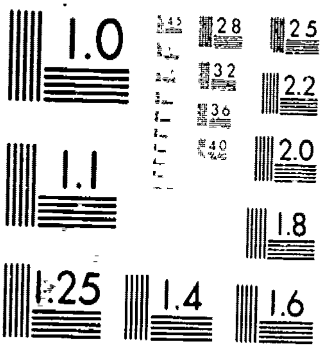

1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

DOCUMENT RESUME

ED 107 675 TM 004 481

AUTHOR Forbes, Dean W.; Ingebo, George S.

TITLE An Empirical Test of the Content Homogeneity Assumption Involved in Rasch Item Calibration.

PUB DATE [Apr 75]

NOTE 9p.; 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 Elementary Education; *Item Analysis; *Item Banks; *Mathematics; *Test Construction; *Tests

IDENTIFIERS *Rasch Item Calibration

ABSTRACT

A project was carried out to determine the degree of content homogeneity that a test item pool must have in order to accomplish successful Rasch calibration. Mathematics item pools were administered to upper elementary children. The items were analyzed under two conditions, with items organized into separate subtests and as a global mathematics test. Results indicate that satisfactory item calibration can be carried out with global mathematics tests, avoiding the necessity of organizing items into separate subtests. The degree to which calibration is identical under the two conditions is a topic for further study. (Author)

An Empirical Test of the Content Homogeneity Assumption Involved in Rasch Item Calibration[1]

by

Dr. Dean W. Forbes, Portland (Oregon) Public Schools - Area II
Dr. George S. Ingebo, Portland (Oregon) Public Schools - Area III

Introduction

Rasch item calibration has great implications for public school measurement. The possible incompatibility between item calibration and the survey assessment of broad instructional disciplines suggests potential problems which must be explored in order to determine the actual utility of item calibration in public school measurement practice.

Rasch calibration permits describing item performance in terms of that degree of capability within the discipline which is required for its successful completion rather than in a normative relationship to some specific population. This permits a degree of flexibility in educational measurement which far exceeds that which is possible by means of conventional normative procedures. In recent years it has been receiving more and more public attention. The few empirical demonstrations of Rasch calibration in public school practice have been very encouraging and suggest that the procedure has great power for educational measurement. Inherent in the process is an assumption of content homogeneity which has been at best rather sketchily defined.

This assumption of homogeneity is extremely important since it raises questions concerning the degree of precision with which curriculum domains must be described in order to permit calibration. For example, in language arts can the various skill elements be combined or must they be calibrated separately for greatest precision (i.e., must comprehension problems be calibrated separately from vocabulary)? In mathematics, to what degree can various sub-categories of mathematical skills (such as arithmetic computations, problem applications, or concepts) be combined into composites which permit calibration? Until some of these questions have been answered any effort to move into calibration in the public school fields bears with it the possibility of potential failure due, not to failure of the model itself, but to excessive heterogeneity of content.

Objectives

The objective of this study is to test the stringency of the content homogeneity assumption of Rasch calibration. Specifically, it intends to determine the degree of congruency between the calibration of individual mathematics test items when treated as: (1) members of mathematics sub-tests (e.g., computation, concepts, or problem applications); or (2) members of a global mathematics survey test.

[1] Paper presented at the American Educational Research Association Meeting, Washington, D. C., April 1975.

This assumption has been tested in one specific context, mathematics at the seventh grade level. It is part of a much larger calibration project in which a global item pool of seventh grade mathematics material (separated into four difficulty levels to be linked together) was calibrated. The item pool had gone through traditional item analysis and all "defective" items had been deleted. The surviving items were arranged in order of difficulty and were broken down into four difficulty level "trial tests", each of which was of a practical length for administration to students. Where "tests" of two difficulty levels came together, a number of items were shared between adjacent levels (between 20 and 25 items in each instance) so that the calibration of the individual levels could be linked together to form one continuous scale extending through all four levels. Up to this time there had been no break out of items in terms of sub-test structure.

Following initial calibration of the various difficulty ranges in the total item pool, a post-mortem analysis subdivided the items into the three conventional sub-test classifications, computation, problem applications, and concepts.

The sub-tests were individually calibrated at each level and the "sub-test only" calibrations were compared with calibrations of those same items when embedded in a composite containing the other sub-tests as well.

The ordering of items within the calibrated scaling was compared from sub-test to composite calibration as was the scale value assigned to the difficulty of each item (again comparing sub-test calibration values to composite calibration values). The basic question to be answered concerned itself with whether or not sub-test calibration alters scale values between and among sub-test items relative to scale values achieved by the same items as part of a composite.

All comparisons were made in terms of logit scales. Since this procedure computes a scale in terms of values at hand (as based on the specific items involved) the center of the calibration scale for any specific situation is arbitrarily set at 0.00, proceeding up and down from this point to those positive and negative limits necessary to span the operating range of the particular group of items being calibrated. For this reason, the individual calibration values for a given group of items can vary slightly from one situation to another (in this case when the items are calibrated separately as an intact sub-test as opposed to their calibration when embedded in a context involving items from other sub-tests).

In effect, if a composite had a slightly higher proportion of more difficult items than did a selected group (a sub-test) the maximum range and zero point would vary somewhat from that of the sub-test calibrated in isolation. The values for individual items calibrated in each of these two situations would vary by a constant factor if the scale did not suffer from some type of distortion. If one or the other scale were distorted such differences between calibration for individual items would vary somewhat in size from one region of the total scale to another as a function of the magnitude and nature of the distortion which was involved.

In order to translate scale values for each sub-test and level which was involved in this analysis into comparable terms the mean difference in calibrated value (of individual items) from sub-test calibration to composite calibration was computed and, in each case, the scale of values in the sub-test calibration was adjusted by this amount to bring calibration in both situations into comparable terms.

The calibration of items in each specific sub-test (computation, problem applications, and concepts) at each of the four difficulty levels (W, X, Y, and Z) was compared to the corresponding composite calibration with respect to two characteristics, the ordering or ranking of item difficulties, and the maximum range from least difficult to most difficult. In all cases the scaling of items in the isolated sub-test situation was compared to the scaling of those same items when embedded in a composite consisting of that sub-test plus the items from the other two sub-tests.

Results

In every situation (all three sub-tests and all four difficulty levels) identical ordering of all items took place in both sub-test and composite calibration. There were no changes whatsoever in rank of any item and virtually no difference in the spacing of adjacent items or in the pattern of spacing of inter-item difficulty differences along a sequence of items. (This identity of ordering of items from one situation to the other leads to rank difference correlations of +1.00 in all 12 of the situations which were examined.)

Upper and lower difficulty limits (as well as ranges) for each of the various sub-tests and difficulty levels are presented in Table 1 for both the sub-test and composite calibrations. Table 2 summarizes the ranges and gives the differences in maximum range between the two calibrations.

An examination of the tables indicates that maximum range (from the easiest to most difficult item) is practically identical in all but two situations. Range is virtually identical for all levels of problem applications and concepts as well as for the intermediate difficulty levels (X and Y) of computation. It is apparent that the difficulty ranges of the items for the easiest level (W) and most difficult level (Z) of computation increase somewhat in the sub-test calibration. This "stretch" is on the order of 1/12 of the total range (Plate 1) (for W the distortion equals 1.8 times the average difference in scale value between adjacent items in the scale; for level Z it is equal to 2.4 times the average difference between adjacent items).

Since this distortion is approximately equally distributed between the upper and lower end of the scale and is progressive from the zero point to either extreme, it means that any actual displacement of an individual item would rarely exceed one ranking place when sub-test calibration is compared to that of composite calibration. In all other sub-test/level combinations the maximum possible displacement of an item would be less than the equivalent of one ranking position.

Table 1

Composite and Sub-test (Adjusted) Difficulty Limits and Ranges for each of Four Difficulty Levels[1] of Seventh Grade Mathematics Expressed in Logits.

	Computation		Problem Applications		Concepts	
	Adjusted Sub-test	Composite	Adjusted Sub-test	Composite	Adjusted Sub-test	Composite
Level W						
Top Bottom	1.029 -2.151	0.912 -2.034	1.427 -1.053	1.437 -1.063	1.391 -0.092	1.406 -0.094
Range	3.180	2.946	2.480	2.500	1.483	1.500
Level X						
Top Bottom	0.725 -0.723	0.721 -0.696	1.099 -1.262	1.097 -1.246	0.803 -1.035	0.806 -1.037
Range	1.448	1.417	2.361	2.343	1.838	1.843
Level Y						
Top Bottom	0.436 -1.019	0.417 -0.993	0.701 -0.627	0.694 -0.614	0.957 -0.545	0.977 -0.556
Range	1.455	1.410	1.328	1.308	1.502	1.533
Level Z						
Top Bottom	0.905 -1.987	0.780 -1.855	1.822 -1.313	1.847 -1.346	1.006 -1.192	1.036 -1.238
Range	2.892	2.635	3.135	3.193	2.198	2.274

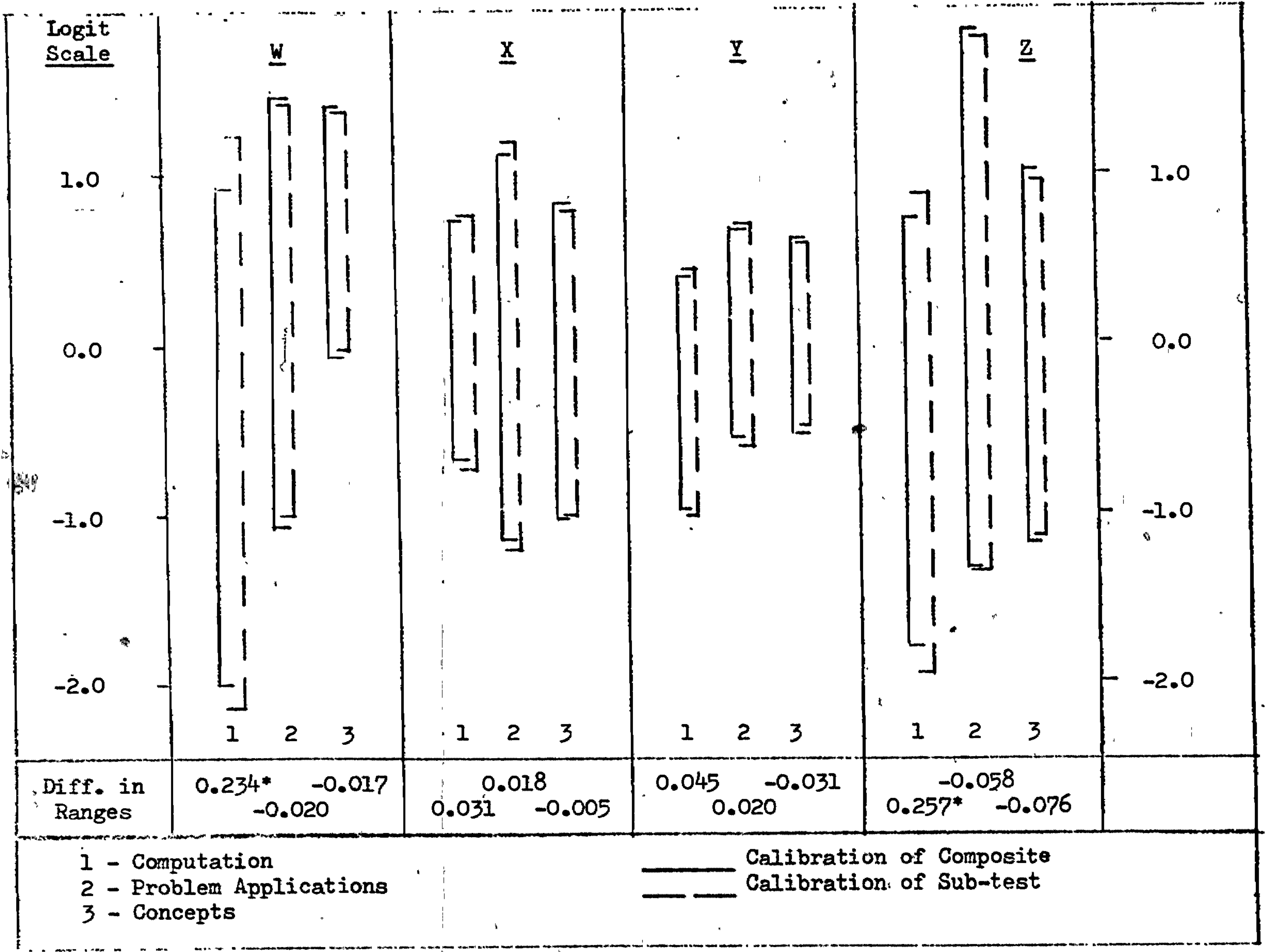
[1] The logit is defined as the metric underlying the logistic curve.

Table 2

Difficulty Ranges for Composite and Sub-test Calibration, and the Differences in Range between Composite and Sub-test Calibration for Four Levels of Seventh Grade Mathematics Expressed in Terms of Logits.

	RANGE IN LOGITS		
Sub-test	Sub-test Calibration	Composite Calibration	Difference in Range (Sub-test Compared to Composite)
	Level W (Easiest)		
Computation	3.180	2.946	+0.234*
Problem Application	2.480	2.500	-0.020
Concepts	1.483	1.500	-0.017
	Level X (Moderately Easy)		
Computation	1.448	1.417	+0.031
Problem Application	2.361	2.343	+0.018
Concepts	1.838	1.843	-0.005
	Level Y (Moderately Difficult)		
Computation	1.455	1.410	+0.045
Problem Application	1.328	1.308	+0.020
Concepts	1.502	1.533	-0.031
	Level Z (Most Difficult)		
Computation	2.892	2.635	+0.257*
Problem Application	3.135	3.193	-0.058
Concepts	2.198	2.274	-0.076

* Difference in excess of mean difference between adjacent items in scale.



* Difference in excess of mean difference between adjacent items in scale.

Plate 1 - Difficulty Level Ranges for Sub-tests at four Difficulty Levels.

Summary and Conclusions

In order to test the homogeneity of content assumption involved in Rasch item calibration, a post-mortem analysis was conducted in which the calibration of a total seventh grade mathematics item pool (comprising a combination of computation, problem applications, and concept problems) was carried out under two conditions, as separate sub-tests and as a global composite. The composite calibration of items was checked against calibration of identical items sorted out into conventional sub-tests. In all cases, item calibration values arranged themselves in identical order in both the composite and the sub-test situations. In most situations the scaling was virtually identical from the one situation to the other. For the easiest and most difficult levels of computation items there was a certain amount of distortion with the total range of scale values stretching slightly in the sub-test calibration. Even in these situations the amount of scale distortion was sufficiently small that it would present no practical problem in assembling tests utilizing the items involved in either calibration.

In view of this data it is concluded that the content homogeneity assumption involved in Rasch item calibration is sufficiently tolerant to permit calibration of general mathematics items (arithmetic) at the upper elementary grade levels in terms of a composite pool of items without the necessity of carrying out a sub-test breakdown.

Summary and Conclusions

In order to test the homogeneity of content assumption involved in Rasch item calibration, a post-mortem analysis was conducted in which the calibration of a total seventh grade mathematics item pool (comprising a combination of computation, prcblem applications, and concept problems) was carried out under two conditions, as separate sub-tests and as a global composite. The composite calibration of items was checked against calibration of identical items sorted out into conventional sub-tests. In all cases, item calibration values arranged themselves in identical order in both the composite and the sub-test situations. In most situations the scaling was virtually identical from the one situation to the other. For the easiest and most difficult levels of computation items there was a certain amount of distortion with the total range of scale values stretching slightly in the sub-test calibration. Even in these situations the amount of scale distortion was sufficiently small that it would present no practical problem in assembling tests utilizing the items involved in either calibration.

In view of this data it is concluded that the content homogeneity assumption involved in Rasch item calibration is sufficiently tolerant to permit calibration of general mathematics items (arithmetic) at the upper elementary grade levels in terms of a composite pool of items without the necessity of carrying out a sub-test breakdown.