DOCUMENT RESUME

ED 189 101	TM 800 184
AUTHOR	Forster, Fred: Inqebo, Georqe
TITLE	Rasch Model Monograph Series. Portland Public Schools Occasional Papers in Measuremen+ No. 20.
INSTITUTION	Northwest Evalua+ion Association, Oreg.: Portland Public Schools, Oreq.
PUB DATE	[78]
NOTF	25p.
EDFS PRICE	MF01/PC01 Plus Postaqe.
DESCFIPTORS	Achievement Tests: Flementary Education: Field Tesss: *Item Banks: *Latent Trait Theory: Sampling: Scalinq: *Test Ccnstruction: Test Validity
IDENTIFIERS	Interval Scales: Item Calibra+icn: Oregon (Portland): *Pasch Model: Pestricticn of Range: Sample Size: Test Linkinq

ABSTRACT

Six monographs on the Rasch model are summarized. The first qives a historical perspective on the application of the Pasch model in +he Portland, Oregon, metropclitan area. The remaining papers summarize research on the Rasch model. The research in Monograph II lead to the conclusicn that random samples are not needed to calibrate item levels in reading and mathematics. The research reported in Monoqraph III supported the contention that Rasch item levels are based on an equal interval scale. The research in Mcnograph IV concluded that students receive comparable achievement levels from different tests, provided that neither test is at a qrossly inappropriate level. Research in Monograph V determined that a sample size of at least 200-300 students should be used to field test new items. Moncgraph VI dealt with the item calibraticn problems that might result if the range of the test were restricted, and concluded that the item calibration would nct be significantly affected. (Authcr/BW)

PORTLAND PUBLIC SCHOOLS

OCCASIONAL PAPERS IN MEASUREMENT

OCCASIONAL PAPER NO. 20

RASCH MODEL MONOGRAPH SERIES

Fred Forster
George Ingebo
Portland Public Schools

© NORTHWEST EVALUATION ASSOCIATION
631 N.E. Clackamas Street
Portland, Oregon 97208

HISTORICAL PERSPECTIVE ON THE APPLICATION OF THE RASCH MODEL IN THE PORTLAND METROPOLITAN AREA

The work described in this monograph series grew from 20 years of cooperative projects in Oregon and Washington.

The first step in a cooperative, area-wide testing program began in fall, 1957, when Victor Doherty and George Ingebo developed local norms for a high school testing program in the Portland Public Schools. These local norms were reported as standard scores (a mean of 50 and standard deviation of 10)--a significant advance over grade equivalent and percentile scores. The following year, through the leadership of Bernice Tucker, this local norming program was adopted by the Multnomah County Intermediate Education District and made available to districts throughout the country. In 1960, the three counties in the Portland area established the Metropolitan Area Test Program Board, to make the new program available and provide cooperative planning for new testing programs in Clackamas, Multnomah, and Washington counties. In addition to increasing efficiency and reducing costs of current programs, the foundation of MATPB made it possible to establish comprehensive up-to-date Northwest norms considerably more relevant than those provided by test publishers.

At this time, the MATPB program was extended to include reading and mathematics in the elementary grades, three through eight. At these levels, tests were developed, field tested, revised, formated and printed by MATPB committees, providing MATPB control over the content of tests as well as the norming base. An important innovation in these tests was the even distribution of items at the low and high ends which provided valid measurement for the most-and least-able students, as opposed to publishers' tests, which are heavily weighted toward items of average difficulty. At this time, MATPB was so successful that it involved districts comprising two-thirds of Oregon's students.

By 1966, the desire for more flexible testing led to the Computer Based Testing Project (COMBAT) with Teaching Research as the supervising agency. As part of this project, teachers wrote thousands of test questions which were entered into a computer file. Each question was tagged by a behavioral objective so that teachers could phone in one or more key words and receive a ditto master of all relevant items in one or two days. Although COMBAT was short lived, it provided valuable experience for future item bank efforts. It demonstrated the need for careful validation, through field testing, of the items entered into the bank as well as the need for an organized structure to describe the content covered by each item.

In 1970, Dr. Victor Doherty recognized the need to develop a comprehensive system for detailing content in school subjects. This led to the formation of the Tri-County Course Goals Project. Since beginning work in 1971, this group has published collections of course and program goals in twelve subjects -- language arts, mathematics, science, social studies, art, music, physical education, health, industrial education, business education, home economics and second language -- covering grade levels K-12.

In the late-Sixties, the desire to establish a more flexible base for test development inspired interest in the Rasch model. For example, provisions of the-then new Title I program stipulated testing the least able students (most functioning two or more years behind normal) and yet out-of-level testing as promoted by several test publishers produced unsatisfactory results. Motivated by an article written by Benjamin Wright, Peter Wolmut and James Beaird attended a conference organized by Dr. Wright to introduce the Rasch model. Based on their reports, and the results of initial data analysis, the Rasch model appeared to provide the flexibility that was needed.

In 1973, a meeting of task forces in Washington and Oregon led to the foundation of the Northwest Evaluation Association (NWEA) whose mission was to develop goal referenced item banks in all school subjects. Simultaneously, Dr's Ingebo, Forster and Forbes, began carefully researching the important properties of the Rasch model. With the help of the Office of the Superintendent of Public Instruction in Washington and the Northwest Regional Educational Laboratory in Portland, NWEA sponsored Rasch conferences in February and September, 1974. At the same time, the NWEA reviewed available sources of items and began field testing to develop its initial item banks in reading and mathematics. Not unexpectedly, the complications involved in field testing coupled with those of applying and expanding the Rasch methodology were sufficient to delay publishing the initial versions of the mathematics and reading item banks until Spring 1977. In the interim, the foundation of using the banks was laid in several NWEA workshops covering the techniques for applying the Rasch, the characteristics of the NWEA Rasch scales and the characteristics to be built into the item banks.

The Northwest Evaluation Association has made Rasch calibrated item banks available in reading and mathematics, and language arts.. Work on the development of item banks in other subjects has already begun, and will be an important activity of the Association for the next several years.

CAN RASCH ITEM LEVELS BE DETERMINED WITHOUT RANDOM SAMPLING?

Background

Historically, the difficulty of a test question has been tied to the performance of a specific group on that question. For example, we might determine an item has a difficulty p-value of 75% (correct) when taken by a particular group of students. The problem with this approach is that we must specifiy a "comparison" group (such as suburban fifth graders) and are never quite sure how the item might work with a different group. For this reason, those concerned with developing good tests have attempted to use random samples to insure the p-value for a test item is representative of the comparison population. Unfortunately, this course is fraught with difficulty. If students are sampled randomly, they must be pulled away from their regular classwork and school routine is likely to be disrupted. If volunteers are used in the sample, the question arises as to what might have happened if the holdouts had been included. On the other hand, if teachers and principals are forced to participate, there is no way to judge how carefully they may have followed the proper standardized procedures or prepared students for the tests. In short, random sampling (or almost any sampling), has serious limitations.

For this reason, we wanted to explore the Rasch test model which supposedly eliminates the need for random sampling. The Rasch approach makes this possible by using p-values and student scores to determine the "true" difficulty level of each item and the "true" achievement level of each student on an underlying curriculum scale. Going beyond this, Ben Wright (the father of the Rasch model in the U.S.) made what seemed to be a preposterous claim; that an item would be calibrated to the same level regardless of the students used in field testing. Our doubts about this claim launched the research described in this monograph.

The Research Question: Can the Rasch Level on an Item be Determined Without Random Sampling?

The analysis was based on student responses to standardized tests in fourth grade reading (approximately 1400 students), We divided the student responses in two ways: (1) students scoring above the average score vs. students scoring below the average, and (2) students from inner-city schools vs. students from the more advantaged schools. Then we calculated the Rasch item difficulty levels separately for each of the groupings, and for the total group of 1400 students. Finally, we correlated the Rasch item levels as calibrated for the different groups against the total group to determine the degree of match.

Results

The results of our research are shown in Figure 1. In addition to the correlation between item levels (which insured they are in the same order), it is also important to examine the ratio of the standard deviations (which insures they are of the same magnitude). These two values can be combined to form a "restricted" correlation; i.e., one in which the pairs of item levels are required to have identical values as well as to be in the same order. As can be seen from the figures, the results indicate that the item levels agreed quite well, even in the case of the extreme split of students above and below the average.

Replication

This research was repeated on approximately 4000 students at fourth grade in reading and mathematics and 4000 students at eighth grade in reading and mathematics. As shown in Figure 2, the replication confirmed the results of the original research.

Conclusions

Based on this research, it appears that random samples are not needed to calibrate item levels in reading and mathematics.

Figure 1

Comparison of Rasch Item Levels Calibrated on Student Subgroups

	FOURTH READING	
COMPARISON	CORRELATION	RESTRICTED CORRELATION
Girls above the average vs. Total	.98	.74
Girls below the average vs. Total	.96	.89
Boys above the average vs. Total	.97	.75
Boys below the average vs. Total	.95	.80
Inner-City Schools vs. Total	.99	.96
Top Schools vs. Total	.93	.89
Inner-City Schools vs. Top Schools	.97	.90

*Based on approximately 1400 students

Figure 2

Correlations and Restricted Correlations* of Item Levels for Subgroups and the Total Group#

	SUBJECT							
	FOURTH READING		FOURTH MATH		EIGHTH READING		EIGHTH MATH	
GROUP	r	Restricted r	r	Restricted r	r	Restricted r	r	Restricted r
Above the Average vs. Total	.99	.83	.98	.88	.98	.75	.97	.80
Below the Average vs. Total	.96	.81	.97	.89	.98	.81	.95	.85
Inner-City Schools vs. Total	1.00	.96	.99	.98	1.00	.96	1.00	.99
Top Schools vs. Total	1.00	.95	.99	.97	1.00	.97	1.00	.98
Inner-City Schools vs. Top Schools	.99	.92	.98	.95	.99	.92	.98	.96

*Restricted correlations between Rasch Item Levels requiring that the values be equal in magnitude.

#Based on approximately 4000 students for each grade and subject.

IS THE RASCH ITEM LEVEL SCALE EQUAL INTERVAL?

Introduction

One of the claims made for the Rasch model is that its items are measured on an equal interval scale. This is important since making comparisons among traditional measures of item difficulty is a shaky affair. The p-value (percent correct), for example, varies significantly for different student groups. A test item may have a p-value of 85% for a high level fourth grade group and 45% for an inner-city fifth grade group. Which of these is the "correct" difficulty level?

After some consideration, the reason for this confussion is clear: P-values are tied to a specific group rather than the underlying curriculum. For this reason Georg Rasch proposed his model to transform p-values into an equal interval item level based on the underlying learning scale.

The Research Question: Is the Rasch Item Level Scale Equal Interval?

Background

When the Rasch item levels are calculated for a test, they are centered on the average level for that test. Each calibration of an item is the same except for the correction needed to reflect the average level for the test. For example, if one test has an average level ten points above that for another, then an item calibrated at the 180 level on the first test would be expected to be calibrated at the 190 level on the second test.

By including several of the same items on both of the two tests, the average difference of the calibrations provides an unbiased estimate of the difference in the average level of the two tests. The equal interval nature of the Rasch scale can be checked by the consistency of the linking values among several tests.

Method

To test this research question, a pool of 250 previously field tested seventh grade mathematic items were formated in seven tests. The items were divided into two groups: an easier group which had p-values over 20% and a difficult group which had p-values of 20% or lower. The easier group of items was formated into four sixty-item tests of graduated difficulty (forms W, X, Y, Z). These tests were constructed so that twenty items were shared between adjacent levels. The difficult group of items were formated into three tests (forms D, E, F). These tests were linked to each other as well as to forms W and Z (see Figure 1).

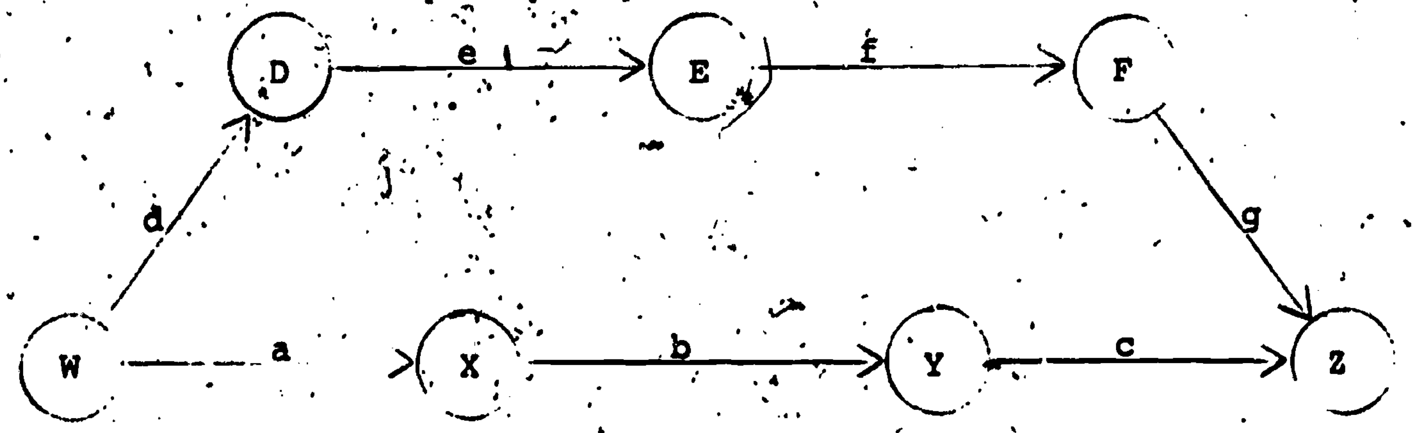
The linking values between the tests are summarized in Figure 2. Note, that a different block of common items was used between each pair of forms. The linking value was calculated by subtracting the average calibration for a block of items on the first test and the second test. Figure 3 shows the composite values for the links from W to Z through two independent pathways (W-X-Y-Z and W-D-E-F-Z). The two pathways led to a discrepancy of 1.1 Rits or 4.4% of the total linking value.

Conclusions

These findings support the contention that Rasch item levels are equal interval.

Figure 1

Linking Network Among Tests



Capital letters designate test forms.
Lower case letters designate linking items.

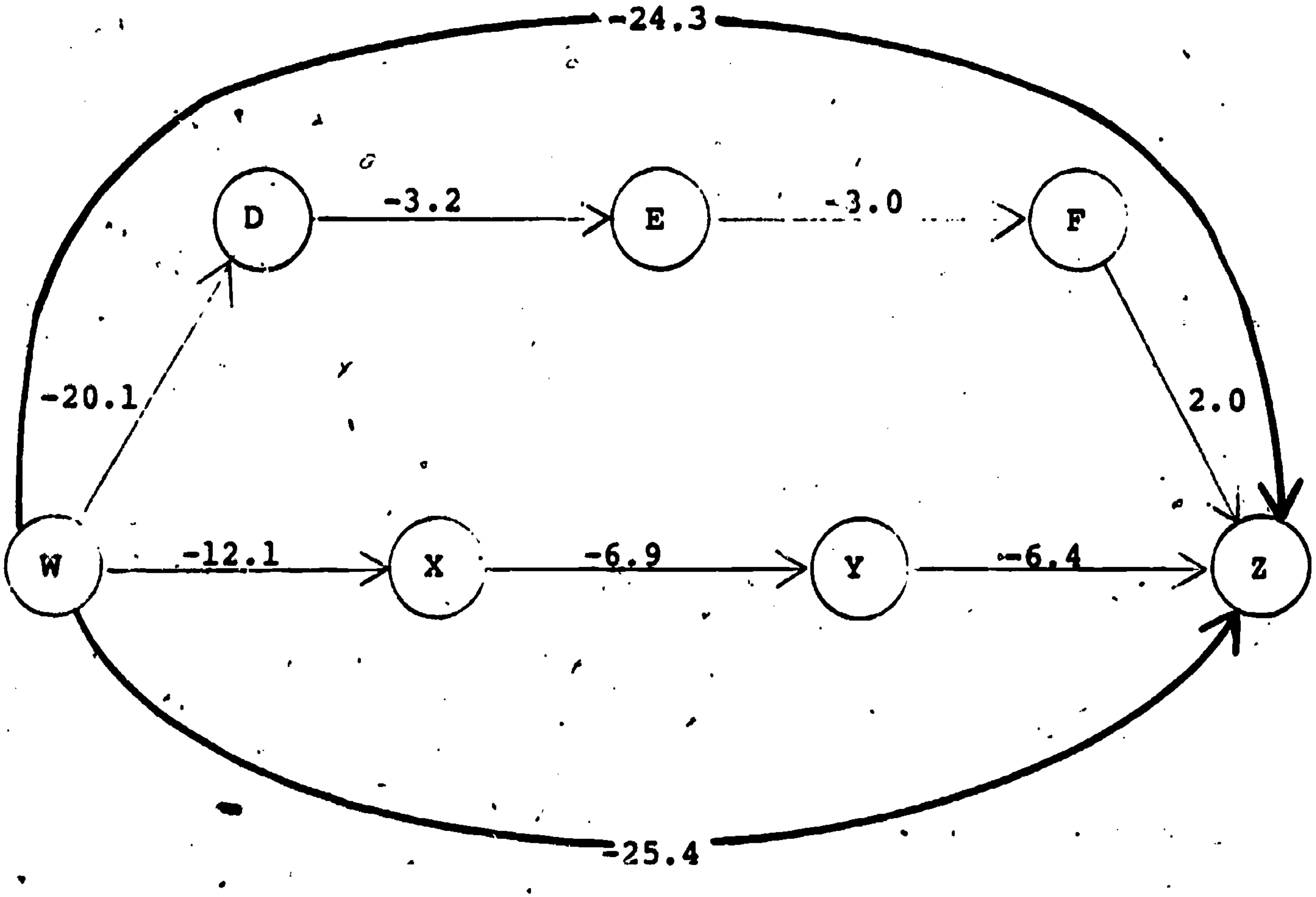
Figure 2

Linking Values for the Network

First Form	Second Form	Linking Block	Average First Form	Average Second Form	Linking Constant
W	X	a	207.5	195.4	-12.1
X	Y	b	204.5	197.6	- 6.9
Y	Z	c	202.5	196.1	- 6.4
W	D	d	207.0	186.9	-20.1
D	E	e	206.9	203.7	- 3.2
E	F	f	203.4	200.4	- 3.0
F	Z	g	194.0	196.0	2.0

Figure 3

Linking Values Shown on Network



DO STUDENTS RECEIVE THE SAME RASCH ACHIEVEMENT LEVEL FROM DIFFERENT TESTS?

Background

Traditional methods for equating tests are complicated and unreliable for two reasons. First, there are too many variables to handle (differences in test length, test level, and test range). Second, they are based on sample dependent statistics which shift dramatically for different groups.

The Rasch model, on the other hand, circumvents these problems, by scaling each test on the same underlying curriculum scale of student achievement. The purpose of this this research study was to verify this characteristic of the Rasch model.

The Research Question:	Do Students Receive the Same Achievement Level from Different Tests?

Method

In fall 1977, two reading tests were given to students in grades three through eight. The first was a thirty-item field test, and the second was an 80 to 100-item standardized survey test in reading. Since two different field tests were used at all grade levels except three and five, there were ten independent comparisons in the study.

Following field testing, the field tests and the survey tests were linked to the Northwest Evaluation Assoication Rasch Reading Scale. After eliminating low quality items from the analysis, each student's achievement was scaled separately for his performance on the field test and on the survey test. The resulting pairs of raw scores and achievement levels for each student were then averaged to make the required comparison.

Results

As shown in Figure 1, the Rasch averages agree quite closely for nearly all the test pairings. In those instances where the difference was more than one Rasch unit, it was found that the field test was poorly matched to the level of the students in the sample and "ceiling" or "floor" estimates were introduced inadvertently.

Replication

This study was replicated on a more extensive basis in fall 1977. Each student in grades four through eight took two short achievment tests in reading and in mathematics. The first test, the competency progress test (CP), was administered on a grade level basis, while the second test, the achievement level test (AL), was assigned to the student based on previous measures of his achievement. The tests were scaled independently and a pair of raw scores and achievement levels were calculated in each subject for each student as in the original study. As shown in Figure 2, the results of the replication confirmed the findings from the previous study. (It should be noted that this comparison is based on a single form of the CP test, but up to ten different AL forms at each grade level.) In those cases where the averages for the two tests differ by more than a point, it was found that the raw score distribution of the CP test was severly truncated at either the top or bottom end indicating the inappropriateness of that test for much of the student sample.

Conclusions

These results lead us to conclude that students do receive comparable achievement levels from different tests, provided that neither test is at a grossly inappropriate level.

Figure 1

Comparability of Rasch Achievement Levels

Grade	Field Test Form	AVERAGE RAW SCORE		AVERAGE RASCH ACHIEVEMENT LEVELS	
		Field	Standardized Test	Field Test	Standardized Test
3	R12	18.1	16.2	186.8	187.2
4	R17	19.6	33.5	194.9	194.6
4	R35	19.6	36.1	195.8	196.8
5	R36	15.6	31.7	199.3	199.1
6	R25	14.4	38.6	204.4	205.3
6	R37	11.0	38.4	206.2	205.3
7	R27	14.1	41.0	213.1	211.4
7	R38	12.9	35.7	206.6	207.4
8	R32	12.7	61.8	211.7	211.7
8	R39	15.0	57.5	211.2	209.4

*Each comparison is based on a sample of approximately 250 students at the indicated grade level.

Figure 2

Comparison of Results Between the Competency Progress and Achievement Level Test Programs#

READING			MATHEMATICS		
Grade	CP* Average	AL** Average	Grade	CP Average	AL Average
4	194.12	193.76	4	191.28	190.36
5	200.00	200.00	5	200.81	200.00
6	205.53	205.74	6	207.22	208.36
7	209.77	211.21	7	213.49	214.87
8	214.74	216.83	8	222.15	221.41

*Competency Progress

**Achievement Level

#Based on the analysis of test scores for approximately 4000 students at each grade level.

WHAT IS THE SMALLEST SAMPLE SIZE NEEDED FOR FIELD TESTING?

Background

To field test the many items needed in building a comprehensive item bank, it is important to take full advantage of the limited number of participating students and teachers by using the smallest sample which will yield reliable item level calibrations. This research study was intended to address this practical issue.

The Research Question: What is the Smallest Sample Size Needed for Field Testing?

Method

In the fall of 1976, approximately 1400 students responded to a fourth grade mathematics test and the same number responded to an eighth grade reading test. A computer program was developed to randomly select five samples each of sizes 50, 100, 200, and 300. The calibrations for the five samples of a given size were then correlated with those for the total group of 1400 students. Figure 1 shows these correlations together with the ratio of the standard deviations of the calibrations (which should be 1 if the metrics are equal), and the average discrepancy which is the average of the absolute value differences between the calibrations for the samples of a given size and the total group.

Results

Based on the data in Figure 1, we concluded that a sample size of 200 provided nearly as accurate information as 300, and yet was significantly more accurate than lower sample sizes.

Replication

Using the responses of approximately 3800 students at fourth grade and seventh grade to a reading and a mathematics test, five random samples were drawn of each of the sizes 50, 100, 150, 200, 250, and 300. The results of the correlations between the calibrations for the five samples of a given size and the total group are shown in Figure 2. Based on these data we again concluded that a sample size of 200 appears to maximize the information available from a limited field test population.

Conclusions

Based on the research, we use 200-300 students in field testing new items for the NWEA item banks.

Figure 1

Fourth Grade Mathematics and Eighth Grade Reading 1974-1975

	FOURTH MATHEMATICS			EIGHTH READING		
Number of Items	81			100		
Total Population	1478			1808		
Standard Deviation	14.15			15.25		
Sample Size*	Correlation	$ Ratio	Avg.# Disc.	Correlation	$ Ratio	Avg.# Disc.
50	.956	1.102	3.108	.965	1.034	3.042
100	.978	1.029	1.950	.982	1.019	2.074
200	.987	1.107	1.314	,989	,994	1,381
300	,989	1,010	1.138	,991	,996	1,103

*Entries represent five samples drawn of each size.

#The average absolute value difference in the calibrations.

$The ratio of the standard deviations of the two sets of calibrations

<u>Figure 2</u>

Fourth and Seventh Grade Reading and Mathematics 1977

	FOURTH READING			FOURTH MATH			SEVENTH READING			SEVENTH MATH		
Number of Items	35			37			36			30		
Total Population	3771			3827			3846			3773		
Standard Deviation	13.51			11.84			6.17			8.30		
Sample Size*	Corr.	Ratio	Avg.# Disc.	Corr.	Ratio	Avg.# Disc.	Corr.	Ratio	Avg.# Disc.	Corr.	Ratio	Avg.# Disc.
50	.956	1.045	3.367	.950	1.100	3.291	.939	1.099	2.689	.901	1.205	2.531
100	.977	1.019	2.430	.975	1.024	2.201	.961	1.104	2.086	.941	1.096	1.884
150	.986	.990	1.822	.985	1.007	1.577	.978	1.020	1.463	.955	1.087	1.639
200	.990	.992	1.623	.990	1.021	1.388	.983	1.014	1.299	.971	1.038	1.244
250	.992	1.006	1.353	.992	1.020	1.173	.982	1.054	1.407	.972	1.008	1.219
300	.994	.993	1.148	.994	1.022	1.113	.986	1.045	1.170	.978	.999	1.067

*Entries represent five samples drawn of each size.

#The average absolute value difference in the calibrations.

IS THE CALIBRATION OF AN ITEM AFFECTED BY THE RANGE OF ITEMS ON A TEST?

Background

Every item is calibrated in the context of a test. This led to the question of how the calibration for an item might shift if the range of the test were restricted. Thr purpose of this study was to answer that question.

The Research Question:	Is the Calibration of an Item Affected by the Range of Items on a Test?

Method

Approximately 1400 student responses to an eighty-item fourth grade standardized survey test were used in this study. First, the total test was calibrated. Then, as shown in Figure 1, the ten highest level and ten lowest level items were dropped to yield a sixty-item subtest. Next, the five highest and five lowest items were successively dropped to yield subtests of 50, 40, 30, 20, and 10 items of decreasing range of item levels. By this procedure, the items were calibrated on several different subtests, the ten middle level items being calibrated on all the subtests.

The calibrations for the items were then correlated across subtests to identify any shifts or inconsistencies. The correlations, which were essentially 1.00, were then further restricted to require the equality of metrics between the subtests. The results of this analysis are shown in Figure 2.

Conclusions

Based on these data, it appears that the calibration of an item is not significantly affected by the range of item levels on a test.

Figure 1

Comparison of Item Levels for Subtests of Varying Ranges and Difficulty Drawn from a Fourth Grade Mathematics Test

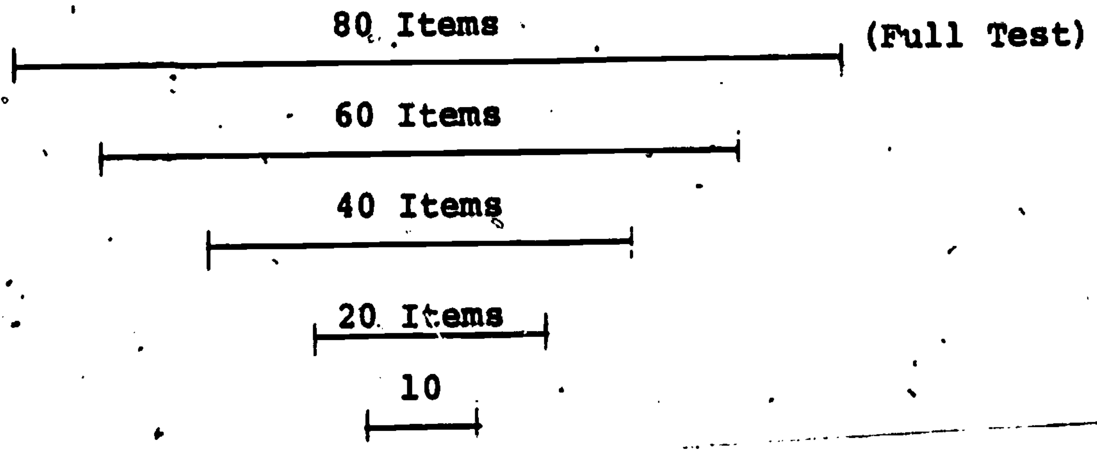


Figure 2

Correlations Among Subtests*

Subtest	Full Test	60	50	40	30	20	10
60	.998						
50	.996	.999+					
40	.995	.999	.999+				
30	.993	.999	.999+	.999+			
20	.989	.997	.999	.999	.999+		
10	.991	.998	.999	.999+	.999+	.999+	

*Restricted to require equal metrics as well as equal orderings.