in the handicapped and nonhandicapped samples.

3. Validity for tests taken under special administrations. These validation studies are more difficult to do since relatively few people with a given handicap enroll in a given college, thus providing comparable grade data. Yet from the perspective of maximizing fairness and usefulness of tests, these studies are the most important; no other design assesses the validity of the specific measure represented by a modification. ETS has taken some important steps toward making studies of this sort practical. First, Braun and Jones have experimented with empirical Bayes procedures for aggregating data across institutions and across cohorts in a given institution. Second, under Ragosta's direction, we are in the process of accumulating a data base across years that would make it possible to combine samples across years in a given college.

There are some limitations to all of these studies. First, there is a kind of circularity built into predictive validity studies of handicapped students. The disability that has interfered with past learning--and thus test scores--may also interfere with future learning--and criterion performance. If educational conditions were modified to overcome the disabilities, as by providing interpreters for the deaf and readers for the blind, this "created validity" might disappear. Second, the scales and scores used for handicapped applicants are based on data for the non-handicapped. The fact that mathematics items with diagrams have known correlational and factorial structures for sighted students does not necessarily imply that the same relations would hold for blind students.

To overcome these limitations will require new efforts to build the most effective tests possible for specific groups of the handicapped. This means beginning at the item level and constructing new scales and

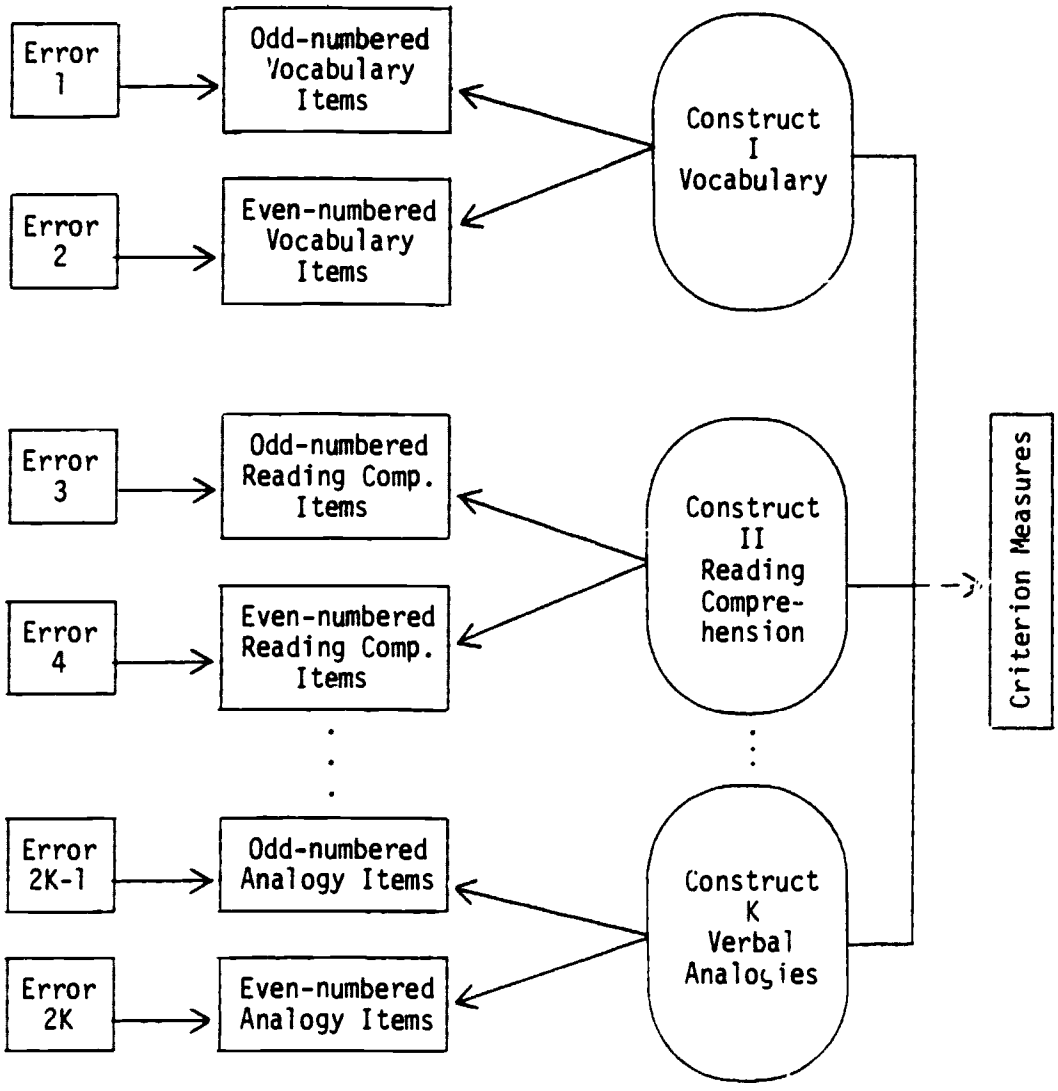
scores designed to improve validity. For example, certain items might be omitted and the remainder rescaled; items may be substituted for other items; and weightings may be changed. To improve predictive validity, items would be analyzed against total scores for specific handicapped groups and against criterion performance. To improve construct validity, items would be analyzed in relation to alternative measures of the same and different constructs and items selected and combined to provide the intended factor structure. This kind of work is needed to approach the goal of measuring what handicapped students can do rather than what they cannot. In the long range, one may envision collections of items that vary in format but are calibrated in difficulty and categorized by domain. An individual's test might be tailored according to individual sensory and motor abilities to produce a result comparable in meaning and predictive value to those of other test takers.

Mary Anne Nester has done pioneering work in this area for the Office of Personnel Management. ETS is now assembling a data base for such work at the item level, under the leadership of Marjorie Ragosta.

Construct Validity

To reach the goals just outlined, predictive validation studies will not be sufficient. A comprehensive program of construct validation research will be required. The handout illustrates some of the questions of construct validation research. The diagram represents a hypothetical structure of a verbal ability test battery. The battery is divided into scales such as vocabulary, reading comprehension, and verbal analogies, each yielding a subscore. In practice, a battery would probably cover a larger domain than verbal ability; the problems and issues of construct validity generalize to any set of constructs. If the scales in fact measure what they are expected to measure, then the split-halves (or any other subdivisions of the scales) will demonstrate convergent and discriminant validity, as the diagram suggests. That is, the component-scores will show evidence that they have satisfactorily high relationships with the predicted factors and satisfactorily low relationships with the others. The statistical methods of maximum likelihood confirmatory factor analysis (Jöreskog, 1970) provide tests of goodness of fit to the postulated factor model.

Hypothetical Structure of A
Verbal Ability Battery



This analysis can be carried out using any particular populations, such as all examinees, all applicants to a particular institution, all examinees who use a particular special form (such as Braille), or examinees identified independently as sharing a specific handicap. Perhaps the most pertinent questions, however, involve comparison across groups. Do the tests measure the same constructs in different populations? Many comparisons are possible: blind and sighted students, deaf and hearing students, combined handicapped students and non-handicapped students, one handicapped group and another. There are statistical techniques for testing hypotheses about the comparability of factor structure, scale units, and precision of measurement across populations. For example, one could test the hypothesis that a set of verbal analogy items measures the same factor in deaf and hearing populations.

Construct validation research would substantially add to our knowledge gained by predictive validity research but there are substantial problems. In one way, the sample size problem is less severe than is the case with predictive validity research. Since criterion data are not required, subjects can be pooled across years and institutions and systematic record keeping and data retention would provide a valuable data base. The difficulty lies in the sheer number of variations to be examined. Handicaps vary in severity as well as in type. For example, degrees of partial sightedness and partial hearing may have significant impact on test and educational performance. Findings may not generalize over groups of students with the same kind of handicap but different degrees of severity. Therefore, the groups that must be examined include those differing in severity as well as type of handicap.

Validation As A Model

The Panel on Testing of Handicapped People recommends that the required validity studies be reported in four years. It concludes that "current psychometric theory and practice do not allow full compliance with the regulations as currently drafted," but that "the technical problems of developing and validating tests that accommodate specific handicaps, while difficult, are not insurmountable."

What can we expect from the four years of psychometric work called for by the Panel? Predictions are hazardous but it seems probable that the Panel's report will stimulate substantial interest and activity and that in four years the research community will have produced new methods of measuring performance of the handicapped, new validation research on those measures, and many new insights into the test performance of handicapped people. But it seems unlikely that this work will have advanced far enough to permit confident testing by modified techniques of a major proportion of the relevant handicapped populations.

The actual and suggested research approaches I have referred to concentrate on a particular arena for test use and evaluation: the use of tests for competitive selection and the evaluation of tests by way of test validation. That emphasis will undoubtedly continue. This arena is the center of work by test developers and the focus of most of the issues that have arisen. It is the subject of Section 504 of the Rehabilitation Act of 1973, and therefore the focus of recommendations by the Panel on Testing of Handicapped People. In our comprehensive selection and validation, however, we should not lose sight of the fact that there are other approaches to the use and evaluation of tests for handicapped persons.

The major research effort that the Panel placed at the heart of its major policy recommendations should include efforts to develop ways of evaluating tests for handicapped persons to supplement traditional predictive validation methodology. Most test validation techniques are designed to assess statistical evidence of success in decision making over a large number of individuals. To apply that model to the handicapped who require modified tests is to assume that decisions concerning them will be made in large numbers; that, with modifications in tests, test scores will be comparable with those of nonhandicapped persons; and that the decisions to be made are comparable for handicapped and nonhandicapped persons. Each of those assumptions requires closer examination.

In a broader sense, the problems are indeed problems of validity because as Messick (1980) has persuasively argued, a comprehensive view of validity must include evaluation of the consequences of test use; that evaluation inevitably requires dealing with values. What I am suggesting is that values be introduced explicitly and examined openly; that participants in the process go beyond test users and test developers; and that the review not be limited to technical psychometric concerns. In particular, we need to consider optional decision-making mechanisms and optional test-evaluation models.

Decision-making mechanisms. Most of the discussion of test validity for handicapped persons has assumed that the situation is one of competitive selection; optional ways of making decisions can be envisioned. I assume that the values include maximum access by handicapped and fairness to non-handicapped persons who might be competing for a scarce acceptance.

1. No testing of handicapped persons. In many instances the best strategy might be to waive test requirements for an applicant whose handicap prevents taking the standard test and who meets educational requirements. This in effect means giving the handicapped applicant a chance to try the course of study or job--in effect a job-sample test. This policy would not put other students at a serious competitive disadvantage in cases where selectivity is absent or not severe--e.g., in most undergraduate admissions. Thus while this strategy would not be applicable to all selection situations, it would be applicable to a very large proportion of them.

2. Clinical testing of handicapped persons. In cases in which waiver of testing is not considered appropriate, a more valuable array of information might come from individualized testing of the handicapped applicant. Such assessment would include a number of perspectives--e.g., of educators, clinicians, specialists in the handicapping condition--and would involve professional judgment. The goal would be to make decisions in the best interest of the individual, rather than that of the institution, although the two need not conflict.

3. Guidance as a model. Locating the best opportunity for the handicapped applicant is as important and difficult as selecting among applicants for a given opportunity. Assessment for handicapped applicants could focus on identification of strengths and needs directed toward identifying educational programs and institutions best capable of meeting those needs.

4. Matching students and institutions. This strategy, an extension of the guidance model, would provide a data base on institutions permitting detailed assessment of how well they would meet particular needs. The data would include information on programs of study, counseling, and availability of special services to the handicapped.

Many other approaches are possible. These approaches have in common an increased focus on the individual, as compared to the institution. Perhaps these functions could be carried out by a new kind of service organization that provides assessment and counseling services to the handicapped and information and, where appropriate, advocacy to institutions.

Test evaluation models. If decision-making procedures were modified in any of these directions, new approaches to the evaluation of tests would be called for. Test validation, in its traditional institution-focused form, would be insufficient. The goal would be to evaluate decision-making methods with regard to their value to the individual and to society at large as well as to educational institutions. New concepts of test evaluation would be needed. Likely approaches, for example, would be ones that study the experiences of handicapped persons longitudinally and that permit evaluation of a decision in terms of its effects on the individual.

References

Braun, H. & Jones, D. H. The Graduate Management Admissions Test prediction bias study. Final Report to the Graduate Management Admissions Council. Princeton, NJ: Educational Testing Service, 1981.

Jones, D. H. & Ragosta, M. Predictive Validity of the SAT on Two Handicapped Groups: The Deaf and the Learning Disabled. Princeton: ETS, 1981.

Jöreskog, K. G. A general method for analysis of covariance structures. Biometrika, 1970, 57 (2), 239-251.

Messick, S. J. Test validity and the ethics of assessment. American Psychologist, 1980, 35, 1012-1027.

Nester, M. A. Testing Handicapped Persons for Employment. Paper presented in American Psychological Association, 1980.