DOCUMENT RESUME

ED 400 284 TM 025 552

AUTHOR	Gershon, Richard
TITLE	The Effects of Person-Item Mismatches on the Integrity of the Item Characteristic Curve.
PUB DATE	Apr 92
NOTE	14p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
AVAILABLE FROM	Computer Adaptive Technologies, 2609 West Lunt Ave., Suite 1, Chicago, IL 60645.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Ability; Difficulty Level; Elementary Education; *Elementary School Students; *Estimation (Mathematics); *Guessing (Tests); Item Response Theory; *Vocabulary
IDENTIFIERS	BIGSTEPS Computer Program; *Item Characteristic Function; Rasch Model

ABSTRACT

In 1990 a routine study confirming the accuracy of the Rasch model in predicting item difficulties led to an in-depth analysis of the impact of guessing on the Rasch model when used with multiple-choice items. This paper reviews the highlights of that research. Seventeen linked vocabulary tests of 110 items were each administered to groups of 400 to 600 elementary school students. The obtained data were then analyzed using the BIGSTEPS computer program (B. Wright and M. Linacre, 1992) in order to obtain item difficulties and person ability estimates for the 1,304 items and 7,711 students. The actual performance of each person-item interaction was compared to predicted performance at each interaction point in order to construct an empirical item characteristic curve. The empirical curve was compared to the theoretical Rasch based curve and the discrepancies noted. When restrictions were placed on the analysis, the differences between the empirically derived curve and the theoretical curve were decreased. (Contains two figures, two tables, and five references.) (Author/SLD)

The Effects of Person-Item Mismatches on the Integrity of the Item Characteristic Curve

Richard Gershon

Computer Adaptive Technologies, Inc.
and
Johnson O'Connor Research Foundation

Paper presented at the Annual Meeting of the American Educational Research Association, April 1992, San Francisco. For additional copies of this paper or additional information please contact the author at Computer Adaptive Technologies, 2609 West Lunt Avenue, Suite 1, Chicago, IL 60645 (312) 274-3286.

Abstract

In 1990 a routine study confirming the accuracy of the Rasch model in predicting item difficulties led to an indepth analysis of the impact of guessing on the Rasch model when used with multiple-choice items. This paper will review the highlights of that research. Seventeen linked vocabulary tests of 110 items were each administered to groups of 400-600 students. The obtained data was then analyzed using BIGSTEPS (Wright and Lincacre, 1992) in order to obtain item difficulties and person ability estimates for the 1,304 items and 7,711 students. The actual performance of each person-item interaction was compared to predicted performance at each interaction point in order to construct an empirical item characteristic curve. The empirical curve was compared to the theoretical Rasch based curve and the discrepancies noted. When restrictions were placed on the analysis, the differences between the empirically derived curve and the theoretical curve were decreased.

The Johnson O'Connor Research Foundation (JOCRF) is actively engaged in determining the Rasch based difficulties of all non-technical words in the English language. Each year thousands of multiple choice vocabulary items are written and then assembled into test forms which contain 74 experimental items of similar pre-estimated difficulty, and 36 pre-calibrated linking items of appropriate difficulty for the sample population. Each test form is then administered to 400-600 public and private elementary school students. In 1989, 3500 items were analyzed with approximately 500 persons per item (a total of 22,000 persons). In an effort to demonstrate the accuracy of the Rasch model, all of the new items that had good fit statistics were examined relative to the persons who took the items. Figure 1 was constructed by examining each item-person encounter. A tally was made of all persons at 40 different distances from the difficulty of the item. The distance was defined as person ability minus item difficulty for each item that a person takes. Close to two-million person-item encounters were examined (Gershon, 1990). For reference purposes, a "theoretical" line was drawn according to the basic Rasch probability curve (Wright and Stone, 1979):

$$P(CorrectResponse_{vi}) = \frac{\exp(\beta_v - \delta_i)}{1 + \exp(\beta_v - \delta_i)}$$

Where beta = person ability and delta = item difficulty for each person v and each item i.

items are answered correctly 50% of the time, but the empirically derived curve was lower at this point. The area of discrepancy was observed to be greatest when person ability was approximately 1.5 logits above the item difficulty. The discrepancy at the lower end of the scale is relatively easy to explain in light of the use of multiple choice items. It is relatively intuitive that where multiple choice items are concerned, test takers are likely to have a greater than zero chance of answering an item correctly. It should also not be surprising that the level of success is less than the purely random chance which in this case would have been twenty percent. This is not the first study to show that extremely low able persons are actually less likely to choose the correct answer than any of the other choices. The discrepancy in the middle of the curve is perhaps the most disturbing. While clearly within the error of measure, applications using the Rasch model should be most accurate when ability equals item difficulty. This discrepancy is most likely to be problematic in testing situations which are designed to target the difficulty of the items exactly at the person ability as desired when administering computer adaptive tests.

The discrepancy at the top of the scale is also disturbing. It may be possible that the observed anomaly in this region is due to the same type of activity that occurred at the lower end of the scale. High able persons may be being distracted by an incorrect distractor, in the opposite way that low able persons were less likely than random chance to select the correct answer.

A working hypothesis was constructed in order to explain the discrepancies encountered between the empirical and theoretical item characteristic curves. The first assumption made was that it was inappropriate to use persons whose ability is so far below that of a given item difficulty that guessing plays a role determining the true difficulty of the item. In the case of the Rasch model which does not directly account for behavior of persons within this guessing range (particularly for multiple choice items), the discrepancy created when many persons fall within this low B-D range actually bends the item characteristic curve at both the middle and far ends of the curve. This appears to result in the B-D = 0 range of the empirical curve bending downwards and the B-D = 2 range of the empirical curve bending upwards.

The second assumption was that some of the discrepancy between the observed and expected curves was due to the presence of poorly fitting items. The Wordbook vocabulary testing program does not pre-test items, and therefore there are always misfitting items in the test forms. If these misfitting items were identified by an initial analysis and then removed from later analyses, the resultant item difficulty estimates and related person ability estimates should improve.

Method

A new data set was collected in the fall of 1991 as part of the regular JOCRF Wordbook program. Seventeen forms of increasing pre-estimated item difficulty were constructed. Each test consisted of a single set of 36 linking items which were the same for all 17 forms. An additional 74 unique experimental items were included on each form for a grand total of 1,294 items. The item difficulties were pre-estimated using the ability estimates for vocabulary words found in *The Living Word Vocabulary* (Dale & O'Rourke, 1981) or as pre-estimated by the item writers.

The raw data from each form were combined into one large data matrix using The CAT System Software package (Computer Adaptive Technologies, 1992). This program placed the items into the correct position in the data matrix. The linking items for each form were in the same column position regardless of form, and each unique experimental item had a unique column within the matrix. The program also created BIGSTEPS (Wright & Linacre, 1992) control files with the correctly constructed key for the new data matrix. The 1,184 items which each person *did not* encounter were automatically marked as missing data within the matrix.

The data were then analyzed using three different control files. In Condition 1, the analysis was allowed to proceed using the usual system defaults. No items were anchored, and there was no restriction of data (persons or items). In Condition 2, two new BIGSTEPS parameters were selected. "CUTLO" was set to 1 and "CUTHI" was set to 2. These parameters instructed the program to: a) estimate all person ability and item difficulty parameters using PROX; b) examine each person-item interaction, if the person's ability minus item difficulty (B-D) was less than -1 (CUTLO = 1), or was greater than 2 (CUTHI = 2) the item was marked as missing within the matrix; c) the analysis continued by re-estimating the item difficulties and persons abilities using PROX and then UCON iterations as usual. Condition 3 was the same as Condition 2 with the elimination of 110 items which after the analysis performed for Condition 2 had Mean-Square Infit or Mean-Square Outfit values greater than 1.2 or had a sample of less than 100.

Results

The output files generated from the two BIGSTEPS runs were graphed using the Item Characteristic Curve option in the CAT System. The software examines each person-item interaction using the person and item files generated by BIGSTEPS as well as by comparing the key with the raw data file. For every item the item difficulty is subtracted from the person ability. A tally is then kept for each quarter logit range on the B-D scale of the number of times items were answered correctly versus the number of attempts made in that B-D range. For example, when the person ability is the same as the item difficulty, the Rasch model predicts that 50% of the item person interactions observed will result in a correct response.

The empirically derived results from the analyses undertaken for the three experimental conditions are listed in Table 1, and graphically illustrated in Figure 2. The distribution of person-item interactions are listed in Table 2, and graphically illustrated in Figure 3. Overall one should note that regardless of the constraints placed upon the analysis, the empirically determined Rasch model estimates are extremely good in the $-1 < B-D < 1$ range.

The empirical unedited data set shows responding consistent with that found in the 1989 study. However, the empirical edited data set where person-item interactions fell outside of the -1 < B-D < 2 range, leads to a significantly improved item characteristic curve. This is particularly true in the middle of the range where most of the discrepancy between the theoretical curve and the empirically unedited curves has been eliminated (see Table 1). It is interesting to note that at the lower end of the scale, the edited empirical curve is now closer to the expected random guessing level. In addition, at the other end of the scale, the edited data curve also shows some minor improvement.

Discussion

The reactions of persons viewing this data have ranged from "that's much ado about nothing" to "AHA! Here's proof of the necessity of including a guessing parameter." The aforementioned analyses show that neither of these extreme statements are true.

It is clear that, particularly for data sets where there are many items of extreme difficulty (hard or easy) relative to the sample tested, item difficulty estimates and person ability estimates can be improved by eliminating extreme responding from the calculation of the item parameters. The advent of adaptive testing will eventually eliminate this issue altogether, as persons will [hopefully] be tested only using items which are appropriate for their ability. In the meantime, pre-calibrated item banks are being used for adaptive testing which often maximizes the importance of accurate ability estimates at the B-D = 0 point. This situation increases the importance of having accurate item difficulty estimates, and reinforces the need to follow the procedures outlined for editing data sets.

The Rasch model was used to accurately predict performance across the ability continuum, in spite of a potentially problematic data set where guessing was an issue. This is particularly notable in light of the relatively small sample sizes used (especially when compared to the sample sizes necessary for predicting multi-parameter models), and also considering the accuracy of the difficulty estimates calculated prior to the elimination of bad items. In almost any field, researchers are frequently hard pressed to find a theory which best describes their data. The Rasch model was clearly able to describe the items and persons in this data set with ease.

Table 1

835,000 Grade School Responses to 5-Choice Vocabulary Items

	Rasch			No Misifts
B-D	Theory	Unedited	Edited	Edited
-4.00	0.02	0.136	0.160	0.174
-3.75	0.02	0.152	0.162	0.187
-3.50	0.03	0.135	0.172	0.180
-3.25	0.04	0.146	0.164	0.186
-3.00	0.05	0.136	0.176	0.192
-2.75	0.06	0.141	0.175	0.194
-2.50	0.08	0.151	0.189	0.204
-2.25	0.10	0.158	0.197	0.209
-2.00	0.12	0.172	0.209	0.226
-1.75	0.15	0.188	0.221	0.236
-1.50	0.18	0.209	0.227	0.245
-1.25	0.22	0.231	0.236	0.248
-1.00	0.27	0.261	0.285	0.296
-0.75	0.32	0.302	0.327	0.342
-0.50	0.38	0.348	0.368	0.388
-0.25	0.44	0.400	0.423	0.440
0.00	0.50	0.466	0.479	0.490
0.25	0.56	0.534	0.544	0.547
0.50	0.62	0.600	0.606	0.600
0.75	0.68	0.668	0.667	0.654
1.00	0.73	0.737	0.731	0.711
1.25	0.78	0.789	0.785	0.762
1.50	0.82	0.839	0.834	0.813
1.75	0.85	0.875	0.873	0.849
2.00	0.88	0.909	0.910	0.890
2.25	0.91	0.934	0.938	0.926
2.50	0.92	0.950	0.943	0.930
2.75	0.94	0.962	0.950	0.940
3.00	0.95	0.974	0.960	0.951
3.25	0.96	0.981	0.963	0.957
3.50	0.97	0.986	0.971	0.968
3.75	0.98	0.990	0.977	0.963
4.00	0.98	0.990	0.976	0.972

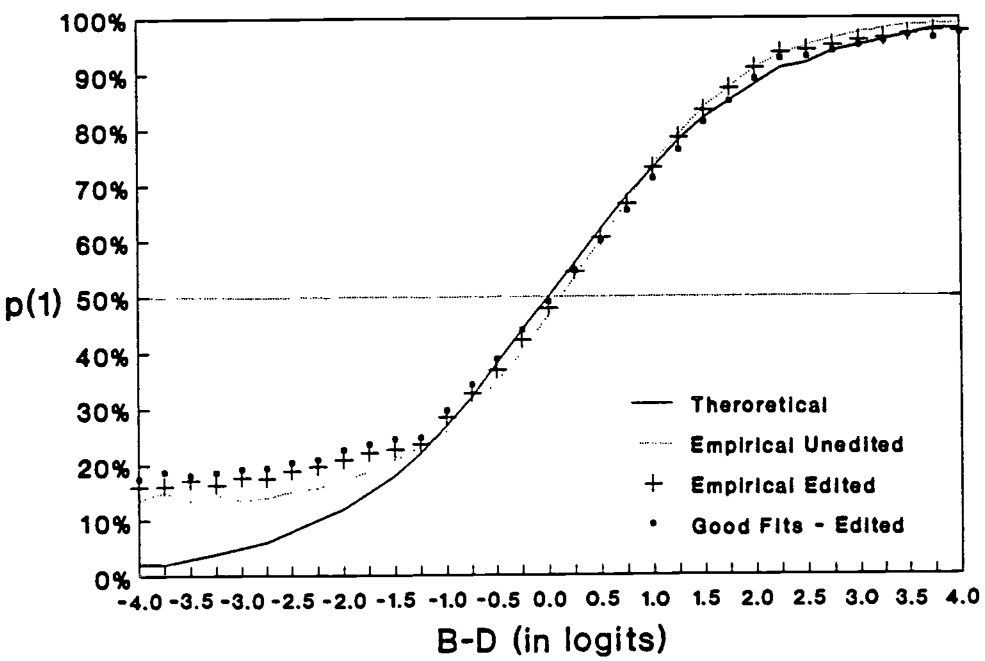
Table 2

Distribution of Person-Item Interactions

			No Misifts
B-D	Unedited	Edited	Edited
-4.00	354	2175	1342
-3.75	718	3043	1949
-3.50	1282	4196	2728
-3.25	2049	5471	3881
-3.00	3269	7089	4983
-2.75	5277	9379	6973
-2.50	7405	11846	9119
-2.25	11047	14840	11504
-2.00	14963	18437	14847
-1.75	20245	22401	18332
-1.50	25862	26610	22675
-1.25	32276	31529	27704
-1.00	39088	36758	33028
-0.75	45035	40935	37738
-0.50	51151	45265	42488
-0.25	55327	48562	46453
0.00	58616	51208	49020
0.25	59093	53022	51785
0.50	59601	53605	52133
0.75	56658	52294	50433
1.00	53288	49937	48136
1.25	48575	47005	44280
1.50	42412	42815	39468
1.75	36185	37720	34754
2.00	29611	32325	28706
2.25	23170	25875	22738
2.50	17085	19350	16370
2.75	12734	13851	11275
3.00	8582	9198	7253
3.25	5540	5668	4330
3.50	3363	3233	2277
3.75	2155	1882	1238
4.00	1207	991	577

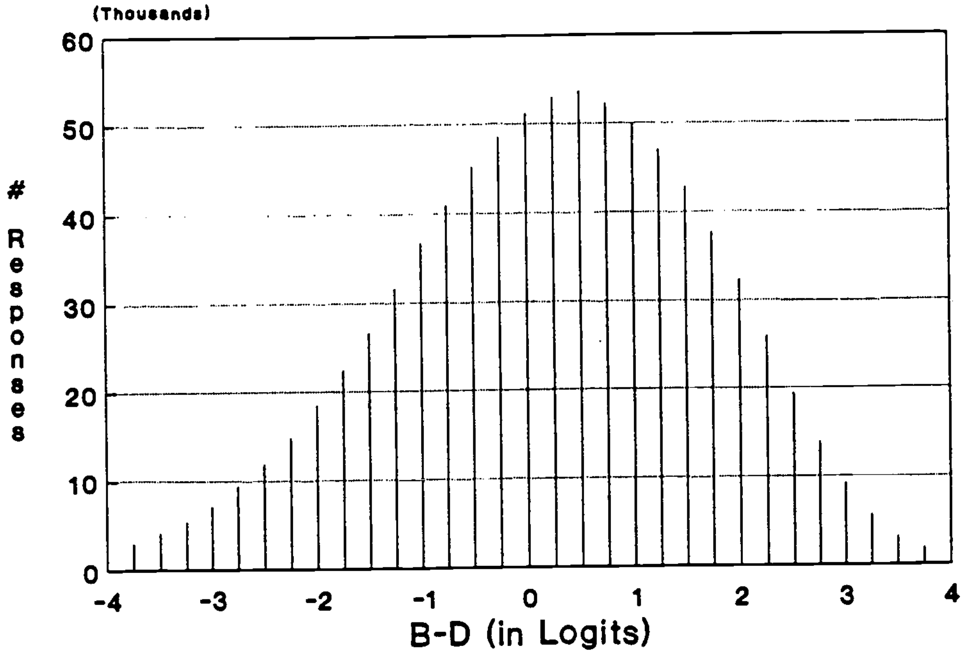
Figure 2

835,000 Grade School Responses to 5-Choice Vocabulary Items
With and Without Response Editing



835,000 Grade School Responses to 5-Choice Vocabulary Items

Distribution of Person-Item Interactions



References

Dale, E. & O'Rourke, J.(1981). *The Living Word Vocabulary*. Chicago: World Book-Childcraft International, Inc.

Gershon, R. (1992). *The CAT System* [software program]. Chicago: Computer Adaptive Technologies, Inc.

Gershon, R. (1990). *Rasch-Model Procedures Used to Build the JOCRF Vocabulary Item Bank*. Technical Report 1990-3. Chicago: Johnson O'Connor Research Foundation.

Wright, B. & Linacre, M. (1992). *BIGSTEPS* [software program]. Chicago: MESA Press.

Wright, B. & Stone, S. (1979). *Best Test Design*. Chicago: MESA Press.

NCME Annual Meeting, April 9-11, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Effects of Person-Item Mismatches on the Integrity of Item Characteristic Curve

Author(s): Richard Gershon

Corporate Source: Computer Adaptive Technologies

Publication Date: 4/92

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑	◀ **Sample sticker to be affixed to document**	**Sample sticker to be affixed to document** ▶	□
Check here Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction	"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."	"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."	**or here** Permitting reproduction in other than paper copy.
	Level 1	**Level 2**	

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: B. Bergstrom	Position: Dir, Psychometric Services
Printed Name: Betty Bergstrom	Organization: Computer Adaptive Technologies
Address: 2609 W Lunt Ave #2E Chicago IL 60645	Telephone Number: (312) 274-3286
	Date: 4/22/96

OVER