DOCUMENT RESUME

ED 141 383 TM 006 270

AUTHOR Petrosko, Joseph M.; Shani, Esther
TITLE Structural Components Revealed by Evaluating the Quality of Elementary School Tests.
PUB DATE Apr 77
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, New York, April 5-7, 1977); for related document, see Journal of Educational Measurement, v13 n4 p283-96, Winter 1976

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Comparative Analysis; Correlation; Elementary Education; *Evaluation; *Evaluation Criteria; Factor Analysis; Guides; Rating Scales; *Standardized Tests; Statistical Analysis; *Test Reliability; Test Reviews; *Test Selection; Test Validity
IDENTIFIERS *MEAN Test Evaluation System

ABSTRACT

An analysis was made of quality ratings of elementary level standardized tests. Applying multidimensional scaling to intercorrelations of quality ratings, it was found that the criterion of content and construct validity was a central element in the evaluation of elementary tests. In addition, several types of validity and reliability were found to be closely related to one another. The study confirms previous results obtained with secondary tests, and has implications for test selection by evaluators and program directors. (Author)

Structural Components Revealed by Evaluating the Quality of Elementary School Tests*

Joseph M. Petrosko
University of Louisville

and

Esther Shani
Jerusalem, Israel

* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, April 1977.

This study was performed to test a theory and, on a more practical level, to report results useful to people who must select tests. Finding the appropriate test for a particular educational purpose presents problems for many researchers. Difficult judgments are involved. Although lip service is paid to the principle that selecting a standardized test requires a thorough consideration of many factors, in fact, the principle is rarely followed. And not surprisingly so. There are thousands of tests on the market, and trying to judge which one might be best for a particular purpose is simply beyond the resources of many test users.

In order to provide a simple-to-use but detailed quality guide in this area, a comprehensive rating system for the evaluation of tests has been developed (Hoepfner, Conniff, Petrosko, Watkins, Erlich, Todaro, Hoyt, McGuire, Klibanoff , Stangel, Lee, Rest, Hufano, Bastone, Ogilvie, Hunter, & Johnson, 1974; Hoepfner, Stern, & Nummedal, 1971; Hoepfner, Strickland, Stangel, Jansen, & Patalino, 1970). Using this system, numerical ratings of educational and psychometric quality can be used to compare standardized tests. The ratings reflect criteria grouped into four general areas of test quality: Measurement Validity, Examinee Appropriateness, Administrative Usability, and Normed Technical Excellence (yielding an acronym for the evaluation system - MEAN).

With some variations, the procedure used in implementing the MEAN system was similar each time that it was applied. The evaluation process was initiated by the acquisition of virtually all published tests at the relevant grade levels. Tests were then categorized into educational

goal areas and evaluated against the MEAN rating scales. At least two persons, working independently, performed the ratings. (A third rater was used when there were disagreements between the first two.) The final outcome was the publication of the ratings in books available to test users.

The ratings were primarily conceived as a source of comparative information on those tests which were designed to measure the same general outcome. Using this "consumer's guide," a person could, for example, compare various tests in reading comprehension with one another. After examining the strengths and weakensses of various instruments, a selection could be made of a test suitable for a given set of educational circumstances.

The ratings are also useful for another purpose, however. They can be used to examine the quality of tests in general, and to discover how the various elements of test quality relate to one another. Questions like these can be addressed. How do the rated validity, reliability, and score distribution characteristics of tests relate to one another? Are reliable tests valid? Are tests with good norms also generally possessed of a good physical format?

As a vehicle for answering questions like these, Esther Shani proposed a theory for the quality structure of standardized tests (Shani & Petrosko, 1976). Using data from evaluations of secondary school tests (Hoepfner et al., 1974), the theory successfully predicted a structural configuration to explain the correlations of quality ratings.

To explore the generalizability of this theory, the present study was undertaken. Correlations obtained from elementary level tests (Hoepfner et al., 1970) were analyzed to determine if the theory

developed for secondary level tests would still be applicable to another age level. The analysis could also provide useful information for test users.

METHOD

The Theory

The theory employed in this study follows directly from the study of Shani & Petrosko (1976). Adaptations were made, where necessary, to reflect the differences between the MEAN evaluation system as employed with elementary tests.

Given the requirements of the study--conceptualizing and relating a number of variables to one another--the obvious need was for a technique of conceptualization and analysis suitable for a set of multivariate data. Facet theory developed by Guttman (1965) offered the advantage of a well developed method for linguistically processing the many variables involved and also providing a link to an analytic technique for mathematical processing of the data. Facet theory has been applied to such content areas as attitude measurement (Mori, 1965) and intelligence testing (Schlesinger & Guttman, 1969) and is a general approach to research applicable to any content area where sets of variables can be identified in terms of more basic sets or facets.

Examination of the MEAN test evaluation criteria for elementary school tests revealed an emergent theory about a structure for evaluating standardized tests. The overall outlines of this theory might be drawn by asking two questions: (a) What are components of a test evaluation that are inherent in the construction and development of the test?; (b)

What is the relationship of the test development process to the examinee?

The two considerations could be expressed as two facets (or sets).

Facet A Components of a test evaluation inherent in the test and its development. (Five elements)

a_1 Theoretical conceptualization of the test

a_2 Characteristics and format of items

a_3 Test instructions

a_4 Empirically determined validity and reliability

a_5 Test scores and norms

Facet B Relationship of the test development process to the examinee. (Three elements)

b_1 Initial test construction activities

b_2 Standardization and refinement of a test through sampling from a population

b_3 Direct contact between the test and examinee

The elements of facet A are assumed independent of one another and, therefore, define it as a polarizing facet. Criteria related to these elements would emerge as independent factors in a factor analysis. Facet B could be defined as an ordered facet, with each element showing a different degree of relationship between the examinee and the test's development.

The elements of facet A relate to independent aspects of a standardized test about which quality assessment can be made. For example, one can ask the question: what evidence does a particular test present that sufficient efforts were taken in its theoretical conceptualization (a_1)

or in the way items were written (a_2) to operationalize the theoretical conceptualization? Similar questions can be asked about the remaining three elements.

Facet B contains three elements--all related to the degree in which the development of a test relates to an examinee. Element b_1, *Initial test construction activities*, has the smallest relationship between an individual examinee and the test's development. During item writing activities, the authors typically have no specific individual in mind, and construct items for a broad spectrum of examinee types within the general constraints of the age level intended. *Standardization and refinement of a test through sampling from a population*, element b_2 involves activities which are more closely related to an individual examinee who will eventually take the published test. For example, validity are reliability studies carried out by a test developer would be associated with this element. Finally, element b_3 shows the closest relationship between the test and the test-taker. *Direct contact between the test and examinee* generally involves aspects of the actual test-taking situation, e.g., format and clarity of tems. In summary, the elements of facet B may be considered to lie on a continuum spanning the degree of relationship between a test developer and a person actually taking a developed test.

A structure for evaluating a standardized test in terms of facets A and B can be defined in the following mapping sentence:

The quality of test (x) with respect to component

a_1 Theoretical conceptualization

a_2 Item characteristics and format

a_3 Test instructions

a_4 Empirical validity and reliability

a_5 Test scores and norms

at the

least (initial construction)

medium (standardization on population)

highest (direct contact

level of relationship with the examinee → very high to very low quality

According to Guttman (1970), concepts dealt with by two facets, one of which is polarizing and the other ordered, tend to show a radex structure in the analysis of empirical data based on the facets. It was hypothesized that analysis of data from the MEAN evaluations of elementary school tests would yield such a radex structure.

The analysis that was seen as most appropriate was Guttman's Smallest Space Analysis (SSA). The latter, as are several other nonmetric multi-dimensional scaling techniques, is based upon a simple principle: the higher the correlation between two variables, the smaller is the represented distance between two points representing the variables. If $r_{12} > r_{34}$, then $d_{12} < d_{34}$, where: r = correlation coefficient; d = distance in space.

Test evaluation criteria

For the elementary school test evaluations, tests were acquired and evaluated for grades 1, 3, 5 and 6. Trained raters used only the specimen

tests and other supporting material sent by the publisher. Each test was first categorized into one of 145 goals of elementary education. These goals constituted a comprehensive taxonomy of elementary education in terms of student outcomes. After this categorization, evaluators rated the test on the 24 criteria of the MEAN system. For each criterion, each test was awarded zero to a specified number of points, depending on its possessing the desired trait in question.

Table 1 shows each criterion, its facet profile (each criterion being a structuple of facets A and B), and the range of possible points a test could receive for the criterion. Complete descriptions of the criteria are contained in Hoepfner et al. (1970). It might be noted that the criteria differ somewhat from those used with secondary school tests and analyzed by Shani and Petrosko (1976).

Table 1

Elementary Test Evaluation Criteria with Facet Profiles and Ranges of Points Awarded

Profile		Criterion	Range
a_1b_1	1.	Content/Construct Validity	0-10
a_4b_2	2.	Concurrent/Predictive Validity	0-5
a_2b_2	3.	Content Comprehension	0-4
a_3b_3	4.	Instructions Comprehension	0-4
a_2b_3	5.	Visual Format	0-2
a_2b_3	6.	Quality of Illustrations	0-1
a_2b_3	7.	Time and Pacing	0-1
a_3b_3	8.	Response Recording	0-2
a_5b_3	9.	Test Administration (Group)	0-2
a_5b_3	10.	Training of Administrators	0-1
a_5b_3	11.	Administration (Time)	0-1
a_5b_2	12.	Scoring	0-2
a_5b_2	13.	Norm Range	0-1
a_5b_2	14.	Score Interpretability	0-1
a_5b_2	15.	Score Conversion	0-2
a_5b_2	16.	Norm Representativeness	0-1
a_5b_2	17.	Score Interpreter	0-1
a_1b_3	18.	Decision-Making Utility	0-3
a_4b_3	19.	Test-Retest Reliability	0-3
a_4b_2	20.	Internal Consistency Reliability	0-3
a_4b_3	21.	Alternate Form Reliability	0-3
a_5b_2	22.	Replicability	0-1
a_1b_1	23.	Range of Coverage	0-3
a_5b_2	24.	Gradation of Scores	0-3

The first two ratings deal with validity. Content and construct validity referred to whether the test measured the specific educational objective that the test was categorized under. Concurrent and predictive validity referred to evidence that such validity studies had been performed.

Evaluation criteria 3 through 8 were related to the general theme of Examinee Appropriateness. Content comprehension and Instructions comprehension dealt with the perceived clarity of the items themselves and of the test's overall instructions. The criteria Visual format and Quality of illustrations had to do with physical arrangement of items on the page and quality of printing and graphics. Time and pacing required a judgment about whether an instrument was a power test or was unnecessarily speeded. Response recording related to whether there was a simple and direct connection between the item stem and the recording of a response.

The next set of evaluation criteria - 9 through 18 fell under the general area of Administrative usability. Criterion 9, Test Administration, gave tests a positive rating if they were designed for group rather than individual or small group administration. Training of Administrators was used to downgrade those tests requiring a psychometrist to administer. The criterion Administration credited those tests that could be administered in a typical class period of time. Criterion 12, Scoring, gave tests optimal points for simple and objective scoring procedures. Norm Range was used to evaluate if the norm sample was broad in age range. Score interpretability related to whether converted scores were of a well known type (e.g. percentiles). Score Conversion gave credit to tests with a simple conversion procedure from raw score to standard score.

Norm Representativeness credited those tests with well represented norm samples from the student population. Score Interpreter gave a point to tests that could be interpreted by the school staff. Finally, criterion 18, Decisions gave maximum credit to tests where a definite prescriptive decision could be made about a student (the more prescriptive, the better).

The last set of criteria were in the area Normed Technical Excellence. Criteria 19 through 21 were used to give tests credit if they reported high coefficients of Test-Retest, Internal Consistency and Alternative Form Reliability. The criterion Replicability gave tests more credit for replicable procedures for obtaining scores. The Range of coverage criterion was used to award points to those instruments aimed at providing information for a wide range of some behavior domain. Finally, Score Gradation gave tests maximal credit for useful converted scores such as centiles rather than crude scores like pass/fail.

Analysis

A 24 x 24 matrix of correlations was derived from a report by Hoepfner (1971). The matrix was generated by correlating ratings on each criterion with one another. Ratings for sixth grade tests (N = 508) were analyzed. The matrix was used as input for the multidimensional scaling program, SSA-1 (Guttman, 1968; Lingoes, 1973; Roskam & Lingoes, 1970). The latter represents the distances between points in space so that positively correlated items are close together and items that correlate zero or negatively are far apart. Two measures of the adequacy of the solution are provided, both of which the program algorithm attempts to minimize in iterative steps: Kruskal's stress coefficient and the Guttman-Lingoes coefficient of alienation.

RESULTS

A solution for three dimensions was selected for presentation (Kruskal's stress = .11, Guttman-Lingoes coefficient of alienation = .12). A plot of two dimensions of this solution (vector 1 against vector 2) is presented in Figure 1. The numbers in Figure 1 correspond to the 24 variables listed in Table 1.

The plot reveals a radex pattern very similar to that obtained by Shani & Petrosko (1976). The plot, generally speaking, shows most variables located in space where they would be expected, based on the theory.

There were several reasons for discrepancies from theoretically predicted locations. First, an obvious reason presents itself - there was an imperfect match between the theoretical conception and the empirical reality. The rational considerations used in constructing the theory were not in all cases borne out by how tests are actually rated on their quality. Secondly, several of the criteria in this analysis were not represented in the analysis of secondary school tests. Such variables were assigned facet profiles based on a more-or-less common sense consideration of Shani's theory. For example, variables 9 through 11 in this study had no clear equivalents among the 25 variables analyzed by Shani and Petrosko (1976).

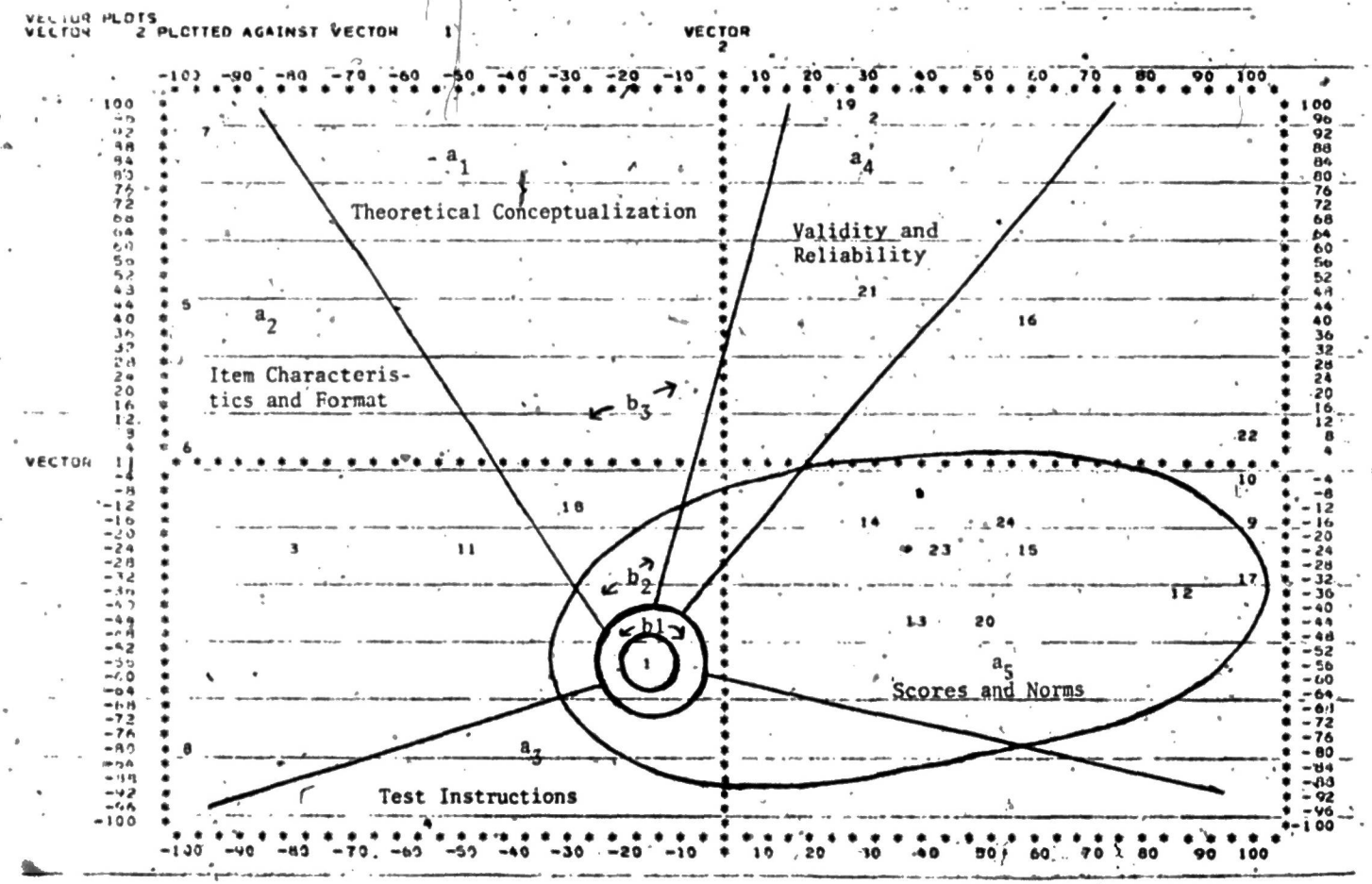

Figure 1. Radex for 24 test evaluation variables.

DISCUSSION

It is important not to lose sight of the practical implications of this study. Smallest Space Analysis produces configurations that show how variables are interrelated - the closer the relationship, the smaller the distance between them. An inspection of Figure 1 shows which aspects of elementary school tests are related to one another and which are not. Some of the more pertinent results bear discussion.

Note zone a_5b_2. Many of the variables related to quality of scores and norms were closely related to one another. Based on empirical analysis of an actual population of tests, the quality of norm range (variable 13) was closely related to such things as quality of score graduation (variable 24) and score interpretability (variable 14). Tests strong in one of these areas also tended to be strong in the other areas.

An interesting finding was the great divergence between variable 1, Content/Construct validity and variable 2, Concurrent/Predective validity. Variable 1 formed the center of the radex and variable 2 landed up toward the top part of area a_4b_3. In effect, whether a test was judged as adequate in covering the content of a goal area and as having "face valid" items, had little to do with existence of empirical validity studies for the test.

Variable 19, Test-Retest Reliability and variable 21 were found as hypothesized, in the a_4b_3 zone of Empirical Validity and Reliability. However, variable 20, internal consistency reliability, was relatively independent of the other reliability types.

Finally, some of the often neglected aspects of tests - physical format of items, quality of printing - were not related to the central issues of validity and reliability and only somewhat related to one another.

As a concluding note, it might be well to pay heed to the results in terms of practical decisions about tests that many of us make. The quality of a standardized test is not a unitary concept, but multivariate. Whether a test might be strong in one type of reliability may have little to do with its strengths in other areas. Mundane, but in some cases crucial aspects - like a test's format for recording student responses - should be assessed seperately from its other characteristics. Especially when a test will be used for a special purpose, or with a special population, it should be judged on many independent criteria.

REFERENCES

Guttman, L. A faceted definition of intelligence. In R. Eiferman (Ed.), Studies in psychology, scripta hierosolymitana (Vol. 14). Jerusalem, Israel: The Hebrew University, 1965.

Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, 1968, 33, 469-506.

Guttman, L. Integration of test design and analysis. In P.J. Dubois (Chair), Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, J.J.: Educational Testing Service, 1970.

Hoepfner, R. A test of tests. Center for the Study of Evaluation Report No. 69. Los Angeles, University of California, 19

Hoepfner, R., Conniff, W., Jr., Petrosko, J.M., Watkins, J., Erlich, O., Todaro, R.S., Hoyt, M.F., McGuire, T.C., Klibanoff, L.S., Stangel, G.F., Lee, H.B., Rest, S., Hufano, L., Bastone, M., Ogilvie, V.N., Hunter, R., & Johnson, B.L. CSE secondary school test evaluations (3 vols.) Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hoepfner, R., Stern, C., & Nummedal, S.G. CSE/ECRC preschool kindergarten test evaluations. Los Angeles: Center for the Study of Evaluation, University of California, 1971.

Hoepfner, R., Strickland, G., Stangel, G., Jansen, P., & Patalino, M. CSE elementary school test evaluations Los Angeles: Center for the Study of Evaluation, University of California, 1970.

Lingoes, J.C. The Guttman-Lingoes nonmetric program series. Ann Arbor, Michigan: Mathesis Press, 1973.

Mori, T. Structure of motivation for becoming a teacher. Journal of Educational Psychology, 1965, 56, 175-183.

Roskam, E., & Lingoes, J.C. MINISSA-I: A Fortran IV (G) program for the smallest space analysis of square symmetric matrices. Behavioral Science, 1970, 15, 204-205.

Schlesinger, I.M., & Guttman, L. Smallest space analysis of intelligence and achievement tests. Psychological Bulletin, 1970, 15, 204-205.

Shani, E., & Petrosko, J.M. Structural components derived from evaluating standardized tests. Journal of Educational Measurement, 1976, 13, 283-296.