DOCUMENT RESUME

ED 109 152 TM 004 595

AUTHOR Stiles, Richard; And Others
TITLE Use of the Rasch Model With Content Referenced Tests.
PUB DATE [Apr 75]
NOTE 6p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Achievement Tests; Arithmetic; *Comparative Analysis; Educational Objectives; Measurement Techniques; Models; *Norm Referenced Tests; Performance; Reading Tests; *School Districts; Student Evaluation; *Testing Problems; Test Reliability
IDENTIFIERS Content Referenced Tests; *Rasch Model

ABSTRACT

The purpose of this study was to explore the usefulness of the Rasch test analysis model for providing consistent scale statistics between content-referenced and norm-referenced tests across school districts. A number of content-referenced tests were created to assess student performance relative to basic skills goals identified in reading and mathematics. Performance on these tests was related to performance on norm-referenced tests in two school districts. Based on preliminary analyses, it appears that the Rasch model offers promise for a fully flexible content-referenced test system providing information now available only through the use of norm-referenced tests. (Author)

3.12

USE OF THE RASCH MODEL WITH CONTENT REFERENCED TESTS*

by
Richard Stiles, Tacoma Public Schools
Walter Hathaway, Portland Public Schools
Fred Forster, Portland Public Schools

Objectives

This paper will review seven of the measurement development activities of the Northwest Evaluation Association. The overall system that some school districts in the states of Oregon and Washington are trying to develop and implement is the one described by Doherty and Hathaway (1972) in NCME's publication, Measurement in Education. The overall system contains three cross-referenced banks (a collection of course level curriculum goals; a collection of calibrated goal referenced measurement items/scales; and a collection of goal-referenced instructional strategies), the second of which is the major concern of the Northwest Evaluation Association.

For the past several months, members from the Northwest Evaluation Association have been empirically testing the Rasch model to explore its usefulness in providing scale statistics for content referenced measures. The Rasch model was examined to see if it would make the following possible:

(1) the identification of "problem" items;

(2) the identification of scaled scores between tests;

(3) the identification of the relationship between the content referenced tests and existing norm referenced tests; and

(4) the scaling of individual items for inclusion in an "item pool."

Theoretical Framework

The Rasch test model offers a promising approach to the scaling of tests, but as yet has seen only limited application in actual school settings. The most important aspect of the model is the interval scaling it provides for both test scores and individual test items on the underlying latent dimension. Using the interval scales provided by the model, it is theoretically possible to equate different tests or alternate forms of the same test by including a set of common items between the tests. In addition, the Rasch analysis provides an indication of those items which perform "questionably" and may be inappropriate or invalid. The present paper was intended to take advantage of the many desirable features of the Rasch model in analyzing content referenced tests developed to assess specific basic skills learning goals.

Studies

The first study was spearheaded by Drs. Hathaway and Forster of Portland Public Schools and involved a field test of 200 reading and 200 mathematics items approximately evenly divided between the fourth and eighth grades. The items were arranged into 26 "interlocking" tests so that each test shared approximately ten items with a preceding and a succeeding test. This arrangement of the items was intended to

*Prepared for presentation at the American Educational Research Association Annual Meeting, Washington, D. C., March 31-April 3, 1975.

make it possible to use the Rasch methodology to link together the analyses of each test. Each test form was given to approximately 250 students to insure as much stability as possible in the results. There were no other conditions, however, with regard to the students taking a test, since earlier studies (Wright 1967; Stiles 1974) had demonstrated that the Rasch yielded virtually identical results for bright students and slow students.

Each test was analyzed separately using the Rasch technique. First, each item was screened with respect to its discrimination and its fit to the Rasch model. The discrimination index used was the familiar point biserial correlation between the item and the raw score. The fit to the model is based on the discrepancy between the actual percent of students at each achievement level who get the item correct, and the predicted percent specified by Rasch equations. It was found that the two methods generally identified the same items as defective, but that the Rasch fit to the model approach did pick up some item defects missed by the point biserial. In particular, the fit appeared to be more sensitive to items which gave cues to the less able students that enabled them to guess the right answer, and subtle item defects that were only recognized by the most able students and therefore caused them to miss the item. After eliminating the faulty items, the next step was to "link" together the analyses from the separate tests. This was accomplished by using the characteristic of the Rasch that each item in a test is assigned its specific difficulty level, regardless of the students taking the test. Thus, while the average percent correct on an item differs dramatically between groups of slower students and brighter students, the Rasch difficulty value assigned that item remains constant (i.e., within the error of the estimate). Since this is also true of the same items when they appear on different tests, the organization of the tests (described above) made it possible to determine the average Rasch difficulty value for the same items appearing in the different tests. This, in turn, made it possible to equate the difficulties for the items which the tests did not have in common. For example, assume that test A and test B shared ten items whose average Rasch difficulty value was 55 on test A and 45 on test B. In equating the items on test B to the items on test A it would be necessary to add 10 to the Rasch difficulty value for each item on B. Similarly, if tests B and C had ten items in common and the average difficulty was 50 on B and 45 on C, then it would be necessary to add 15 to the Rasch difficulty of the C items to equate them to A. Following this procedure it was possible to develop single difficulty scales for fourth grade reading, fourth grade mathematics, eighth grade reading and eighth grade mathematics.

In a second study, an effort has been made to relate the difficulty of these items to the items in the current Portland citywide testing program. Since the citywide test was locally developed by districts in the Portland metropolitan area and free of publishers' copyrights, the completion of this research program would free us to use our items as a flexible item pool. This, in turn, would make it possible to develop several shorter tests aimed at specific achievement levels that would yield higher reliability and a smaller error of measurement than our current survey tests.

It should be mentioned here that an extensive effort has been made to relate each item to a learning goal which curriculum staff have indicated it measures. This procedure would make it possible to check the content coverage of a test as well as its measurement characteristics and to add calibrated items to "fill in the gaps." In this latter case, the Rasch shows promise in making it possible to equate the new test to the old test and maintain continuity of the normative data reported to the public. Experimentation is still in process with this equating aspect of the Rasch analysis.

In another vein, for a third study, two cooperative projects have been undertaken with the Parkrose (Oregon) and Tacoma School Districts to equate the Portland Basic Skills mathematics items to mathematics tests developed by them. By employing a similar design to that used in the original field test, each district has developed tests containing their items and ten of the Portland items. When the results of their test administration are available, it will be possible to equate the Rasch difficulties for all the items on a single scale. As well as enlarging the item pool, it will make the items of each district available to the other districts as well as available normative data on these tests. In this way, it is hoped it will be possible to explore cross-district comparisons based on a sound foundation.

A fourth major test development effort was spearheaded by Drs. Forbes and Ingebo of Portland Public Schools, and has focused on mathematics at the seventh grade level. Approximately 250 items were available from a pool of items developed for the seventh grade Portland Metropolitan Area Test. These items had been previously field tested, and percent right information was available for each item. Using this data, the items were arranged into four interlocking tests of increasing difficulty. The links between tests were also included in three additional tests which made it possible to cross-validate the link between each pair of tests. Each of these seven tests was given to approximately 300 students ranging from sixth graders to ninth graders, based on the relative difficulty of the test.

The tests were analyzed in much the same manner as that described for the Portland Basic Skills tests. In addition, the tests were analyzed separately by subtest (computation, concepts, and problem solving). The comparison of the overall links and these subtest links is the subject of an important paper being presented by Dr. Forbes at this conference. The results of the overall links between these tests were extremely encouraging. It was shown that the values used to equate items between the four basic tests agreed closely with the cross-validation values. This study established both the single scale for item difficulties and the robustness of the Rasch linking procedure in a single operation.

In a fifth study, Dr. Forbes has initiated an even more ambitious effort designed to explore the feasibility of using the Rasch analysis in designing shorter tests focused at a student's performance level to replace our current survey testing program. He has designed seven interlinking tests, each with approximately 30 items, arranged to represent increasing difficulty levels and balanced to represent computation, concepts and problem solving. These tests will be administered with the regular survey test in the spring of 1975 and the results analyzed to compare the reliability of these shorter more specific tests with that of the general survey tests. Obviously, this study will have significant implications for all future test development activities which the Northwest Evaluation Association will undertake.

In a sixth study, the Metropolitan Area Test Planning Board (supported by several school districts in the Portland metropolitan area) under the leadership of Dr. Forbes has undertaken a project designed to Rasch calibrate the previously developed survey tests. The mechanics of this study closely resemble the previously described studies (i.e., the development of overlapping tests administered to approximately 300 students each). This project will provide valuable data concerning the capability of the Rasch to equate scores from several tests to the same interval scale. This is extremely important since there is no existing adequate methodology for combining student scores on different tests. In addition, this study will make it possible to compare the norms developed for these tests across grades and possibly use them to establish checkpoints on continuous developmental scales in reading and mathematics.

Through the cooperative efforts of Leonard Winchell and William Conley of Tacoma Public Schools, Dr. Stiles conducted the seventh study. In this study an attempt was made to link a locally developed 48-item content referenced arithmetic computation test with a commercially developed norm referenced test. In the fall of 1974 fourth, fifth and sixth grade pupils were administered both the Mathematics Management By Objectives Test (MAMBO) and Form Q2 Arithmetic Section of the CTB/McGraw-Hill Comprehensive Tests of Basic Skills (CTBS).

Initially, correlations were computed on the item difficulties obtained through the Rasch analysis between each of the grade levels. The correlations ran from 0.900 to 0.976 showing that the Rasch item difficulty index did maintain stability of item difficulty from grade level to grade level. In addition, Rasch Achievement Scale Indices were calculated for each subtest of the CTBS and compared to the publisher's Expanded Scale Scores, which was produced through the Thurstone absolute scaling procedure. The correlations from this comparison ranged from 0.985 to 0.998, showing the Rasch procedure has comparability with the Thurstone absolute scaling procedure.

In the second phase of the study, Rasch scale parameters were estimated for the MAMBO and each of the three arithmetic subtests of the CTBS. Linking equations were then developed to equate each of these measures to one another and to the total arithmetic scale of the CTBS.

Additionally, Dr. Stiles is heading an activity sponsored by the Washington State Office of the Superintendent of Public Instruction to identify the basic skills measures (both commercial and non-commercial) currently administered in the states of Washington and Oregon along with dates and places of administration. Volunteers for additional content referenced measurement testing will be obtained and from this a plan will be developed for linking these basic skills measures by content area during the 1975-76 school year.

Summary

Based on the analysis of data cited in this paper, it appears that the Rasch model has met the criteria set for it. The items indicated as poorly fitting the model have evidenced defects which warrant revision or elimination. The linking equations between tests appear to provide consistent estimates of item performance and test scaling based on the cross-validation of results. Finally, these data are supportive of the feasibility of obtaining estimates of individual item difficulties which are consistent across testing situations.

The efforts of the Northwest Evaluation Association provide considerable optimism about the usefulness of the Rasch test analysis model in solving a variety of important school testing problems. It appears that the model can provide the necessary link between content referenced and norm referenced tests. Support is also offered for the feasibilit" of scaling each item independent of student performance, and independent of the scaling of all other items. This latter result may provide the basis for a fully flexible content referenced testing program providing the information now only available through norm referenced tests.

References

Doherty, Victor W. and Walter E. Hathaway, "Goals and Objectives in Planning and Evaluation: A Second Generation," NCME Measurement in Education, Fall 1972, Vol. 4, No. 1

Stiles, Richard L., "Estimating Achievement" a presentation made in the symposium; The Rasch: A New Approach to School Measurement; at the Second Annual Pacific Northwest Educational Research and Evaluation Conference, May 23-24, 1974, Seattle, Washington.

Wright, Benjamin D., Sample-Free Test Calibration and Person Measurement, Proceedings of the 1967 Invitational Conference on Testing Problems (Princeton: Educational Testing Service), 1968.

RLS/jh 3-75