

Chapter 4

Diagnostic Opportunities with the Rasch Model for Ordered Response Categories

David Andrich, John H.A.L. de Jong and Barry E. Sheridan

Murdoch University, Western Australia
CITO, The Netherlands
Edith Cowan University, Western Australia

1. Introduction

Items with ordered response categories are used in social and other sciences when ideally measurements would be used, but no measuring instrument is available. They are similar to measurements in that they partition a latent unidimensional continuum into adjacent intervals. In elementary treatments of ordered response categories, they are in fact treated as measurements in that successive categories are simply assigned successive integers. This assumes that the distances between the thresholds which define the intervals are identical. In more advanced treatments, statistical models which parameterize the thresholds and values of objects of classification are applied.

Central to the understanding of what constitutes more or less of the property on the continuum is the definition of the successive categories which reflect successively more of the property. However, there is no guarantee that the categories will operate in the way intended. Therefore, this ordering must be treated as an hypothesis about the data and it is important that the statistical model applied has the property that it can reject this hypothesis. That is, the ordering must be a property of the data themselves, not simply the model. The significance of this feature was articulated by Ronald Fisher (1958) who proposed a least squares method for analyzing such data. Following a demonstration of his procedure on a small example involving 12 serological readings classified into 5 levels of reaction, he wrote: *It will be observed that the numerical values . . . lie . . . in the proper order for increasing reaction. This is not a consequence of the procedure by which they have been obtained, but a property of the data examined* (Fisher, 1958, p. 294).

Even though it was none other than Fisher who made the above point, it seems to have been generally ignored in the analysis of ordered categorical data, both in the choice of models and in exploiting model which can reveal disordering in the data. This paper illustrates, with two examples, how the Rasch model for ordered categories can be used to reveal whether or not ordered categories operate as intended, and the diagnostic opportunities this provides.

2. The Rasch Model for ordered response categories

The Rasch model for $m + 1$ ordered response categories can be expressed in different forms, all equivalent to

$$p(X_{vi} = x) = \frac{1}{\gamma_{vi}} \exp \left[-\sum_{s=1}^x \tau_{is} + x(\theta_v - \sigma_i) \right] \quad (1)$$

where $X_{vi}, x_{vi} \in \{0,1,2,\dots,m\}$, is a random variable taking the values of successive integers for the successive categories, θ_v and σ_i are respectively the locations of person v and item i , $\tau_{is}, s = 1, \dots, m$, are thresholds which partition the continuum, and $\gamma_{vi} = \sum_{x=0}^m \exp \left[-\sum_{s=1}^x \tau_{is} + s(\theta_v - \sigma_i) \right]$ is the normalizing factor (see eq. (14) in chapter 1.4).

Because the mean of all thresholds is parameterized by σ_i , all thresholds add up to zero, $\sum_{x=1}^m \tau_{ix} = 0$, for each item i . This form of the model will be termed simply the unidimensional

Rasch model (URM). Figure 1 shows the category response functions (CRFs) for the case of five categories and four thresholds.

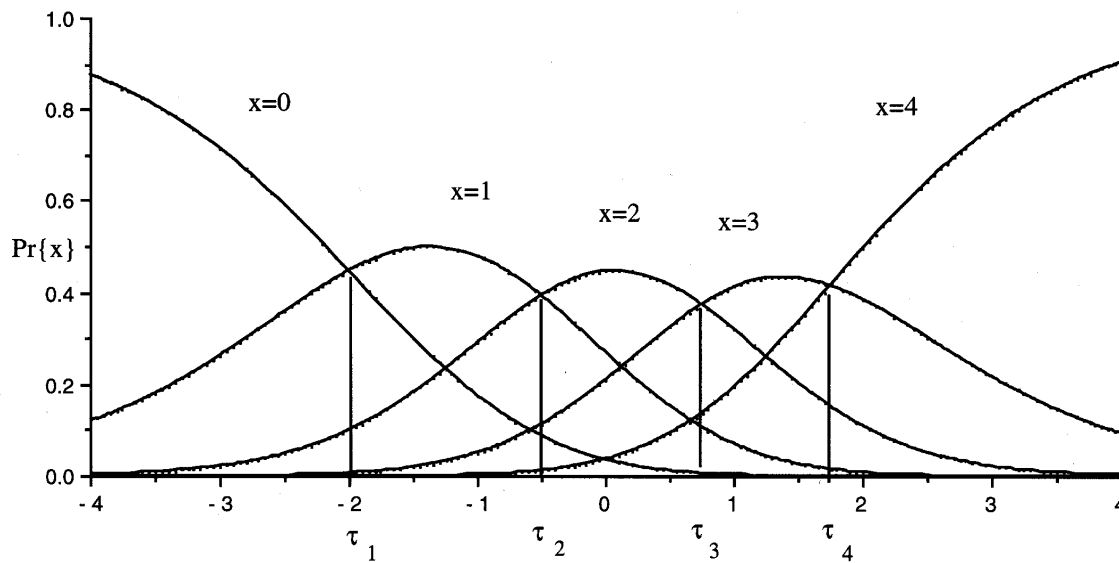


Figure 1: Response functions for five ordered categories in the unidimensional Rasch model

The construction of this model has been described elsewhere (Andrich 1978, 1979) and has been compared with the more traditional model for ordered response categories, that based on the work of Thurstone which does not have the property that it can reveal whether or not the intended ordering of the categories is reflected in the data (Andrich 1995a, 1995b), and therefore only certain features of this construction, and properties relevant to this paper, will be highlighted.

First, the model is constructed to satisfy the criterion of specific objectivity (Rasch 1961, 1977), a criterion which is independent of data. This criterion is consistent with the criteria

set for measurement by both Thurstone (1927) and Guttman (1950), which were also independent of data, but is more comprehensive.

Second, because the denominator contains all thresholds, the probability of a response in any category depends on the locations of all thresholds, not just the ones defining the category. This reflects the property that the model characterizes a response at each threshold which, however, must be constrained to satisfy the Guttman structure (Andrich, 1985).

Third, the model has the feature that if the data fit the model with $m + 1$ categories, then if the data are collapsed into $m' < m + 1$ categories, they will not fit as well the model with m' categories (Rasch, 1966; Andersen, 1977; Andrich 1978, 1995c; Jansen and Roskam, 1986). This means that the model is sensitive to the number of categories in the data collection.

Finally, the threshold estimates can be in any order, and not simply the intended order (Andrich, 1982). Although this is the feature of the model most directly relevant to this paper, all of the above features are related. These are briefly reconsidered following the exposition of the examples.

3. Examples

This section contains two examples which illustrate applications of taking the ordering of the thresholds in the URM as an hypothesis. In both examples, an original data collection and a second one based on an interpretation of the reversed thresholds in the first data collection, are available.

3.1 Assessment of language proficiency

In a research project (The Dutch Institute for Educational Measurement, CITO), a tape-mediated, semi-direct, oral assessment test (Clark 1979, 1986, 1988) was developed to test functional communicative oral proficiency. A full description of the tests, which consisted of five sections, is given in De Jong and Van Ginkel (1992). In this example, only the data gathered from section 2 consisting of 10 trichotomously scored items is reported, although the results arise from a simultaneous analysis of the first two sections where the first section consisted of 20 dichotomously scored items.

In section 2, each person is presented with 10 questions on a particular theme where the first few words of an answer are provided in the test booklet. The person has to complete these answers with the aid of visual information also provided in the booklet. Responses require use of particular grammatical constructions and are scored as follows: 0 in the case of no response or an illogical/meaningless response; 1 in the case of a meaningful response containing serious errors that may hinder communication; 2 in the case of a meaningful response which is almost error free. Clearly, the greater the score, the greater the implied latent ability.

The test was administered to 25 persons. Three copies of the taped responses were made and assigned at random without replacement so that each response was rated by three independent raters. This design permitted the $(3) (25)$ response vectors to be regarded as 75 independent persons. This is a small sample, but it was large enough for a piloting of a first trial of a test.

The full analysis of these data, including a test of fit, is provided in De Jong (1991, pp. 87-96). Here, only the difficulty and threshold estimates of the item, which were obtained using a pairwise conditional procedure in which the abilities of the persons were eliminated (Andrich, 1988; Choppin, 1983) was used. Table 1 shows these parameter estimates. It is evident that items 5, 7 and 9 have reversed threshold estimates.

Item No <i>i</i>	Difficulty $\hat{\sigma}_i$	Thresholds	
		$\hat{\tau}_{i1}$	$\hat{\tau}_{i2}$
1	0.711	-0.954	0.954
2	0.866	-1.038	1.038
3	0.117	-0.997	0.997
4	-0.041	-1.062	1.062
5*	-1.303	1.260	-1.260
6	-0.066	-1.454	1.454
7*	1.026	0.681	-0.681
8	0.269	-0.438	0.438
9*	0.318	0.172	-0.172
10	0.509	-0.996	0.996

* $\hat{\tau}_{i2} < \hat{\tau}_{i1}$

Table 1: Difficulty and threshold estimates in items from a language proficiency test

A detailed study of the statistics of item 5 (though items 7 and 9 would have been just as useful) is now provided and interpreted. Consider this item's CRFs shown in Figure 2. For a person with ability θ in the range -2.563 to -0.043, where a score of 1 is expected to be most likely, scores of 0 and 2 are both more likely than 1. This implies a bimodal distribution of the scores 0, 1 and 2, which it is stressed, is not a result of a bimodal distribution of abilities across people - the bimodal distribution is for a single person responding to this item with reversed thresholds estimates.

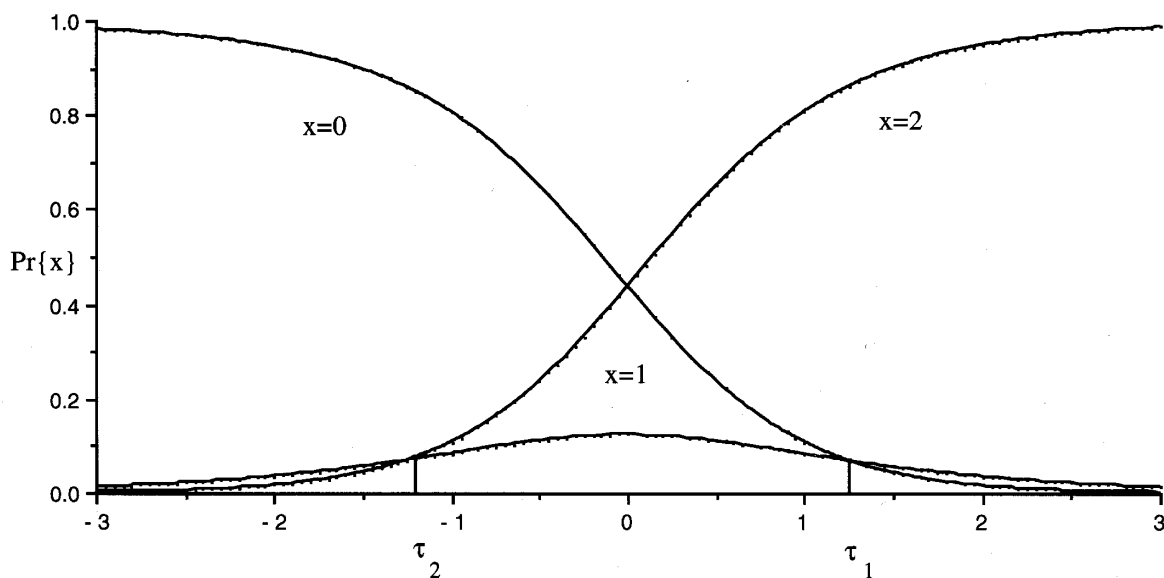


Figure 2: Response functions for a trichotomously scored item with reversed thresholds

Two related points are emphasized. First, the intention is that the categories reflect order in the achievement. Therefore, a student of low ability who might be expected to score 0 should always have a greater probability of scoring 1 than 2 - with reversed thresholds, this is not always the case, as evident from Figure 2. To consolidate this point, consider a person whose ability is exactly equal to the difficulty of an item. Taking item 5 with difficulty -1.303 and threshold values -1.260 and 1.260 as an example, and inserting $\theta_v = -1.303$ into the URM of equation (1), gives response probabilities of $p(0) = 0.44$, $p(1) = 0.12$ and $p(2) = 0.44$ and a mean value of $E[X_{vi}] = 1.0$.

Two of these results are consistent with requirements for three ordered categories - the mean value of 1.0 is the middle score and should be the theoretical average score when a person's ability is at the item's difficulty. However, the probability of a score of 1 is lower than that of a score of both 0 and 2, in this case more than twice as low. In fact, when thresholds are reversed in a trichotomous item it is never the case, no matter where the location of the person, that the probability of a score of 1 is the greatest. In contrast, when they are properly ordered, the probability of a middle score of 1 occurs exactly as expected - when θ_v is in the range $\sigma_i + \tau_{i1} < \theta_v < \sigma_i + \tau_{i2}$.

Second, in the case where the distribution of scores is bimodal, one could simply say that the middle category is less popular, or less attractive (Fischer and Parzer, 1991). However, that is *only a description of the distribution, and not an explanation of why the middle category is less attractive than it should be for those for whom it should be the most attractive*.

Many factors can go wrong when data are collected, and the above analysis itself cannot tell the specific source of the problem - it can only tell that there is a problem. Item 5 was studied in order to try to understand the source of the problem which, in this case, was easy to diagnose. Inadvertently, the test constructor used the same visual stimulus as was used in the example item, and although the verbal stimulus was not identical, the most appropriate response was the same as in the example item. This made the item in general easy. It also meant that those who identified the item with the example scored 2 and those who did not scored 0. Explanations for the source of the problems in items 7 and 9 are provided in De Jong (1991, pp. 93-94), but it can be noted that they all involved a component of knowing *what* to say as well as how to say it, when the intention was to assess only *how* to say it (assuming students knew what to say). Items 5 and 9 were modified to overcome the problem diagnosed as the source of reversed thresholds, item 7 was discarded, and the items readministered to a larger sample of 135 students randomly sampled from 1000 final grade students at 25 different schools in the Netherlands. Each student was rated by two independent raters, and of the $2(135)$ independent response vectors, 265 could be used. The difficulty estimates of both items now became less extreme, but more importantly, the thresholds were properly ordered: -1.957 and 1.957 for item 5 and -0.853 and 0.853 for item 9. These results confirmed (not proved of course) that the source of a problem with each item had been diagnosed and overcome correctly. Moreover, and quite independently of how the diagnosis was made, the problem identified with the items was considered a genuine problem which reduced the validity of the items.

Before concluding this example, it is pointed out that in responding to these items, the students respond holistically and therefore in principle were able to respond at both

thresholds. The task as to whether they responded above both thresholds, below both thresholds, or ahead of the first but not the second, was decided by the rater.

3.2 Teachers attitudes towards direct instruction

The second example involves the very familiar Likert (1932) questionnaires, where the response format is of the form of Strongly Disagree (SD), Disagree (D), Not Sure (NS) or Undecided (U), Agree (A) and Strongly Agree (SA). By being placed in the middle of the response continuum, it is clearly intended that the NS or U category imply an attitude somewhere in between Disagreement and Agreement. However, research which explicitly checks whether this category does in fact work as intended, shows that this is not always the case (Dubois and Burns, 1975). Bock & Jones (1968, p.13) indicate that the consensus then was that such a category should not be included. In particular, it seems to attract responses for reasons other than that the person's attitude is somewhere in the middle of the continuum. Nevertheless, it continues to be used as a middle category. Example 2 specifically considers the operation of the middle category of a Likert response format.

The data for the example arise from a study at Edith Cowan University (Western Australia) which investigated the attitude of teachers towards a strategy of "Direct Instruction" teaching which was new to them. Attitudes towards this strategy were assessed before and after teachers were actually exposed to the strategy, using a 40 item questionnaire covering both general and specific aspects of the strategy. The number of teachers was 144. The statements in the questionnaire were of the kind *Direct instruction meets my students' academic needs*. Clearly, agreeing to the statement implies approval, and disagreeing implies disapproval. Thus the greater the score across the items, the greater the approval. Some statements had a negative construction and were reverse-scored.

Because it was known that some teachers may know nothing about the strategy, it was considered important to have a category such as NS available. However, even before analyzing any data, it seems reasonable to be concerned that if the category really attracts people who are unsure because they know nothing about the strategy, that it will not represent the same meaning as it does for those who know the strategy, but are not sure whether or not it will meet their students needs. In that case, it seemed unreasonable that a NS response should receive a middle rating as if it really meant a moderate level of approval rather than just ignorance about the strategy. Nevertheless, the traditional response format of SD, D, NS, A, and SA was used in the pre-test.

Table 2 shows the threshold estimates for the pre-test and post-test. Because these estimates are expected to be independent of the actual attitudes of the persons if the data fit the model, a comparison of the threshold values and the average location values between the pre - and post-test to check if these have changed is relevant. If they have changed, then the difference in attitude before and after is not just quantitative, but qualitative (Andrich, 1985). However, here the concern is only with the ordering of the thresholds.

It is evident that there are 8 items with reversed thresholds in the pre-test, and 19 in the post-test. Thus the NS category seemed to work better as a middle category when most of the teachers knew nothing about the strategy than when they all knew the strategy. This reinforces the earlier interpretation that this category attracts people who are unsure because they know nothing about the programme. The CRFs for item 20 of the post-test data are shown in Figure 3.

item	Pretest (N = 144)				Posttest (N = 144)					
		1	2	3	4	1	2	3	4	
1	*	-1.16	-1.84	0.72	2.28	*	-2.10	0.03	-0.31	2.39
2		-2.07	-0.49	-0.22	2.78	*	-1.73	1.61	-1.66	1.78
3		-2.00	-0.89	0.64	2.25	*	-2.29	0.12	-0.17	2.34
4	*	-1.55	0.04	-0.23	1.74		-1.91	1.34	0.10	0.47
5		-3.51	-1.27	1.46	3.32		-3.08	-1.10	0.76	3.41
6		-2.52	-0.61	0.37	2.77		-1.79	-0.31	-0.25	2.35
7	*	-1.04	-0.34	-0.66	2.05	*	-1.51	0.24	-1.18	2.45
8		-3.47	-0.47	0.65	3.28		-2.15	-0.15	0.21	2.08
9	*	-1.27	-2.13	1.12	2.28	*	-1.47	-2.15	1.38	2.24
10		-1.53	-0.32	-0.21	2.06	*	-1.36	0.14	-1.24	2.46
11		-1.47	-1.12	-0.15	2.75		-2.15	-0.35	-0.11	2.61
12		-1.87	-0.49	-0.43	2.78	*	-2.34	0.99	-1.63	2.98
13	*	-2.57	-0.29	-0.32	3.18	*	-1.96	0.51	-0.87	2.32
14	*	-1.33	0.00	-0.70	2.04	*	-0.89	-0.95	-1.05	2.89
15		-1.91	-1.81	0.67	3.05		-2.35	-1.38	1.41	2.33
16		-1.62	-1.01	0.20	2.43	*	-1.25	-1.26	-0.13	2.64
17		-2.48	-0.78	0.24	3.01		-1.78	-0.60	-0.08	2.46
18		-2.03	-0.97	0.25	2.75	*	-1.87	0.27	-0.94	2.54
19		-2.13	-0.86	0.29	2.69		-2.68	-0.68	0.71	2.66
20		-3.22	-0.84	-0.07	4.13	*	-2.23	0.68	-1.12	2.66
21		-2.05	-1.08	0.23	2.90		-2.51	-1.26	0.22	3.54
22		-2.68	-0.78	0.64	2.82	*	-2.53	-0.25	-0.45	3.22
23		-2.79	-0.90	0.11	3.59	*	-2.69	0.46	-0.64	2.87
24		-2.11	-1.67	0.20	3.58		-2.18	-0.60	0.04	2.74
25		-2.66	-1.92	0.54	4.05		-2.88	-0.85	0.40	3.33
26		-2.68	-0.23	-0.12	3.04	*	-2.43	1.00	-1.16	2.60
27		-2.56	-1.37	0.57	3.36		-2.54	-0.71	0.80	2.45
28		-2.33	-1.71	0.51	3.53		-2.46	-1.00	0.79	2.67
29		-3.03	-1.51	1.19	3.35		-2.11	-0.83	0.13	2.80
30	*	-0.72	-2.25	0.67	2.29	*	-0.73	-1.88	-0.01	2.63
31		-1.70	-1.65	-0.05	3.40		-1.99	-1.96	0.43	3.52
32		-1.90	-1.06	0.17	2.78		-1.73	-0.88	-0.43	3.04
33		-1.89	-1.26	0.31	2.84	*	-2.50	-0.06	-0.10	2.66
34		-2.54	-1.87	0.80	3.61		-2.24	-0.96	0.69	2.51
35	*	-1.67	-2.06	0.81	2.91		-3.05	-1.04	1.12	2.97
36		-2.13	-1.42	1.03	2.53		-2.37	-0.92	0.48	2.81
37		-2.51	-0.87	0.47	2.90		-2.05	-0.65	-0.58	3.29
38		-2.89	-0.95	0.82	3.01	*	-2.40	0.06	-0.45	2.80
39		-1.93	-1.17	0.06	3.05		-1.49	-0.73	0.02	2.21
40		-2.24	-1.34	0.46	3.12		-2.18	-0.88	-0.22	3.28

Table 2: Threshold estimates from the response format Strongly Disagree, Disagree, Not Sure, Agree, Strongly Agree.

It is evident from these CRFs that the middle category operates in such a way that the probability of someone responding in it is never greater than the probability of responding in the other categories. Thus even people whose location is such that one would expect them to have the greatest probability of responding in the middle category, have a greater probability of responding in other categories. This is evidence that the category is not operating as a category in the middle of the other four categories. Thus if the teaching strategy were to be

assessed for approval, the scoring of the NS category as if it is operating in the middle of the other categories is not tenable.

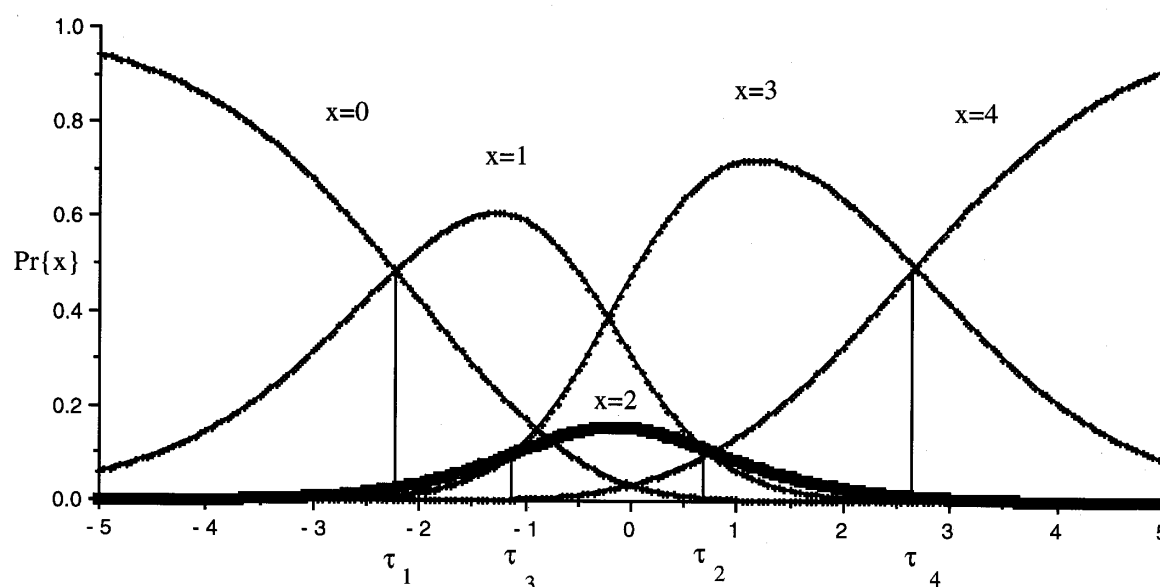


Figure 3: Response functions for a Likert-style item with 5 categories (item 20 of the post test) and disordered threshold estimates

In a second part of the study, 167 teachers received a response format which had the NS category to one side and distinctly separate from the other categories, and 108 had a response format which did not have the NS category available at all. The two formats were (SD, D, A, SA; NS) and (SD, D, A, SA). Table 3 shows the threshold estimates for this set of responses, where in the first response format the NS response was treated as missing data. It is evident that now only three items have reversed thresholds when the NS category is available, and only one item when the NS category is not available at all. As already indicated, other analyses could also be carried out to study the effects of response format, but here the focus is on the NS category and its symptoms in the URM when placed in the middle of the Likert format.

Two points arise from this analysis. First, it confirms concerns with the middle category designated as Neutral, Not Sure or Undecided in the Likert-style response format, and indicates that in this case it should not be treated as an attitude more or less somewhere between a negative and positive attitude. Second, the elimination of the NS category has reduced the number of items with reversed thresholds to a point where it seems that the few items with reversed thresholds themselves deserve closer study.

It can be seen, even from a cursory glance at Table 3, that the threshold estimates are further apart when the NS category is not available, than when it is available but the responses to it are treated as missing data. (An NS response is treated as missing at the level of the response - responses of the person to other items are included in the analysis). Thus the very inclusion of the NS category seems to affect the responses in the other categories, and deserves further investigation.

This sensitivity of the responses to all categories as a function of the change in one category is a property of the data collection, not simply the model used to analyze it. However, the URM is particularly suitable for this purpose because the model too, as stressed earlier, is sensitive to the number of categories - the probability of a response in any category is a function of all of the thresholds, and therefore all of the categories. From a response process point of view, it is also clear that a person can see all categories and can consider all categories simultaneously, and then make a response. That is, there is an opportunity for the person to make a decision at all thresholds (constrained by the Guttman pattern), *making it a simultaneous response process*.

Item	<i>NS data suppressed</i> (<i>N</i> = 167)			<i>NS not included</i> (<i>N</i> = 108)				
	1	2	3	1	2	3		
1		-1.51	-0.47	1.98		-4.29	0.26	4.03
2		-1.44	0.26	1.18		-2.94	0.19	2.76
3		-2.09	-0.05	2.14		-3.70	-0.18	3.88
4		-1.48	0.33	1.14		-3.50	0.39	3.12
5		-2.52	0.40	2.12		-4.15	-0.36	4.50
6		-1.43	0.39	1.04		-3.37	-0.56	3.94
7		-1.50	-0.31	1.80		-2.47	-0.19	2.67
8		-2.01	0.21	1.80		-4.53	-1.37	5.90
9		-1.15	-0.80	1.96	*	-2.22	-2.74	4.96
10		-1.58	-0.30	1.88		-1.90	-0.40	2.30
11		-0.99	-0.65	1.64		-3.34	-1.20	4.54
12		-1.55	-0.13	1.68		-3.24	-1.42	4.66
13		-1.05	0.09	0.96		-2.77	-0.98	3.75
14	*	-0.69	-1.47	2.16		-1.65	-1.60	3.25
15		-1.71	0.35	1.36		-3.81	-0.25	4.06
16		-0.73	-0.54	1.28		-2.29	-0.71	3.00
17		-1.02	-0.84	1.85		-2.18	-0.57	2.75
18		-1.13	-0.63	1.76		-2.36	-0.45	2.81
19		-1.12	-0.90	2.03		-3.45	-0.21	3.66
20		-1.49	0.14	1.35		-3.54	-0.85	4.40
21		-2.16	-0.28	2.44		-3.61	-0.72	4.33
22		-1.72	-0.18	1.90		-3.20	-1.02	4.22
23		-2.26	0.64	1.62		-3.66	-0.37	4.02
24	*	-1.52	0.18	1.33		-2.76	0.41	2.34
25		-1.96	-0.03	1.99		-4.84	0.29	4.56
26		-1.73	1.14	0.59		-3.29	-0.95	4.24
27		-1.75	-0.10	1.86		-3.97	-0.01	3.97
28	*	-1.99	0.24	1.75		-4.03	-0.96	4.99
29		-1.63	-0.17	1.80		-2.82	-1.15	3.97
30		-0.57	-1.18	1.74		-1.73	-1.57	3.30
31		-1.39	-1.26	2.65		-2.24	-2.00	4.24
32		-1.57	-0.61	2.18		-3.47	-2.59	6.06
33		-1.86	0.07	1.79		-3.92	-0.45	4.37
34		-1.84	-0.23	2.07		-4.41	-0.43	4.83
35		-2.02	-0.68	2.70		-3.84	-0.70	4.53
36		-1.30	-0.26	1.56		-4.18	-0.22	4.40
37		-1.11	-0.64	1.75		-2.42	-1.16	3.58
38		-1.30	-0.54	1.84		-1.74	-1.14	2.87
39	*	-0.60	-1.08	1.68		-2.36	-0.36	2.72
40		-1.65	-0.43	2.07		-3.31	-0.42	3.73

Table 3: Threshold estimates from response formats Strongly Agree, Agree, Disagree Strongly Disagree and Not Sure where Not Sure is treated as missing data and from the response format Strongly Agree, Agree, Disagree, Strongly Disagree

4. Summary and Discussion

This paper considers the unidimensional Rasch model for ordered response categories, emphasizing that the model is established independent of any data and that it can be used as a criterion which data should meet if they are to be transformable into measurements. Accordingly, it is argued that the model is an hypothesis about the data. In particular, it shows that the ordering of the thresholds which divide the latent unidimensional continuum into categories itself is a hypothesis about the data which is embedded in the model. Although formal statistical tests of fit could be constructed to test that the thresholds are ordered in the data, reversed threshold estimates is sufficient evidence to conclude that the empirical ordering is not consistent with the intended ordering.

Two examples are provided, one involving achievement testing in the form of a graded response, and the other Likert-style questionnaires with a NS category, showing how the reversal of thresholds permits diagnosing incorrect empirical ordering of categories.

Because the very construction of the model requires an ordering of thresholds, it is argued here that whenever the threshold estimates are reversed, it provides evidence that the ordering is not operating as intended. In addition, it is emphasized that a perspective that does not take the reversal of threshold estimates to be a real problem of ordering, would provide no basis for examining the statements in the first place with a view to explaining and understanding the incorrect ordering.

It is stressed that the cause of the reversal of threshold estimates cannot be determined by the statistical analysis. There are many possible causes, but descriptively the violation of the order may result from multidimensional rather than unidimensional responses, from different discriminations at the thresholds, from the lack of capacity to use all of the assigned number of categories, or from a genuine empirical disordering of the categories. Whatever the source, the disordering of the threshold estimates is clear and unambiguous evidence of problems in the empirical ordering.

One final comment is necessary. The argument that items with ordered response categories requires the threshold estimates to be ordered can be troublesome because the experience of the authors is that many data sets have items that reveal disordering. This implies at least that some collapsing of categories is required in order that, first an overestimate of the precision of the locations is not presumed, and second, that the actual operation of the categories is understood better. This is entirely consistent with the property of the URM that if the data do fit it with a given number of categories and the thresholds estimates are ordered, then any collapsing of the data to a smaller number of categories will reveal a poorer fit to the model. On the other hand, if the thresholds are not ordered, or if other fit statistics reveal misfit between the data and the model, then some collapsing may reveal the effective number and ordering of categories *post-hoc* (Wright, 1994), which ideally would be checked with a follow-up study using the new format. Whichever approach is taken, collapsing existing categories or restructuring the format for further data collection, understanding and correcting disordering provides unique diagnostic opportunities for understanding the variables of measurement operationalised by items with ordered response categories.

References

- Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42, 69-81.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 357-374.
- Andrich, D. (1979). A model for contingency tables having an ordered response classification. *Biometrics*, 35, 403-415.
- Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, 47, 105-113.
- Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon Tuma (Ed.), *Sociological Methodology*. Chapter 2, pp. 33-80. San Francisco: Jossey-Bass.
- Andrich, D. (1988). *Rasch models for measurement*. Sage University Papers Series on Quantitative Applications in the social sciences, 07-001. Beverley Hills: Sage
- Andrich, D. (1995a). Measurement criteria for choosing among models with graded responses. In A. von Eye & C.C. Clogg (Eds) *Categorical Variables in Developmental Research*. Academic Press. Chapter 1, pp 3- 35.
- Andrich, D. (1995b) Distinctive and incompatible properties between two models for graded responses. *Applied Psychological Measurement*, 19, 101- 119.
- Andrich, D. (1995c) Models for measurement, precision and the non-dichotomization of graded responses. *Psychometrika*, 60, 7-26.
- Bock, R.D., & Jones, L.V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden Day.
- Choppin, B (1983) A fully conditional estimation procedure for Rasch model parameters. Report No. 196, Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.
- Clark, J.L.D. (1979). Direct vs. semi-direct tests of speaking proficiency. In: E.J. Briere and F.B. Hinofotis (Eds) *Concepts in Language Testing: Some Recent Studies*. Washington, DC: Teachers of English to Speakers of Other Languages.
- Clark, J.L.D. (1986). Development of a tape-mediated, ACTFL/ILR scale-based test of Chinese speaking proficiency. In: C.A. Stansfield (Ed.) *Technology in Language Testing*. Washington, DC: Teachers of English to Speakers of Other Languages.
- Clark, J.L.D. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency, *Language Testing*, 5, 187-205.
- De Jong, John H.A.L. (1991) *Defining a variables of foreign language ability: An application of item response theory*. Den Haag: Cip-Gegevens Koninklijke Bibliotheek.
- De Jong J.H.A.L. & and Van Ginkel, C.W. (1992) Dimensions in oral foreign language proficiency. In L.T. Verhoeven & J.H.A.L. de Jong (Eds.) *The construct of language proficiency: Applications of psychological models to language assessment*. Amsterdam: John Benjamins.
- Dubois, B. & Burns, J.A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, 35, 869-884.
- Fisher, R. A. (1958) *Statistical methods for research workers*. (13th edition) New York: Hafner Publishing Co.
- Fischer, G.H. & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika*, 56, 637-651.
- Guttman, L (1950) The basis for scalogram analysis. In S.A. Stouffer, et al (Eds.) pp. 60-90. *Measurement and Prediction*. New York: Wiley.
- Jansen P.G.W. & Roskam, E.E. (1986) Latent trait models and dichotomization of graded responses. *Psychometrika*, 51, 69-91.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.) *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. IV, 321-334. Berkeley CA: University of California Press.

- Rasch, G. (1966) An individualistic approach to item analysis. In *Readings in Mathematical Social Science*. P.F. Lazarsfeld and N.W. Henry, (Eds.) Chicago: Science Research Associates pp. 89-108.
- Rasch, G. (1977) On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy* 14, 58-94.
- Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, 34, 278-286.
- Wright, B.D. (1994) Rasch sensitivity and Thurstone insensitivity to graded responses. *Rasch Measurement Transactions*, 8, 3, 382-383.