DOCUMENT RESUME

ED 111 862 TM 004 829

AUTHOR Kocher, A. Thel
TITLE An Empirical Investigation of the Stability and Accuracy of Flexilevel Tests.
PUB DATE [Apr 74]
NOTE 9p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS *Ability; Comparative Analysis; *Computer Oriented Programs; Feedback; Individual Differences; *Response Style (Tests); Scoring; Simulation; Student Testing; *Testing; *Test Reliability; Test Validity
IDENTIFIERS *Flexilevel Test

ABSTRACT

The purpose of the present study was to empirically investigate the stability and accuracy of one suggested method for matching test difficulty to examinee ability level. Students' answers to traditional classroom tests were rescored by computer as if the examinations had been flexilevel tests. The scores thus obtained were found to correlate highly with the traditional test scores (0.8994 to 0.9478), thereby indicating that flexilevel test scores are sufficiently stable and accurate to allow their use for classroom evaluation purposes. (Author)

AN EMPIRICAL INVESTIGATION OF THE STABILITY AND ACCURACY OF FLEXILEVEL TESTS

A. Thel Kocher

Minneapolis Public Schools
Southeast Alternatives Project

Presented at the 1974 Convention of the National
Council on Measurement in Education, Chicago, Illinois

Introduction

Conventional educational testing procedures are, to some extent, inefficient and inaccurate because they require that a heterogeneous group of individuals attempt every test item. Wood (1969) has suggested that measurement might well be made more efficient and more accurate if pupils could be routed through a test so that they spend most of their time working on items appropriate to their ability level. In the case of classroom evaluation, the use of a testing procedure which matches the difficulty of the items administered to student ability level seems desirable from at least two points of view. First of all, the student will not be required to answen as many questions as in a conventional test, thereby reducing testing time. Secondly, since students will not be required to answer items not geared to their general ability level, they will encounter fewer failures which seems psychologically desirable.

If a procedure for matching test difficulty to examinees on the basis of the examinee's ability level is to be acceptable for use as a classroom evaluation technique, then questions about the stability and accuracy of scores on such an instrument must be investigated. More specifically, studies should investigate the correlation between scores from such an instrument and scores obtained from a conventional test of the same objectives. The purpose of the present study was to do this empirically for one suggested method of matching test difficulty to examinee ability level.

Review of Related Research

Tests which permit this kind of measurement have been called "branched," "computer assisted," "individualized," "programmed," "sequential," and, perhaps most recently, "tailored." Although these different types of tests may differ somewhat, basically they require all pupils to begin with the same item; however, the items they subsequently encounter are always dependent upon their response to the item they have just answered.

One of the earliest studies on the subject was a dissertation by Patterson (1962). Patterson used probability models and hypothetical populations and found that, for the models considered, the sequential test discriminated better at the extremes than did the conventional test.

Bayroff and Seeley (1967) administered a verbal and an arithmetic reasoning branching test to 102 subjects. The branching was based on item difficulties and each pupil responded to either eight or nine items depending on the particular branch he followed. As part of the study a conventional 50-item verbal test and a 40-item arithmetic reasoning test were also administered. Correlations between the conventional and branching tests ranged from .74 to .78.

In a study by Wood (1969) three different "tailored" mathematics tests were prepared. The tests consisted of four, five, and six items respectively and were administered to a sample of 91 students. The results on each of the "tailored" tests were then correlated with mathematics grades with the highest correlation being .35.

Linn, Rock, & Cleary (1968a; 1968b; 1969) conducted studies which used existing item data for 4,885 eleventh grade students on the 190 verbal-type items of the SCAT and STEP. In all, the researchers developed seven different programmed tests. The tests differed from each other primarily in the ways in which subjects were routed through the test. For five of the experimental tests, two different scoring procedures were used. Thus, a total of twelve different programmed tests were investigated. Correlations between the programmed and conventional tests ranged from .8738 to .9663. The experiment also examined shortened conventional type tests and it was found that a 50 item test would produce about the same correlation as the best of the programmed type.

Lord (1970) reported the following requirements of "tailored" tests.

1. Development of a large number of items for pretesting, perhaps on the order of several thousand.
2. A very large pretesting to obtain adequate data for statistical analysis of each item.
3. A possibly dubious but very complex statistical analysis of pretest item data to estimate the necessary item parameters in advance of the main testing.
4. A final pool of 500-5000 selected items, for actual administration.
5. Computer simulations of perhaps a hundred different tailoring strategies and scoring methods in order to select item-administration and scoring procedures that will provide accurate measurement at all ability levels.
6. Test administration by a computer at terminals equipped with teletypes and visual display devices.
7. Experimental testings and statistical analyses to demonstrate to the testing agency, to skeptical examinees, and to their lawyers that the scoring method is fair, in the sense of assigning approximately the same score to an examinee regardless of which subtest of items he happens to take (pp. 1-2).

The Flexilevel Test

It would seem that, in light of the aforementioned requirements, the use of the "tailored" test is beyond the reach of the typical classroom teacher; and, indeed, of many standardized test developers. To a large degree the matching of item difficulty level with ability level can also be accomplished by what Lord (1970; 1971) calls the flexilevel test. The flexilevel test avoids many of the disadvantages of "tailored" tests and, thus, seems a more promising instrument especially for locally-constructed tests.

A conventional test may be modified to become a flexilevel test when the items are arranged approximately in order of difficulty. The idea of a flexilevel test is that the examinee begins with the middle item on the test and receives immediate feedback on his response. After each correct response he proceeds to the next hardest unanswered item. When he answers an item incorrectly he attempts the next easiest unanswered item. He continues until he has answered $\underline{n} = (\underline{N} + 1)/2$ items, where $\underline{N}$ is the number of items on the conventional test.

A theoretical study of the measurement properties of the flexilevel test (Lord, 1971) showed that:

> Near the middle of the ability range for which the test is designed a flexilevel test is less effective than is a comparable peaked conventional test. In the outlying half of the ability range, the flexilevel test provides more accurate measurement in typical aptitude and achievement testing situations than a peaked conventional test composed of comparable items (p. 813).

Procedure

In order to empirically investigate the stability and accuracy of flexilevel tests, the present study utilized data from five previously administered conventional objective examinations. Three of the tests considered were classroom examinations administered during a one semester junior level college course in introductory educational measurements. The tests contained 42, 36, and 36 items respectively. The remaining two examinations investigated were semester final examinations in a high school geometry course. These examinations consisted of 100 and 70 items respectively.

For each of the tests, each student's answers were rescored by means of a computer scoring program as if the test had been a flexilevel test. This test is referred to as a simulated flexilevel test. To investigate the relationship between the students' scores on the simulated flexilevel test and the traditional test, the Pearson product-moment coefficient of correlation was then calculated.

In addition to obtaining the correlation between each simulated flexilevel examination and the corresponding traditional test, the following additional evidence was obtained for the introductory educational measurements examinations. Two total scores were obtained for each subject in the course. The first of these total scores was obtained by summing the subject's standard scores on the traditional exams and the second by summing the standard scores on the simulated flexilevel exams. The two sets of total scores were then correlated.

Data Source

The data for the three educational measurements examinations used in the study were obtained from approximately 180 students enrolled in an introductory educational measurements course at the University of Kansas during the Spring semester of 1971. The data on the first geometry examination were obtained from 412 students enrolled in a geometry course at Shawnee Mission South High School, Shawnee Mission, Kansas. The data on the second geometry exam were obtained from 485 students enrolled in a geometry course at the same high school. The first geometry exam was administered at the end of the Spring 1971 semester. The second was administered at the end of the Fall 1971 semester.

Results

For the three introductory educational measurements examinations considered in the present study, the correlations between the original examinations and the corresponding simulated flexilevel test were 0.8994, 0.9464, and 0.9090 respectively. For the two high school geometry exams investigated, the correlations were 0.9353 and .9478. Interpreting the five obtained correlations as measures of parallel forms reliability (measures of equivalence and stability) indicates that flexilevel test scores could validly be substituted for scores obtained by administering a traditional test.

In the introductory educational measurements course the correlation between total scores based on traditional tests and total scores based on simulated flexilevel test scores was 0.955. The magnitude of this correlation further indicates that flexilevel test scores possess the necessary stability and equivalence characteristics that they may be substituted for scores on a traditional classroom examination.

Summary and Implications

The results of the present study demonstrate that scores obtained from simulated flexilevel tests can validly be substituted for traditional test scores of the same objectives. If further investigations using actual flexilevel tests in the classroom show the same high degree of relationship between traditional and flexilevel test scores, then teachers will have an easy to use method of making testing more efficient.

REFERENCES

Bayroff, A. G., & Seeley, L. C. An exploratory study of branching tests. *U. S. Army BSRL Tech. Rep. Note 188*, June, 1967. Cited by Cleary (1969).

Linn, R. L., Cleary, T. A., & Rock, D. A. An exploratory study of programmed tests. *Educational and Psychological Measurement*, 1968, *28*, 345-360. (a)

Linn, R. L., Cleary, T. A., & Rock, D. A. Reproduction of total test score through the use of sequential programmed tests. *Journal of Educational Measurement*, 1968, *5*, 183-187. (b)

Linn, R. L., Cleary, T. A., & Rock, D. A. The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 1969, *29*, 129-146.

Lord, F. M. The self-scoring flexilevel test. *Research Bulletin 68-38.* Princeton, N. J.: Educational Testing Service, 1970.

Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, *8*, 147-151. (a)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 1971, *31*, 805-813. (b)

Patterson, J. J. An evaluation of the sequential method of psychological testing. Unpublished doctoral dissertation, Michigan State University, 1962. Cited by Linn et al. (1969)

Wood, Robert. The efficacy of tailored testing. *Educational Research*, 1969, *11*, 219-222.