DOCUMENT RESUME

ED 191 861 TM 800 467

AUTHOR Canner, Jane M.: Lenke, Joanne M.
TITLE Some Types of Test Items Do Not Fit the Rasch Model: Examples and Hypotheses.
PUB DATE Apr 80
NOTE 20p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests: Elementary Secondary Education: *Goodness of Fit: *Item Analysis: *Latent Trait Theory: Mathematical Models: Mathematics: Phoneme Grapheme Correspondence: Reading Comprehension: Scaling: Spelling: *Test Construction: Test Items: Test Validity: Two Year Colleges
IDENTIFIERS *Rasch Model: Stanford Achievement Tests: Stanford Diagnostic Mathematics Test: Stanford Diagnostic Reading Test: Stanford Early School Achievement Test: Stanford Test of Academic Skills

ABSTRACT

Item response data were obtained from large samples of students in Grades K-community college, taking the following tests: Stanford Early School Achievement Test, Stanford Achievement Test, Stanford Test of Academic Skills, Stanford Diagnostic Reading Test, and Stanford Diagnostic Mathematical Test. Data were classified as fitting or non-fitting the Rasch model, according to mean square fit or adjusted mean square fit statistics. Non-fitting items were examined for consistencies in item content or format. Results indicated the following: (1) high percentages of spelling, reading, and mathematics items in all tests analyzed fit the Rasch Model: (2) "prior notions of likely fit" do not include specific types of item content or format: (3) items measuring knowledge of specific content may not fit if the item content is not always taught or does not follow a regular pattern of instruction at particular grades and times of year. (RL)

Some Types of Test Items Do Not Fit the Rasch Model: Examples and Hypotheses

Jane M. Canner
Joanne M. Lenke
The Psychological Corporation

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, April, 1980.

The Rasch Model has been used in a variety of situations for item analysis/equating purposes. The model has been shown to be appropriate for such different types of tests as achievement tests in reading, mathematics, and other content areas; diagnostic tests of school-related content areas; criterion-referenced reading tests; intelligence or aptitude tests; and writing tests (Rentz and Rentz, 1978). Since the appeal of the Rasch Model is largely due to its characteristics of "item-free" person measurement and "person-free" item measurement, and to its efficient means of handling such major test development phases as equating and scaling, it appears that the Rasch Model will be used more and more as a test development and test analysis model.

The use of the model for equating purposes requires that the items used for equating "fit" the model. Although the model appears to be robust enough to tolerate some degree of departure from its assumptions (Rentz and Ridenour, 1978), it would be helpful to know beforehand which types of items best fit the assumptions of the model. We know about the item characteristics that generally cause a lack of fit to the model. Items with extreme discrimination values generally appear as non-fitting items, for example. When item discrimination values are known, this information can help the test developer choose appropriate items for equating purposes. Items that for one reason or another lead to guessing on the part of examinees also tend not to fit the Rasch Model. This will often be the case for very difficult items on an achievement test that measure concepts that have not yet been taught to the examinees. Items that are confusing, ambiguous, or have more than one correct answer also tend not to fit the model.

Achievement test items may tend not to fit the model if instruction in certain areas has not been continuous and/or if the sample of examinees analyzed has not been exposed to particular instruction. Achievement tests are built from an analysis of textbooks and curriculum guides and assume a certain continuity and progression in instruction. Since this continuity may not always reflect actual classroom instruction, items measuring content that is not consistently taught may show up as non-fitting.

Another major reason that some items show up as non-fitting is that they measure a different skill or content area than the rest of the items in the test. A major assumption of the Rasch Model is that the items being analyzed measure a unidimensional trait. According to Rentz and Rentz (1978), "There are no separate adequate tests of the unidimensionality assumption which are really adequate. . . There is no clear definition of unidimensionality when you go beyond the mathematical definition." This does not mean, however, that test developers have no criteria to review in order to evaluate the unidimensionality of a set of items. Rather, Rentz and Rentz (1978) recommend that the tighter the definition of content, the easier the items, and the more care taken in writing the items, the better the chances are of meeting the unidimensionality assumption. Rentz and Retnz go on to say that "prior notions of likely fit would contribute to efficiency in using Rasch Model methodology."

Although test developers using Rasch Model methods may have to include "non-fitting" items or items in a set that appear to measure more than one trait due to considerations of curriculum coverage, it would be useful for test developers to know in advance which types of items may not be ideal for equating purposes.

In this study, non-fitting items from a variety of sources were reviewed and analyzed. Non-fitting items within tests and across test forms and levels were examined to see if any consistencies in item content or format were apparent. Could any generalizations be made about types of test items that were likely not to fit the Rasch Model?

METHOD

Items from all levels, forms, and subtests of the following tests were analyzed by using Wright's Mesamax program that generates various Rasch statistics:

Stanford Early School Achievement Test (SESAT), 1969 edition
Stanford Achievement Test (SAT), 1973 edition
Stanford Test of Academic Skills (TASK), 1973 edition
Stanford Diagnostic Reading Test (SDRT), 1976 edition
Stanford Diagnostic Mathematics Test (SDMT), 1976 edition

SESAT, SAT, and TASK are achievement tests that cover a grade range of Kindergarten through twelfth grade in a variety of content areas; SDRT and SDMT are diagnostic tests that cover a range of first grade through community college.

Item response data from large samples of students taking these tests were used. In general, these data were taken from standardization or other large-scale research programs.

The mean square fit (MSF) statistic, a χ^2 fit statistic, was used to identify non-fitting items. This statistic is arrived at by determining the expected proportion of examinees at each ability level who should correctly answer an item according to the model and comparing that with actual proportions. The MSFs for sample sizes over 1500 were then adjusted (Adjusted MSF = MSF x $\frac{1500}{N}$) because this statistic tends to inflate as sample

size increases (Rentz and Ridenour, 1978). In addition, this adjustment facilitates comparisons over different samples and analyses. Items with adjusted MSFs (AMSF) greater than two were then classified as non-fitting for the purposes of this study.

The items in these tests were often analyzed in several ways. For example, Reading Comprehension items were analyzed as part of an individual subtest, Reading Comprhension, and as part of a larger total, a Reading aggregate. To facilitate comparisons, most analyses described here are subtest analyses. (An exception is the SAT Mathematics tests. Since items in the three SAT Mathematics subtests were only analyzed as part of a larger aggregate, Total Mathematics, this aggregate analysis will be reported here.

Non-fitting items were then reviewed as a group and in relation to fitting items for possible consistencies in item format, content, and/or skill being tested. Although all subtests were analyzed in this way, only subtests where clear patterns emerged within tests or across test forms and/or levels are presented here.

RESULTS AND DISCUSSION

In general, high percentages of items in all tests analyzed fit the Rasch Model, using a MSF or AMSF less than two as a criterion for fit (see Tables 1 through 6).

In most cases, individual items appeared to be non-fitting for individual reasons. For example, a Reading Comprehension item appeared not to fit because it required some math computation to arrive at the answer; a Listening Comprehension item appeared not to fit because correctly answering this item seemed to depend more on looking at the accompanying picture than on listening to the passage that was dictated. In cases such as these, the

non-fitting items rather clearly stood out as being different in some distinct way from most of the fitting items.

However, in some cases, specific item types tended not to fit, regardless of test form, level, or type of test (i.e., achievement or diagnostic). The consistencies found in several major content areas will be presented.

Spelling

The fit of spelling items to the Rasch Model was generally very good. The analysis of the Spelling subtest of SAT showed percentages of fitting items ranging from 83% to 100% for Primary III through Advanced levels, Forms A and B (see Tables 2 and 3). The Primary II level of the test, however, has a different format for Spelling than the Primary III through Advanced levels, and, at this level, the percentage of fitting items was lower. On Primary II Form A, 67.4% of the items fit, and on Form B, 74.4% of the items fit.

A closer examination of the format showed the following. Primary II Spelling items appear in this format:

	R	W	DK
lat	o	o	o

Students must identify whether the given word is spelled right or wrong, or choose DK if they don't know.

Primary III through Advanced Spelling items appear in this format:

limit	frightin
generation	comment

Students must choose the one incorrectly spelled word from four different words.

At the Primary II level, it appear that two distinct skills may be being measured, depending on whether the stimulus word is spelled correctly or incorrectly. Eighty-nine percent of the non-fitting items were words presented

as incorrect spellings; the fitting items tended to be those that were correctly spelled. One might hypothesize that a skill closer to word recognition was being measured by the correctly spelled words, while the incorrectly spelled words at this level require a skill that more closely approximates that skill required by the upper levels of the test. Perhaps spelling skills are not taught at this level to the extent that word recognition reading skills are taught.

A χ^2 analysis of the fitting and non-fitting items at this level shows clearly that the relationship between item type and fit to the model was a strong one.

Primary II Spelling, Form A

	Fit	Non-Fit
Correctly spelled words	19	1
Incorrectly spelled words	9	14

$\chi^2 = 14.8$ ($p < .001$)

Primary II Spelling, Form B

	Fit	Non-Fit
Correctly spelled words	18	2
Incorrectly spelled words	13	10

$\chi^2 = 6.08$ ($p < .05$)

Reading Comprehension

The fit of reading items to the model was also generally good. The analysis of the Reading Comprehension subtest of SAT showed percentages of fitting items ranging from 81.4% to 97.3% for Primary I through Advanced, Forms A and B, 97% to 99% for TASK I and II Forms A and B, and 68.7% to 91.7% for SDRT, Red through Blue levels, Forms A and B (see Tables 2 through 5).

Of the relatively few non-fitting items, more appear to measure inferential comprehension than literal comprehension. (Items measuring global, implicit, contextual, and inferential meanings according to published item objectives, are classified as inferential items for the purposes of this study; items measuring explicit or literal meanings are classified here as literal items.) This is probably due to the fact that inferential items invite more guessing than literal items and that inferential items may have more of a tendency to be ambiguous than literal items.

Although χ^2 analyses showed no significant relationships, in every case but one (SDRT Blue, Form A) the percentage of fitting literal comprehension items is greater than the percentage of fitting inferential items. (see Table 7).

Mathematics

The fit of Mathematics items to the Rasch Model was very high for SAT, all levels, Forms A and B. Percentages of fitting items ranged from 93.8% to 100% for SAT Primary I through Advanced levels, Forms A and B (see Tables 2 and 3). SESAT I Mathematics had 71.4% fitting items, and SESAT II Mathematics had 83.6% fitting items (see Table 1). TASK I and II, Forms A and B, had percentages of items that fit ranging from 79% to 94% (see Table 4). SDMT had percentages of fitting items ranging from 85% to 100% for all levels and forms (see Table 6).

Consistencies among non-fitting items were not easy to find. On the SESAT I Mathematics test, three of the four non-fitting items were dictated word problems that required computation. Fitting items tested a variety of math concepts but only one required computation. This is a fairly clear example of items that don't fit because they measure a different skill or because

they measure something that hasn't been taught yet.

One consistency noted on several levels and forms of the tests studied involved items requiring knowledge of the metric system. Each of forms A and B of SDMT levels Green, Brown and Blue contain three items that require knowledge of the metric system. Although the total number of metric items is small, 61% (11) of the 18 metric items did not fit the model. This can be compared with the generally high percentages of fitting SDMT items overall (see Table 6). On TASK, the one item per form and level requiring knowledge of the metric system did not fit the model, although high percentages of all mathematics items do fit the model at this level (see Table 4). Metric items were generally not tested on SAT Mathematics tests.

This finding is again likely due to the fact that at the time these item response data were collected (early to mid 1970's) the metric system was not systematically taught and the sample tested had not been uniformly exposed to instruction in this area. It would be interesting to see if recent item response data on metric items still shows this pattern.

Letters and Sounds

On this reading subtest of SESAT I, 75% of the items fit the model (see Table 1). On SESAT II, 78% of the items fit (see Table 1).

Several consistencies in item content for non-fitting items were noted. For example, items testing recognition of the letters "p" and "d" did not fit when either level of the test was analyzed. In addition, items testing the sound of the letter "h" did not fit the model for either analysis. On SESAT II, a χ^2 analysis showed there to be a significant relationship between fit and type of initial sound tested (blend/initial letter). Since blends are most often taught after initial letters, test items measuring blends may lead to

more guessing on the part of examinees.

SESAT II, Sounds and Letters

	Fit	Non-Fit
Blends	4	6
Non-blends	15	3

$\chi^2 = 6.54$ ($p < .05$)

This study indicates that, in general, "prior notions of likely fit" do not include specific types of item content and/or format; varied types of content and item types do fit the model well.

However, the study does reinforce the idea that items measuring knowledge of specific content may not fit the Rasch Model if the item content is not always taught (e.g., metric items) or does not follow a regular pattern of instruction at particular grades and times of year (e.g., spelling skills at second grade level, sounds of blends at first grade level).

An analysis such as this can offer some insight into the skills being tested at various levels (e.g. word reading vs. spelling skills at lower grade levels) and into the differences between various types of items (e.g. literal and inferential comprehension items). It seems, however, that individual test developers would do best to use their own judgment as to what types of items should be tested together. If the items are similar to the vast majority of items analyzed here, they will fit the model regardless of specific item content or format.

Table 1. Percentage of SESAT I and II items that fit the Rasch Model

Subtest	SESAT I	SESAT II
Environment	78.6	79.5
Mathematics	71.4	83.6
Aural Comprehension	75.0	77.8
Letters and Sounds	78.6	78.0
Word Reading	--	82.8
Sentence Reading	--	64.1

Table 2. Percentage of SAT 73 Form A items that fit the Rasch Model

Subtest	Primary I	Primary II	Primary III	Int. I	Int. II	Advanced
Vocabulary	78.4	81.1	88.9	94.0	96.0	90.0
Read. Comp.	93.1	91.4	92.9	84.7	90.1	91.9
Word Study Skills	85.0	90.8	89.1	98.2	96.0	--
Total Math	93.8	97.0	97.9	98.2	99.2	99.2
Spelling	--	67.4	91.5	100.0	96.7	93.3
Language	--	--	89.1	98.7	98.8	93.7
Soc. Sci.	--	92.6	90.9	96.7	94.4	98.3
Science	--	88.9	95.2	91.7	95.0	98.3
Listening	80.8	88.0	94.0	98.0	96.0	--

Table 3. Percentage of SAT '73 Form B items that fit the Rasch Model

Subtest	Primary I	Primary II	Primary III	Int. I	Int. II	Advanced
Vocabulary	54.1	81.1	82.2	90.0	94.0	86.0
Read. Comp.	95.4	92.5	81.4	88.9	88.7	97.3
Word Study Skills	80.0	90.8	83.6	94.4	94.0	--
Total Math	96.9	97.0	95.8	100.0	99.2	99.2
Spelling	--	74.4	83.0	98.0	95.0	96.7
Language	--	--	85.5	94.9	95.0	91.1
Soc. Sci.	--	81.5	86.4	95.0	92.6	90.0
Science	--	92.6	83.3	91.7	98.3	98.3
Listening	92.3	90.0	90.0	92.0	98.0	--

Table 4. Percentage of TASK '73 items that fit the Rasch Model

Subtest	Level			
	TASK I		TASK II	
	Form A	Form B	Form A	Form B
Reading	93.6	98.7	94.9	87.3
English	98.6	97.1	98.6	82.6
Mathematics	91.7	85.4	93.8	79.2

Table 5. Percentage of SDRT '76 items that do not fit the Rasch Model

	Level						
	Red		Green		Brown		Blue
Subtest	Form A	Form B	Form A	Form B	Form A	Form B	Form A
Auditory Vocab.	69.5	77.8	62.5	82.5	77.5	72.5	--
Auditory Discrim.	65.0	75.0	63.9	88.9	--	--	--
Phonetic Analysis	62.5	72.5	72.2	77.8	77.8	66.7	90.0
Structural Analysis	--	--	81.7	90.0	87.0	77.8	91.7
Word Meaning	--	--	--	--	--	--	100.0
Word Parts	--	--	--	--	--	--	83.3
Read. Comp.	68.7	68.7	70.0	78.3	91.7	78.3	90.0
Word Reading	81.0	88.1	--	--	--	--	--
Scan./Skim.	--	--	--	--	--	--	97.0

Table 6. Percentage of SDMT '76 items that fit the Rasch Model

	Level							
	Red		Green		Brown		Blue	
Subtest	Form A	Form B	Form A	Form B	Form A	Form B	Form A	Form B
Number System & Numeration	63.3	63.3	75.0	72.2	50.0	69.4	58.3	66.7
Computation	84.8	66.7	85.4	72.9	77.1	75.0	79.2	60.4
Applications	90.0	83.3	70.0	73.3	57.6	57.6	78.8	78.8

Table 7. Percentages of Reading Comprehension Items that fit the Rasch Model

Test	Literal	Inferential
SAT, Intermediate I		
Form A	92.5	80.0
Form B	95.5	75.0
Intermediate II		
Form A	95.7	87.5
Form B	91.7	87.2
Advanced		
Form A	95.0	90.7
Form B	100.0	96.2
SDRT, Green		
Form A	80.0	60.0
Form B	86.7	70.0
SDRT, Brown		
Form A	96.7	86.7
Form B	80.0	76.7
SDRT, Blue		
Form A	90.0	90.0

Table 8. Sample Sizes of Rasch Analyses

Test	Level	Form	Sample Size (Approximate)
SAT	SESAT I	A	500
	SESAT II	A	800
	Primary I	A	3600
		B	3300
	Primary II	A	4100
		B	3700
	Primary III	A	4200
		B	3400
	Int. I	A	4500
		B	3800
	Int. II	A	8500
		B	6300
	Adv.	A	8000
		B	7500
	TASK I	A	10000
		B	10000
	TASK II	A	4300
		B	1800
SDRT	Red	A	1500
		B	1400
	Green	A	1600
		B	1500
	Brown	A	900
		B	1500
SDMT	Red	A	1500
		B	1600
	Green	A	1500
		B	1600
	Brown	A	1700
		B	2000
	Blue	A	1500
		B	1600

References

Rentz, R.R., & Rentz, C.C. Does the Rasch Model really work? A synthesis of the literature for practitioners. ERIC Clearinghouse on Tests, Measurement, and Evaluation, December, 1978.

Rentz, R.R., & Ridenour, S.E. The fit of the Rasch Model to achievement tests. Paper presented at the annual meeting of the Eastern Educational Research Association, Williamsburg, Virginia, March, 1978.