DOCUMENT RESUME

ED 056 096 TM 000 907

AUTHOR Pascale, Pietro J.
TITLE Innovations in Item Scoring Procedures.
PUB DATE [71]
NOTE 11p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Computer Oriented Programs; *Educational Diagnosis; *Educational Innovation; Item Analysis; Multiple Choice Tests; *Response Style (Tests); *Scoring; Scoring Formulas; Testing Problems; Test Reliability; Test Validity; Thought Processes; Weighted Scores
IDENTIFIERS *Confidence Testing

ABSTRACT

This brief review explains some alternate scoring procedures to the classical method of summing correct responses. The novel procedures attempt in some way to retrieve and use even the information in the wrong responses. (Author)

# INNOVATIONS IN ITEM SCORING PROCEDURES



Pietro J. Pascale

TRENTON STATE COLLEGE

ABSTRACT

This brief review explains some alternate scoring procedures to the classical method of summing correct responses. The novel procedures attempt in some way to retrieve and use even the information in the wrong responses.

In the psychometric literature there have been studies proposing new ways of using item responses other than the method of summing of correct responses.

Research in the diagnostic value of multiple choice tests has usually led to discussion of differential weighting of the incorrect or inappropriate options. The differential weighting assumes a priori the possibility of at least rank-ordering of the incorrect options. Guttman & Schlesinger (1966) have developed a method called facet design which generates systematic construction of distractors which differ in degree of attraction. Facet design solves the problem of assigning meaningful differential weights to each response. Diagnostic development is [illegible] by using a deviate form of Raven's Progressive Matrices. The middle cell of a 3 x 3 matrix is used as the stimulus thereby taking advantage of an added function in the diagonal. The Raven matrix uses the extreme lower right cell as the stem stimulus.

Historically, option weighting goes back to E. K. Strong with his work on interest inventories (Strong, 1943). Strong noted that there are no

"correct" responses. Options were weighted empirically which discriminated various occupational groups. Responses to items then were used as variables in a discriminant function analysis which differentiated occupational groups.

There have been repeated suggestions in the literature of getting at the process involved in a response rather than simply scoring answers as 'right' or 'wrong.' The so-called 'wrong' answers can sometimes convey information (frequently of a diagnostic nature if the test is properly designed) concerning the process of human thinking (Laurendeau and Pinard,1962).

Glaser, Demarin, and Gardner (1954) have developed a procedure called the tab item which reveals the strategy used in problem solving in trouble shooting situation. A record is made of the sequence of steps taken and the number of steps needed to arrive at the correct answer. Coffman (1967) has called the tab item a "test item with feedback."

Nedelsky (1954) has suggested the rewarding of the ability to avoid gross errors. He devised a procedure for distinguishing the D students from the F students. The F students were determined by the incor-

dinate amount of options chosen which were referred to as ridiculously implausible. Poor students who at least demonstrate the ability to avoid gross errors received D's. Lord and Novick (1968) refer to these gross errors as worst distractors:

> If we wish to recognize the possibility of partial information or perhaps misinformation, then we can assign different scores to the various incorrect responses. For example, one distractor might be designed to ferret out common misinformation. We might call such a distractor, which is literally the least correct response, the worst distractor. A possible scoring scheme might assign a score of one to the correct response and a score of -S to a worst distractor response where $0<S<1$ (p. 313-314).

Dressel & Schmid (1953) have derived a scoring formula based on the assurance of a given answer. Schuford & Massengill (1966) led on by the expectation of extracting "all of the potentially available information" devised a scoring system to maximise score if the student expresses his 'degree of belief probabilities'. The formal procedures used in both of these studies can not be applied to testing very young children because of the verbal content of the instructions. However, the rationale of both of these studies closely parallels what Piaget (1929) has called conviction.

Coombs, Milholland, and Womer (1956) devised a novel measurement procedure. The individual selects and marks the distractors rather than the correct answers of each item. The rationale for this technique of scoring is that even though an examinee does not know the correct answer he may, nevertheless, know that one or more distractors are wrong. The phenomenon of knowing that certain distractors are incorrect is called partial knowledge. Testing for partial knowledge has little intuitive appeal for use with very young children but the procedure of forcing a scanning strategy of all options is worth investigating.

Davis and Fifer (1959) a priori weighted options and reported an increase in reliability from .68 to .76. The a priori weighting was devised by a panel of judges who qualitatively ranked the options of each item.

Guttman's (1941) procedure consisting of criterion keying of the options probably holds the most psychometric promise. Criterion keying of options rather than item weighting may give clues concerning the process underlying responses (Sigel, 1963).

The Guttman procedure also seems worthwhile investigating since its main concern is with validity whereas studies such as Davis and Fifer (1959) and Jacobs

and Vandeventer (1970) were primarily concerned with augmenting reliability with its small but concomitant validity effect only a secondary concern.

The Jacobs and Vandeventer (1970) study used the facet design analysis of Guttman to a priori weight options on Raven's Coloured Progressive Matrices. The procedure resulted in a statistical increase in reliability. The authors have little to say concerning the possibility of their technique contributing in the area of validity.

Birnbaum (Lord & Novick, 1968) has developed a three-parameter logistic latent-trait model which weights items by level of difficulty. Birnbaum's model has led to the development of sequential or tailored testing procedures of Novick (Lord & Novick, 1968). Sequential testing is still in the experimental stage but its feasibility has been partly supported by claims of high reliability coefficients. The most promising outcome of sequential testing may prove to be the use of the computer both in test administration and test scoring.

In general, the psychometric studies outlined here minimize their potential contribution to testing by overindulging in the domain of reliability.

This review of the psychometric literature is not directly relevant to the testing problems of young children. The attempt has been made to investigate the novel ways of using the individual options in a test item. The conventional psychometric way of using information in a test item is to score the item as 1 if the response is congruent with the keyed answer and to score the item as 0 otherwise. The total test score is then given as the sum of the correct items. There is potential information imparted within a wrong response. The classical psychometric model ignores this information.

The quest for new ways of using all the information in a test item naturally has led to item weighting and partial scoring procedures. Many of the novel procedures can not be applied directly to the testing of young children because of the language limitations of young children.

Some of the more promising procedures seem to be facet design analysis, branching items, and computer based testing. The hallmark of innovative procedures in item scoring has been the overall concern for reliability and general disregard for the more rigorous treatment of validity. Notable exceptions have been the study of gross error (Nedelsky, 1954) and the tab item (Glaser, Demarin, and Gardner, 1954). These two approaches attempted to get at the process underlying a test response.

The truly diagnostic test should reveal information concerning both what the subject knows and does not know. Diagnostic tests have been around for some time. These tests generally reveal information by the binary situation of subject either passing or failing a test item. Tests are needed wherein each response option of each item reveals a certain amount of diagnostic information.

REFERENCES

Coombs, C. H., Milholland, J. E., and Womer, J. F. The assessment of partial knowledge. Educational and Psychological Measurement, 1965, 13, 13-37.

Davis, F. B. Estimation and use of scoring weights for each choice in multiple choice items. Educational and Psychological Measurement, 1959, 19 291-298.

Davis, F. B. & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

Dressel, P. L. & Schmid, J. Some modifications of the multiple choice item. Educational and Psychological Measurement, 1953, 13, 574-595.

Glaser, R., Damarin, D. E. & Gardner, F. M. The tab item: a technique for the measurement of proficiency in diagnostic problem solving. Educational and Psychological Measurement, 1954, 14, 283-293.

Guttman, L. Supplementary study B, pp. 251-364, in P. Horst (Ed.), The Prediction of Personal adjustment. New York: Social Science Research Council, 1941.

Guttman, L. & Schlesinger, I. M. Systematic construction of distractors for ability and achievement test items. Educational and Psychological Measurement, 1967, 27, 159-170.

Jacobs, P. I. & Vandeventer, M. Information in wrong responses. Psychological Reports, 1970, 26, 311-315.

Laurendeau, M. & Pinard, A. Causal thinking in the child. New York: International University Press, 1962.

Lord, F. M. Formula scoring and validity. Educational and Psychological Measurement, 1963, 23, 663-672.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 1954, 14, 459-472.

Piaget, J. The child's conception of the world. London: Routledge and Kegan Paul, 1929.

Raven, J. C. Guide to using the Coloured Progressive Matrices. London: Lewis, 1965.

Shuford, E. H., Albert, A., and Massengill, H. E. Admissable probability measurement procedures. Psychometrika, 1966, 31, 125-145.

Sigel, I. E. How intelligence tests limit understanding of intelligence. Merrill-Palmer Quarterly, 1963, 9, 39-56.

Strong, E. K. Vocational interests of men and women. Stanford, Calif.: Stanford University Press, 1943.

# D
# HE
# O N
# AR E
# IO E
# T O
# C
# IL M