

A typology of polytomously scored mathematics items disclosed by the Rasch model: implications for constructing a continuum of achievement

**John van Wyke**

**and**

**David Andrich**

**Murdoch University, Western Australia**

**Mailing address**

David Andrich  
Murdoch University  
Murdoch 6150  
Western Australia

**Acknowledgements**

The test materials and data on which this study is based were drawn from the 1992 and 1996 mathematics testing programs conducted by the Education Department of Western Australia as part of the Monitoring Standards in Education project. The work for the Report was supported in part by an Australian Research Council grant with the Australian National Ministerial Council on Employment, Education, Training and Youth Affairs (MCEETYA) Performance Measurement and Reporting Task Force, UNESCO's International Institute for Educational Planning (IEP), and the Australian Council for Educational Research (ACER) as Industry Partners\*. This paper has benefited from the many discussions over the years with many people on the implications of the reversals of Rasch thresholds in polytomous items.

\*Report No. 2 ARC Linkage Grant LP0454080: Maintaining Invariant Scales in State, National and International Level Assessments. D Andrich and G Luo Chief Investigators,  
Murdoch University

## A typology of polytomously scored mathematics items disclosed by the Rasch model: implications for constructing a continuum of achievement

### Abstract

The Department of Education in Western Australian conducts a number of testing programs for the purposes of demonstrating the improvement in the quality of student achievement. One of these programs assesses mathematics, and a scale of achievement across years 3, 5, 7 and 9 of schooling is constructed. The Rasch measurement model for dichotomous and polytomous items is used to establish such a scale. By application of the Rasch model to the items of this program, three types of polytomous items in which integer scores of  $0, 1, 2, \dots, m$  are assigned for increasing levels of performance on the item are identified. Within each type of item, two examples are provided. In one of these examples the Rasch model analysis confirms that the scoring rubric of the item is working as required; in the other the Rasch model analysis discloses that the scoring rubric of the item is not working as required. The Rasch model does this by showing that in the former case a higher score requires a greater ability to achieve than does a lower score, while that in the latter case a higher score does not require a greater ability to achieve than does a lower score. Two implications are considered from this classification and the identification of successful and unsuccessful scoring rubrics: first, implications for constructing an achievement continuum using polytomously scored items when the Rasch model discloses that the scoring rubric is not working as required; second, implications for improving the construction of rubrics so that they do work as required.

## A typology of polytomously scored mathematics items disclosed by the Rasch model: implications for constructing a continuum of achievement

### 1. Introduction

The Department of Education in Western Australian conducts a number of testing programs across state schools for the purposes of demonstrating and improving the quality of the government school system. One of these programs is called Monitoring Standards in Education (MSE). During the period when the data used in this Report were collected, a sample of students in their third, seventh and tenth years of schooling, drawn from schools across the state, were tested in two of the eight learning areas which made up the school curriculum. All learning areas were included on a cyclical basis, although *English* and *Mathematics*, with their links to the critical areas of literacy and numeracy, were the main focus of the testing program.

The testing program was designed to provide both descriptive and comparative information about the performance of the sampled students. Descriptive information is given in terms of the knowledge, skills and understandings described in the Western Australian standards framework, the *Student Outcome Statements*. Comparative information is also provided about their performance in order to monitor their improvement in educational outcomes in successive testing programs.

Provision of both descriptive and comparative kinds of information requires the construction of a suitable measurement scale. This permits the knowledge, skills and understandings required by the testing program to be mapped, as well as the performance of the students who have been tested, to be compared. The Rasch measurement model for dichotomous and polytomous items provides the means of establishing such a scale.

### 2. The Rasch Measurement Model

The Rasch simple logistic model for dichotomous items (SLM) given in Equation 2.1 specifies the way that the probability of success by a person  $n$  on an item  $i$  is governed by both the ability of the person ( $\beta_n$ ) and location of the item ( $\delta_i$ ). This relationship is shown in Figure 2.1, which shows the probability of a range of persons of varying ability responding correctly to a single test item.

$$P\{X_{ni} = 1\} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}. \quad (2.1)$$

As ability varies, the probability of a correct response to the item also varies. The probability that a person with low ability will respond correctly is correspondingly low, approaching 0 asymptotically as ability decreases. Symmetrically, the probability that a person with high ability will respond correctly is correspondingly high, and approaches 1 asymptotically as ability increases.

Figure 2.1 also shows the probability of responding incorrectly, where:

$$P\{X_{ni} = 0\} = 1 - \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} = \frac{1}{1 + e^{(\beta_n - \delta_i)}} \quad (2.2)$$

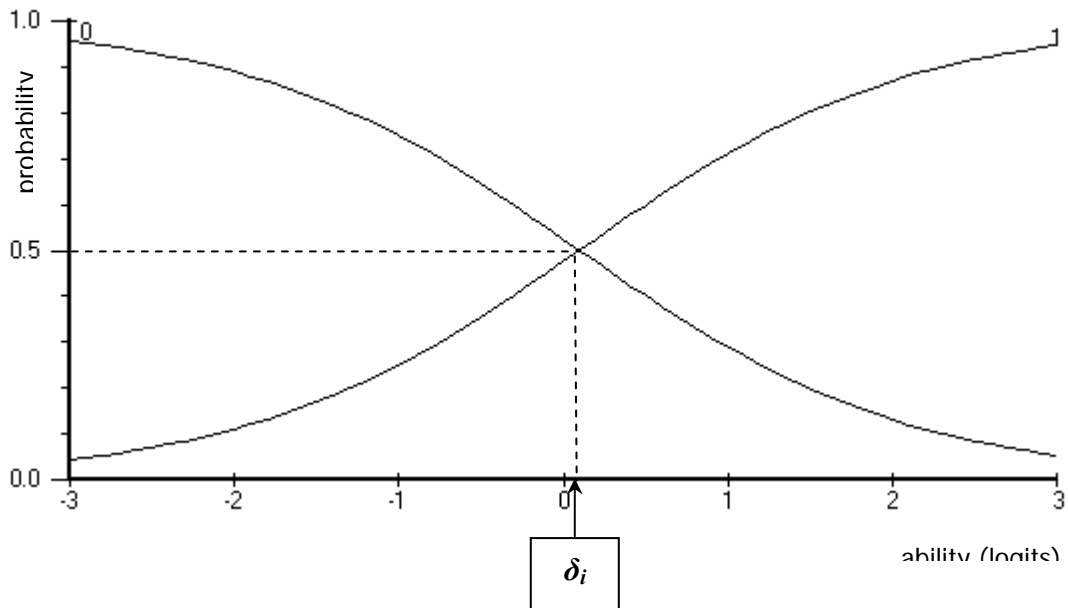


Figure 2.1: Category probability curve showing the probabilities of scores of 0 and 1 on a single item as a function of ability

A key feature of the SLM is that the difficulty of items can be located on the same scale as the ability of the persons attempting those items. This simultaneous scaling of students and items is achieved by locating the difficulty of the item at the point where a person has a 0.50 probability of getting that item right. The difference between item difficulty and person ability,  $\beta_n - \delta_i$ , is expressed in logits: it is the logarithm of the odds that a person  $n$  will get item  $i$  right in

$$\text{Ln} \frac{P\{X_{ni} = 1\}}{P\{X_{ni} = 0\}} = \text{Ln}(e^{(\beta_n - \delta_i)}) = \beta_n - \delta_i \quad (2.3)$$

The location of the item is identified as the point on the ability scale where the probability curves of 0 and 1 intersect. At this point, the probability of a response of either 0 or 1 is equally likely. Because it is a dichotomous item, there is of course a probability of 0.50 of either response. The probability of a correct response decreases as ability decreases, and increases as ability increases, around this point. The item shown in Figure 2.1, for example, has a location of approximately 0.12 logits.

With the single item located on the scale as shown in Figure 2.1, descriptions about the ability of the range of persons shown is obviously confined to their facility in relation to this item. Greater precision and more detail in describing the ability of persons can be achieved by adding more items of the same kind. Figure 2.2 shows how the scale can be built up by locating a number of dichotomous items on the scale.

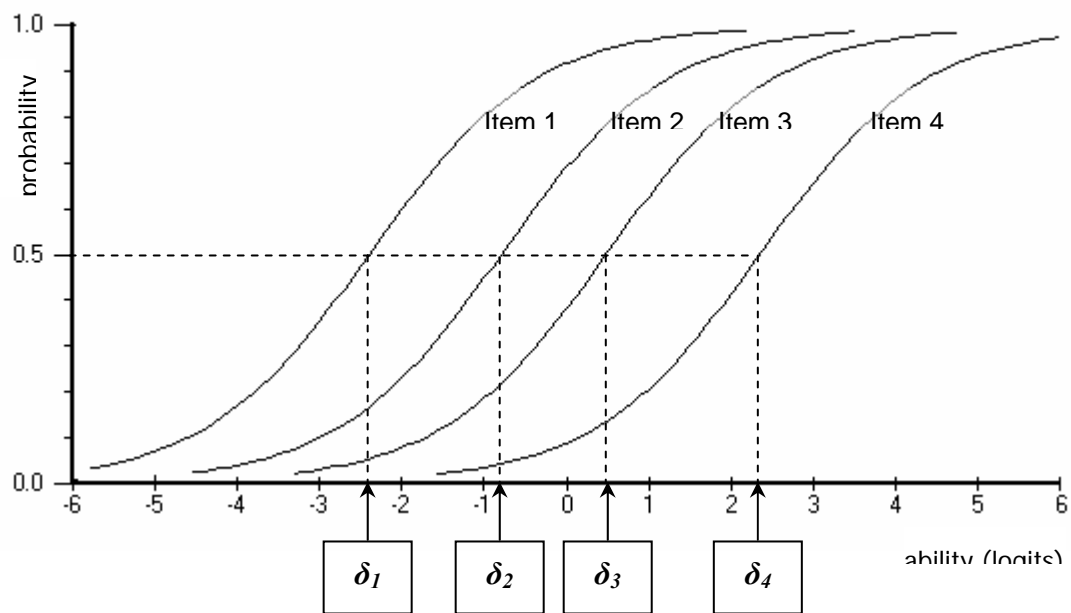


Figure 2.2: Item characteristic curves for four items with increasing difficulties

Figure 2.2 shows the ICC's for four dichotomous items with locations of -2.42, -0.85, 0.50 and 2.37 logits respectively. Their order and location on the scale reflect their increasing difficulty.

## 2.1 Scale building using the Rasch SLM

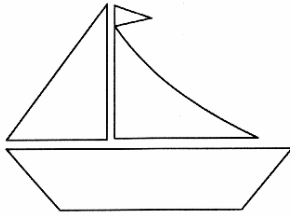
The building up of a scale in terms of dichotomous items is elaborated in some detail in this section. It is used later in studying successful and unsuccessful polytomously scored items.

In practice, the construction of the scale does not explicitly include the ICC's of the items as shown in Figure 2.2. Although they provide the essential theoretical underpinning to the construction of the scale, the purpose of the scale goes beyond a statistical interpretation, to provide a means of describing and interpreting the test performance in terms of the construct which is being measured.

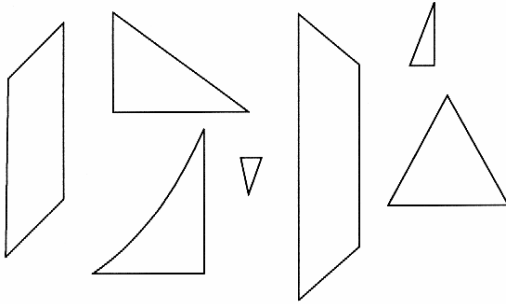
Because the description of performance is referenced directly to the items making up the test, it is common practice to refer to the scale as an achievement scale. This reflects the fact that it is made up of observable behaviours - the students' achievements on the test. The underlying ability scale then takes its meaning from the descriptions and locations of these achievements.

The process of establishing the achievement scale may involve several stages. Initially, at the item construction stage, there is an intention to construct items with a range of difficulties. Then the relative difficulty of each item is estimated from the data. Each item is then located on the scale. Following this, a description of the knowledge and skills addressed by each item is referenced to the item location, so that a profile of the performance demands of the test can be developed. Taken together, the tasks which are described in the profile reveal the progression in achievement measured by the test, from items which are relatively less difficult to those which are relatively more difficult. In this way, it is possible to interpret test results in terms of the knowledge, skills and understandings that are demonstrated on the test, rather than simply in terms of numerical test scores. The process has been reviewed here because it becomes important in applying the same process when polytomously scored items become involved. Often, this process is ignored.

To illustrate the process of constructing the performance scale, four dichotomous items are shown in Figures 3.5 – 3.8. Each of these items, drawn from the *Space* strand of the 1996 MSE mathematics testing program, required students to respond to a question designed to assess their understanding of a single spatial concept. Responses were judged to be either right or wrong, and scored 1 or 0 accordingly.



Put a tick in the shapes you need to make this picture of a boat.



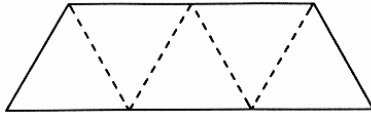
### Marking key

1 ..... tick on all four correct shapes

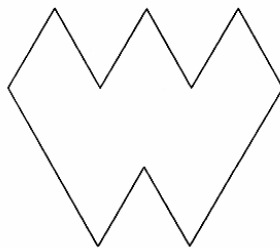
0 ..... incorrect selection

Figure 2.3: Item 1

This shape is made from 5 smaller shapes all the same as each other.



Draw in dotted lines to show how this shape is also made from 5 smaller shapes all the same as each other.



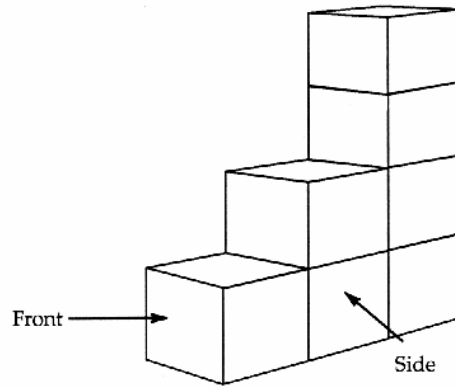
### Marking key

1 ..... shape divided into 5 rhombuses

0 ..... incorrect response

Figure 2.4: Item 2

This drawing shows a shape made of large cubes.



Use this grid to show what the shape looks like from the front.  
Use 1cm for the side of a cube.



### Marking key

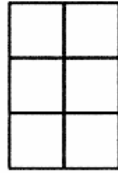
1 ..... front view correct

0 ..... incorrect drawing

Figure 2.5: Item 3



These six squares mark the bottom of a rectangular box to hold 18 wooden one centimetre cubes. The box has no top. Draw the net (or pattern) for the box.



Marking key

1 ..... correct net drawn

0 ..... incorrect drawing

Figure 2.6: Item 4

Figure 2.7 shows each item located on the scale according to its difficulty. Item 1 is located at -2.42 logits; item 2 is located at -0.85 logits; item 3 is located at 0.50 logits; item 4 is located at 2.37 logits. On the right of the scale is a brief description of the knowledge or skill demands of each item, mapped to the item's location. These descriptions chart the progress of conceptual development in the *Space* strand from the least difficult item in the test (Item 1) through to the most difficult item (Item 4).

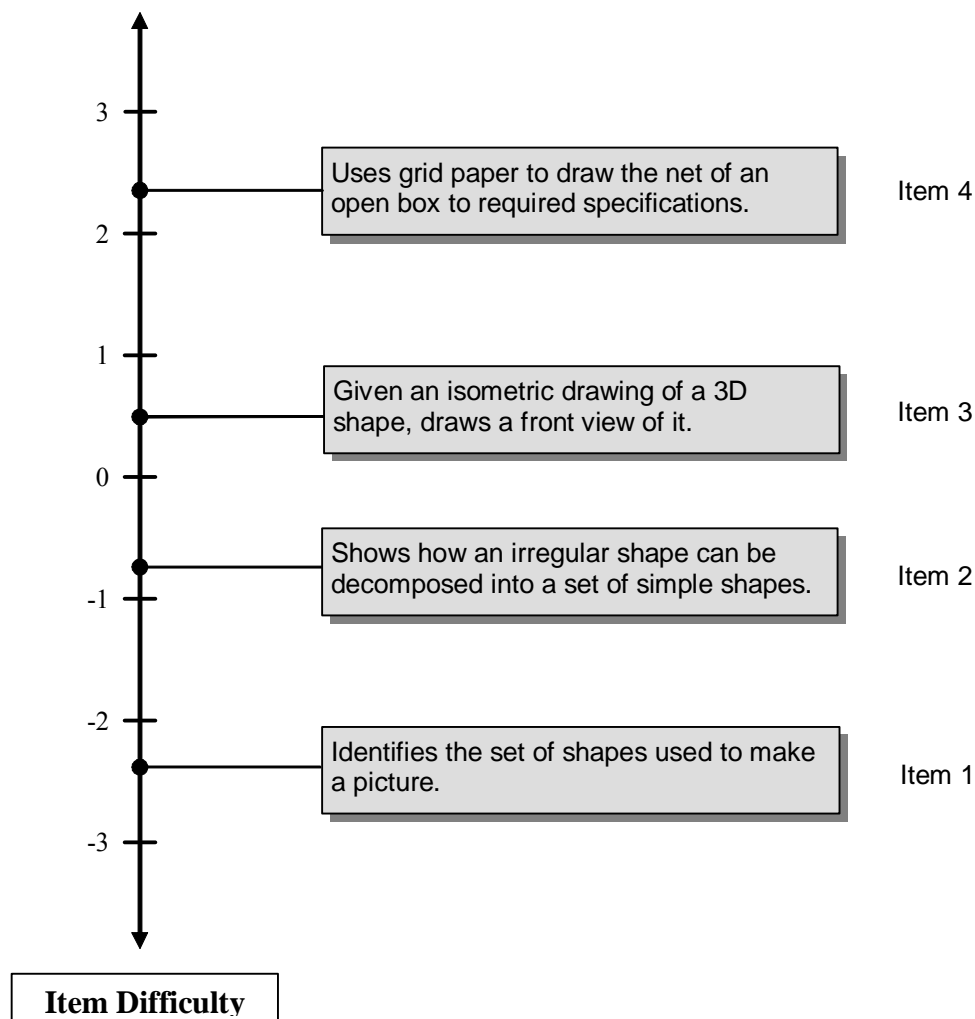


Figure 2.7: Construction of the achievement scale using Items 1 - 4

A further stage may also be added to the scale-building process, although this is not necessary, and is not always included. In this stage, the knowledge and skill demands of the items located along the scale are analysed and then grouped together according to features which they share in common. These groupings can then be used to define bands or levels of achievement along the scale at a higher order of scale and to describe them using more general descriptors, which go beyond simply aggregating the individual item descriptors in each band.

## 2.2 The Rasch model for polytomously scored items

The SLM was developed by Rasch for the analysis of dichotomously scored test items. Many testing programs, however, require greater precision or more information than a simple right/wrong scoring system allows from any particular item. In these cases, polytomously scored items with several levels of performance may be required.

Rasch's (1961) formulation of the model for polytomously scored items is an extension of the SLM. Instead of dealing with dichotomous items with two response categories and possible scores of 0 and 1 only, it provides a model for test items with more than two response categories, with possible scores of 0, 1, ...,  $m$ .

The present familiar form of the model was derived in two related papers by Andersen (1977) and Andrich (1978). Andersen showed that for sufficient statistics,  $\phi_{x+1} - \phi_x = \phi_x - \phi_{x-1}$ . Andrich then resolved the  $\kappa_x$  and  $\phi_x$  in terms of thresholds  $\tau_x$ ,  $x = 1, \dots, m$  on the continuum with discrimination  $\alpha_x$ ,  $x = 1, \dots, m$  at these thresholds.

Andrich's version gives the probability of a person of ability  $\beta_n$  being classified in category  $x$  in a test item of difficulty  $\delta_i$ , with  $m+1$  ordered categories as:

$$\Pr\{X_{ni} = x\} = \frac{e^{(x(\beta_n - \delta_i) - \sum_{k=1}^x \tau_k)}}{\sum_{x=0}^m e^{(x(\beta_n - \delta_i) - \sum_{k=1}^x \tau_k)}}. \quad (2.4)$$

where  $x \in \{1, 2, \dots, m\}$ .

In the dichotomous case, where  $m = 1$ , and  $\tau_1 \equiv 0$ , Equation 2.4 specialises to:

$$\Pr\{X_{ni} = x\} = \frac{e^{(x(\beta_n - \delta_i))}}{1 + e^{(x(\beta_n - \delta_i))}} \quad (2.5)$$

which is simply the Rasch SLM.

In Equation 2.4 the threshold parameters are not subscripted by  $i$ , indicating they are assumed identical across items. If thresholds are different across items, the model takes the form

$$\Pr\{X_{ni} = x\} = \frac{e^{x(\beta_n - \delta_i) - \sum_{k=1}^x \tau_{ki}}}{\sum_{x=0}^{mi} e^{x(\beta_n - \delta_i) - \sum_{k=1}^x \tau_{ki}}}.$$

(2.6)

(Wright and Masters 1982; Masters 1982; Masters and Wright, 1997)

The model of Equation 2.5 has become known as the ratings scale model and the model of Equation 2.6 has become known as the partial credit model. However, at the level of response of a person to an item, the models are identical in structure and response process (Andrich Andrich, 2002; 2005; Luo, 2005). Here we refer to the model simply as the Polytomous Rasch Model (PRM).

### 2.3 Scale building using thresholds

Figure 2.8 shows threshold characteristic curves for a polytomous item with two thresholds, located at -1.22 logits and 0.23 logits. These curves model the probability of a successful response at each of the thresholds, conditional on the response being in one of the adjacent categories. The successful response in each adjacent pair is the one considered to characterise the category which reflects greater ability, e.g.  $x$  rather than  $x-1$ , given that the response is either  $x$  or  $x-1$ . This is modelled by the SLM in which:

$$\begin{aligned} \Pr\{X_{ni} = x | X_{ni} = x-1 \text{ or } X_{ni} = x\} &= \frac{\Pr\{x\}}{\Pr\{x-1\} + \Pr\{x\}} \\ &= \frac{\exp(\beta_n - (\delta_i + \tau_{xi}))}{1 + \exp(\beta_n - (\delta_i + \tau_{xi}))}. \end{aligned} \quad (2.7)$$

Equation 2.7 is readily derived from Equation 2.4 and is identical to the form of the SLM of Equation 2.1 for a successful dichotomous response, except that the threshold parameter  $\tau_{xi}$  included to qualify the location parameter  $\delta_i$ . The reverse process also holds – Equation 2.4 follows from 2.7 (Luo, 2005).

Thus the respective curves reflect the probability of a score of 1 rather than 0 at the first threshold and the probability of a score of 2 rather than 1 at the second threshold. For persons

with an estimated ability less than -1.22 logits, there is less than 0.50 probability that they will succeed at the first threshold and so their most likely score on this item is 0. For persons with an estimated ability greater than -1.22 logits but less than 0.23 logits there is a greater than 0.50 probability that they will succeed at the first threshold and a less than 0.50 probability that they will succeed at the second threshold. Their most likely score is 1. For persons with an estimated ability greater than 0.23 logits, there is greater than 0.50 probability that they will succeed at the second threshold, and so their most likely score on this item is 2. In the same way that dichotomous items are located at the point where there is a 0.50 probability of a correct response to the item, the thresholds of polytomous items are located at points where the probability of a successful response at the threshold is exactly 0.50.

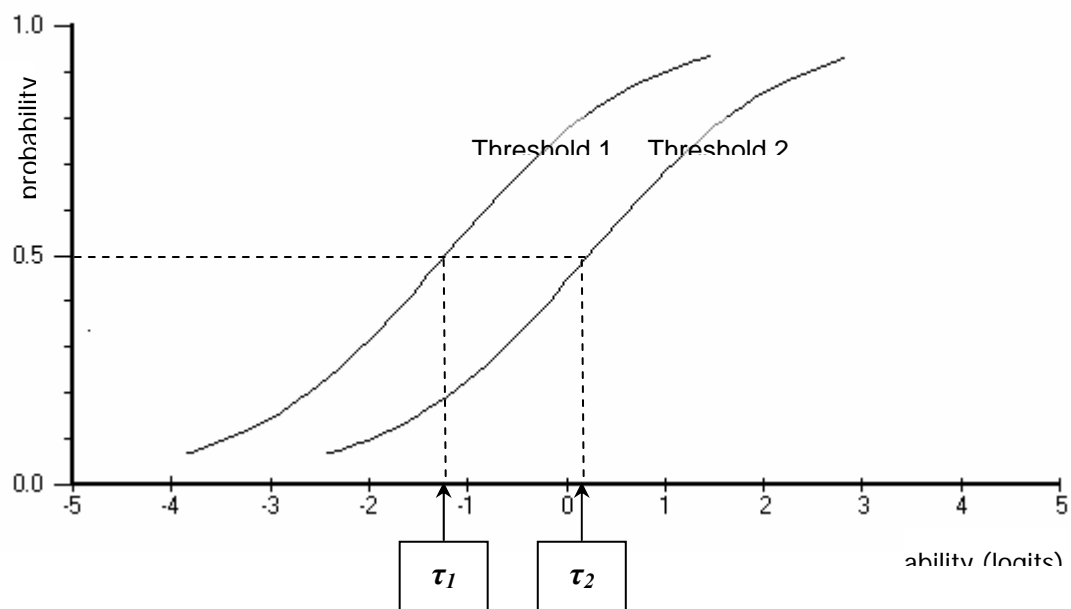


Figure 2.8: Thresholds characteristic curves for a polytomous item with two thresholds

Although the threshold curves have been shown in Figure 2.8 to characterise two dichotomous responses, in the model and in the data, these responses are latent – at the manifest level there is only one response in one of the three categories, not two responses.

The probability of a response in each of the three categories is shown in Figure 2.9, which shows, for a range of persons, the probability of responding in each of the three categories defined by the two thresholds. The three probabilities at any point on the continuum sum to 1.

Figure 2.9 models the probability of a score of 0, 1 or 2 for a typical polytomous item with three categories. As ability increases, the probability of a score of 0 decreases, as in the dichotomous case. In addition, as ability increases, the probability of the maximum score of 2

increases. Between these two curves is the curve which shows the probability of a score of 1. This curve shows that when a person has very low ability relative to the item's location, then the probability of a response of 0 is most likely; when a person is of moderate ability relative to the item's location, then the most likely score is a 1 and when a person has an ability much greater than the item's location, then the most likely response is 2.

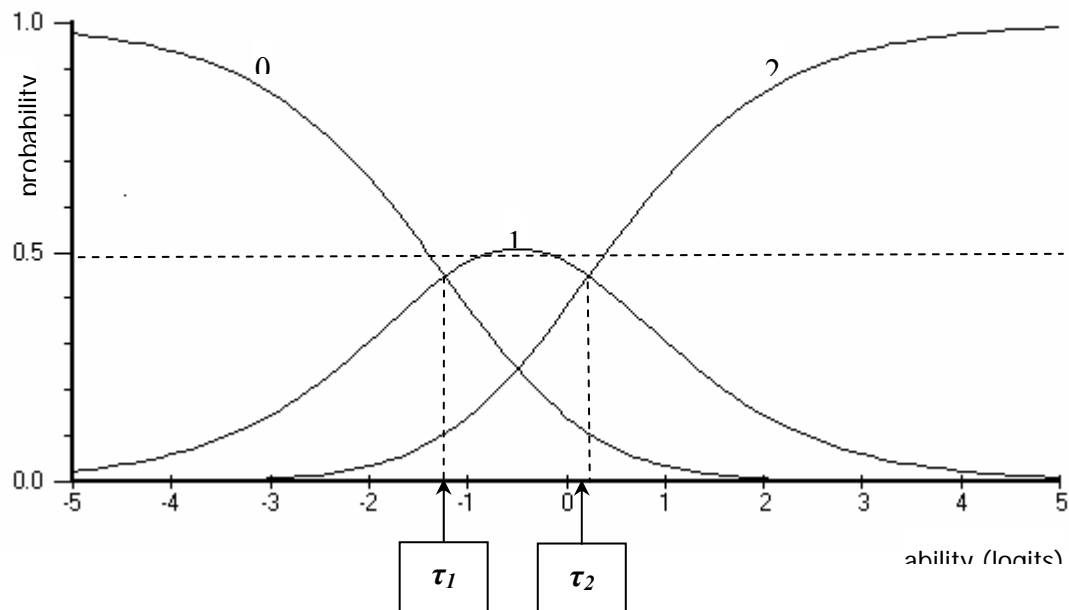
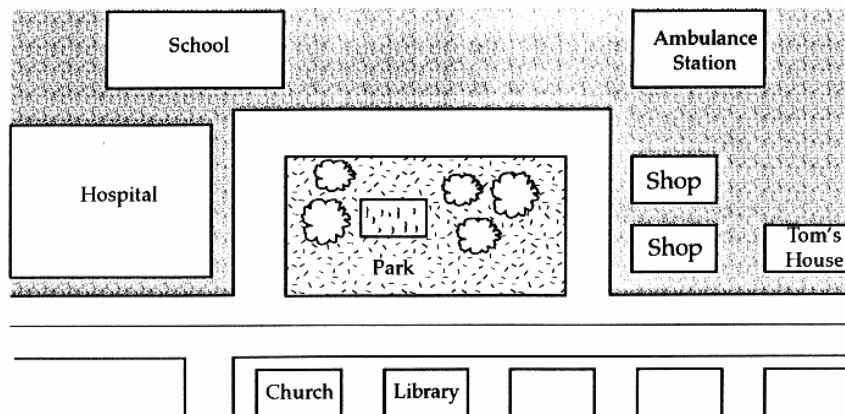


Figure 2.9: Category probability curves for a polytomous item with two thresholds

In Figure 2.9, the thresholds, and the categories they define, are naturally ordered in the sense that the threshold defining the two higher categories of achievement is of greater difficulty than the threshold defining the two lower categories of achievement. The first threshold, which represents the point where a score of 1 becomes more likely than a score of 0, is approximately -1.22 logits. The second threshold, where a score of 2 becomes more likely than a score of 1, is approximately 0.23 logits. In other words, progressively more ability is required to score a 0, 1 or 2 respectively on this item.



Amy is visiting Tom's house. She wants to walk to the school. Give her directions to get from Tom's house to the school.

---



---



---



---

### Marking key

2 ..... one mark for correct use directional language *and* one mark for correct use of landmarks

1 ..... one mark for *either* correct use of directional language *or* correct use of landmarks

0 ..... inadequate response

Figure 2.10: Item 5

In the same way that the achievement scale is built up by locating and describing individual dichotomous items on the ability continuum, polytomous items can also be located and described in terms of their thresholds. To illustrate this, a typical polytomous item is shown in Figure 2.10. Item 5, drawn from the Space strand of the 1996 MSE mathematics testing program, was designed to assess students' ability to give directions using a simple map. As the marking key shows, responses were judged to be at three levels – correct, partially correct, and incorrect - and scored 2, 1 or 0 accordingly.

Figure 2.8 shows the thresholds dividing the response categories 0, 1 and 1, 2 located at -1.22 logits and 0.23 logits respectively. In Figure 2.11 these thresholds have been added to the achievement scale in a manner which is analogous to the mapping of the dichotomous items in Figure 2.7. In the same way that the knowledge or skill demand for success on each dichotomous item is mapped to the item's location, that is, at the point where the probability of success is 0.50, the knowledge or skill demands for success at each level of a polytomous item are mapped to the corresponding threshold location.

In this way, the addition of a single polytomous item, making the same demands on test time and space as a dichotomous item, and still requiring just a single response from the student, has contributed two extra points to the achievement scale. Polytomous items ranging from three to five categories are commonly used in testing programs of this type and offer advantages of economy, efficiency and precision in the process of building up the description of the underlying ability continuum.



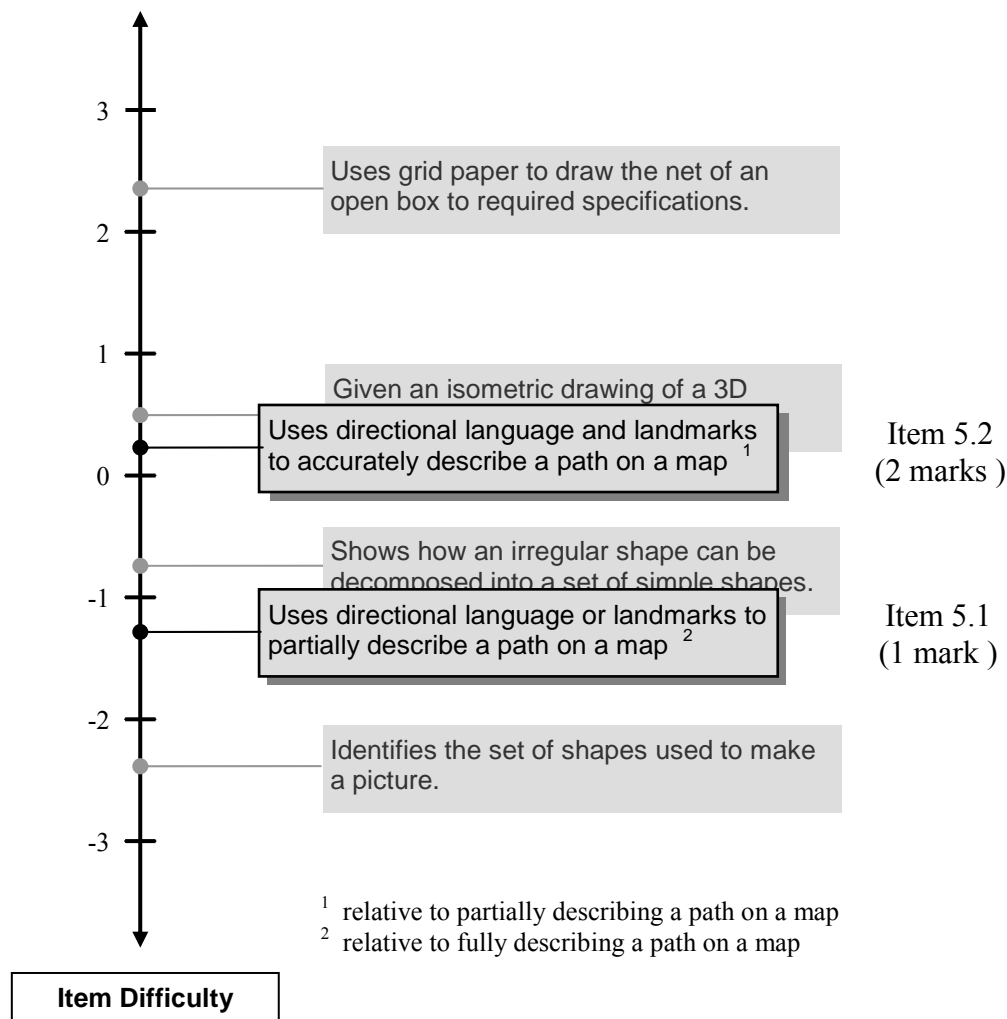


Figure 2.11: Addition of Item 5 to the achievement scale.

## 2.4 Ordered thresholds

Figure 2.9 shows thresholds which are naturally ordered. The first threshold is located at -1.22 logits, the second at 0.23 logits. Persons with ability estimates less than -1.22 logits are most likely to fail the first threshold, and so score 0 on this item. Persons with ability estimates in the range -1.22 logits to 0.23 logits are most likely to pass the first threshold but fail the second, and so score 1. Persons with ability estimates greater than 0.23 logits are most likely to pass the first and the second thresholds, and so score 2. With increasing ability, the probability of exceeding the first threshold and then the second threshold also increases. Conversely, the greater the score, the greater the latent ability that is implied by the model. This is the underlying principle of the ordered categories and therefore the thresholds that are boundaries between the categories reflect an increasing amount of the attribute being

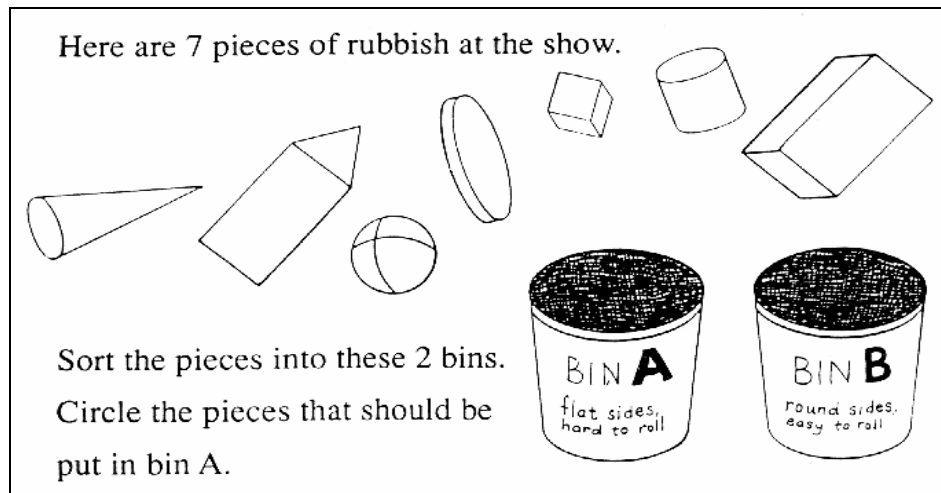
measured (Andrich, 2002).

The intent of Item 5 is to provoke three levels of response in using a simple map. The ordering of the thresholds revealed by the Rasch Model shows the scoring categories working in accordance with the idea that thresholds between higher level categories are more difficult than thresholds between lower order categories. To illustrate this, consider two persons A and B attempting Item 5, where person A has a 0.5 probability of scoring 0 or 1, and person B has a 0.5 probability of scoring 1 or 2. That is,  $\beta_A = \tau_1$ , and  $\beta_B = \tau_2$ . Because the thresholds are in their proper order, Person B is understood to have a greater location, or more ability, than Person A. In terms of their competence in relation to the demands of Item 5, this means that a person who has a 0.5 probability of scoring a 1 or a 2 – being at least able to give partially correct directions, or being able to give fully correct directions – requires more ability than a person who has a 0.5 probability of scoring a 0 or 1 – being at most able to give partially correct directions or being unable to give directions at all.

As the location of the thresholds provides the means of placing the concomitant descriptions on the achievement scale, the order of the thresholds ensures that the placement makes sense. The scale now indicates that an increasing amount of the attribute is required in order to score a 1 rather than a 0, by being able to give at least partially correct directions, and to score a 2 rather than a 1 by being able to give fully correct directions. This is sensible and expected.

## 2.5 Disordered thresholds

It is not always the case, however, that thresholds are properly ordered in the data. Consider Item 6, shown in Figure 2.12, which is also taken from the *Space* strand of the 1996 MSE mathematics testing program. This item was designed to assess students' ability to use functional criteria to classify 3D shapes.



### Marking key

2 ..... student is able to identify all three pieces correctly

1 ..... student is only able to identify one or two pieces correctly

0 ..... student is unable to identify any pieces correctly

Figure 2.12: Item 6

Figure 2.13 shows the empirically derived category probability curves for Item 6. It shows that the location of the first threshold - the intersection of the curves of probability of 0 and probability of 1 respectively - has a difficulty of 0.74 logits. It also shows that the location of the second threshold - the intersection of the curves of probability of 1 and probability of 2 respectively - has a difficulty of -1.63 logits. The problem here is that the location of the first threshold is greater than the location of the second threshold. These thresholds are reversed. Figure 2.13 clearly shows that this is due to the failure of the middle category, corresponding to a score of 1, to function properly - at no point on the graph is a score of 1 most likely. Although persons with low ability relative to the item's difficulty are still most likely to respond incorrectly and score 0, and persons with high ability relative to the items difficulty are still most likely to respond correctly and score 2, persons with ability in the range -1.63 logits to 0.74 logits, where a score of 1 should be most likely, are still more likely to score both 0 and 2.

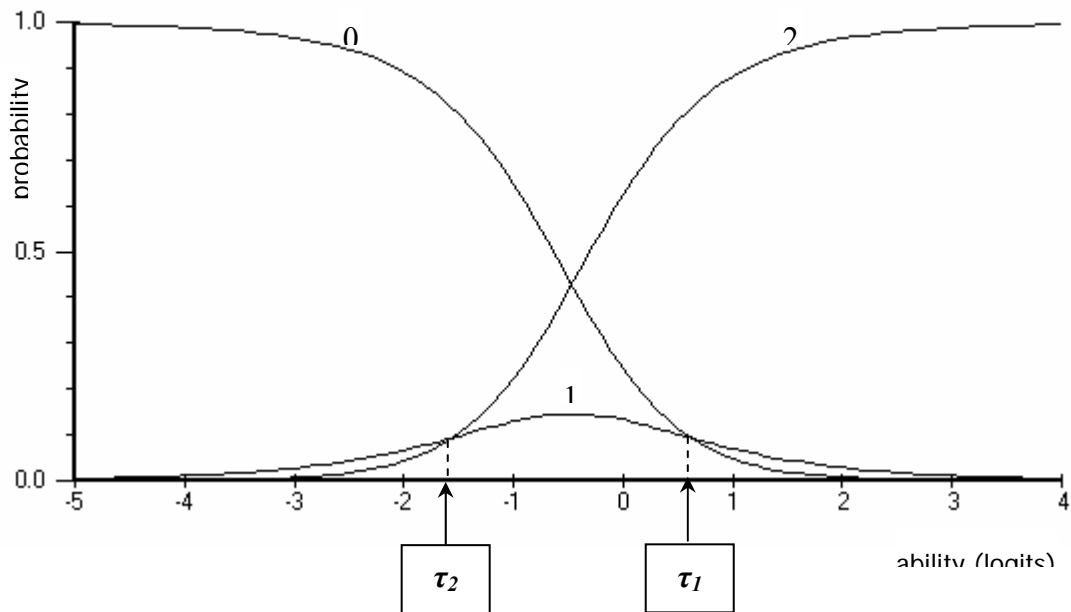


Figure 2.13: Category probability curves for Item 6

The reversed thresholds can be seen more clearly in Figure 2.14 which shows the latent dichotomous threshold characteristic curves for the item. These curves, as discussed earlier, model the probability of an implied or latent successful response at each of the thresholds. Threshold 1, which divides the less difficult response categories 0 and 1, is located at 0.74 logits on the ability continuum. This is greater than threshold 2, however, which divides the more difficult response categories 1 and 2, but which is located at -1.63 logits.

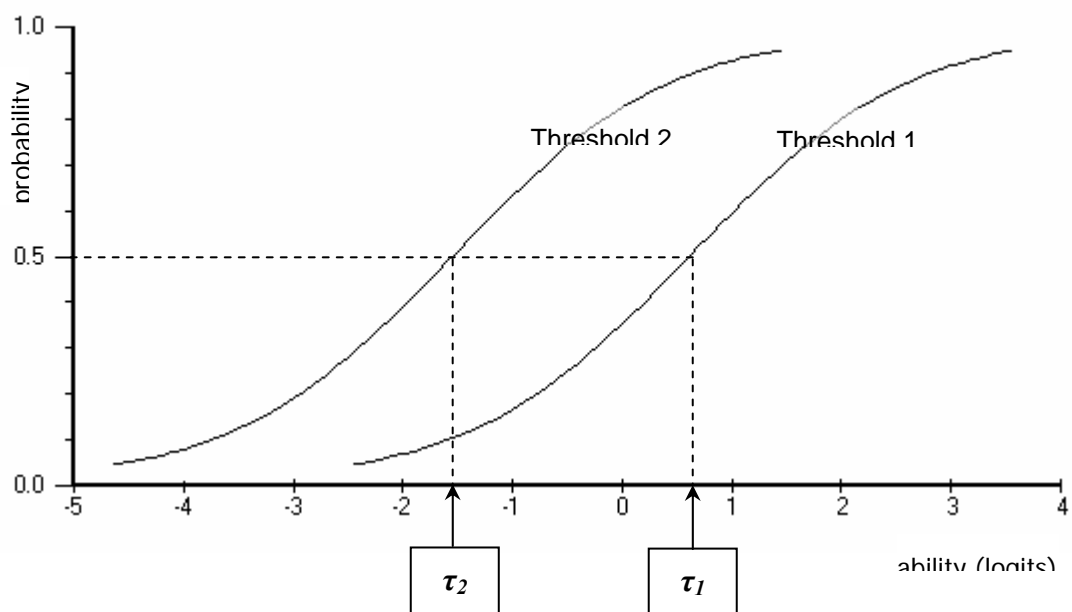


Figure 2.14: Threshold characteristic curves for Item 6

The intent of Item 6 is to identify three levels of response in classifying shapes. Because of the disordering of the thresholds revealed by the Rasch Model however, the scoring categories are not working in accordance with the idea that thresholds between higher level categories are more difficult than thresholds between lower order categories.

To illustrate this, again consider two persons A and B, this time attempting Item 6, where person A has a 0.5 probability of scoring 0 or 1, and person B has a 0.5 probability of scoring 1 or 2. That is,  $\beta_A = \tau_1$ , and  $\beta_B = \tau_2$ . Because the thresholds are in reverse order, Person A appears to have a greater location, or more ability, than Person B. In terms of their competence in relation to the demands of Item 6, this means that a person who has a 0.5 probability of scoring a 1 or a 2 – being able to identify some correct shapes, or being able to identify all correct shapes – requires *less* ability than a person who has a 0.5 probability of scoring a 0 or 1 – being able to identify none or only some of the correct shapes.

The difficulty with reversed thresholds is revealed even more clearly when trying to locate them, and their concomitant descriptions, on the achievement scale. Placed according to the reversed threshold locations, the scale indicates that a greater amount of the attribute is required in order to score a 1 rather than a 0, by being able to identify one or two correct shapes, and a lesser amount is required in order to score a 2 rather than a 1, by identifying all the correct shapes. This representation is clearly inconsistent with the intended ordering of the categories.

## 2.6 Implications of disordered thresholds for scale building

An attempt to represent this on the achievement scale would show a partial understanding (one or two shapes identified), scored 1, located at 0.74 logits. The response showing a full understanding (all three shapes identified), scored 2, would be located below rather than above this, at -1.63 logits.

Figure 2.15 shows the location of the reversed thresholds on the achievement scale. The latent dichotomous response which should be more difficult now appears to be easier than the latent dichotomous response which should be less difficult. Locating the thresholds of Item 6 on the achievement scale in this way shows that a partial understanding of classifying shapes indicates more ability than a full understanding of classifying shapes. This contradicts the underlying principle of the Rasch Model, and of the scoring system of the item, that the

location of the thresholds reflects an increasing amount of the attribute being measured.

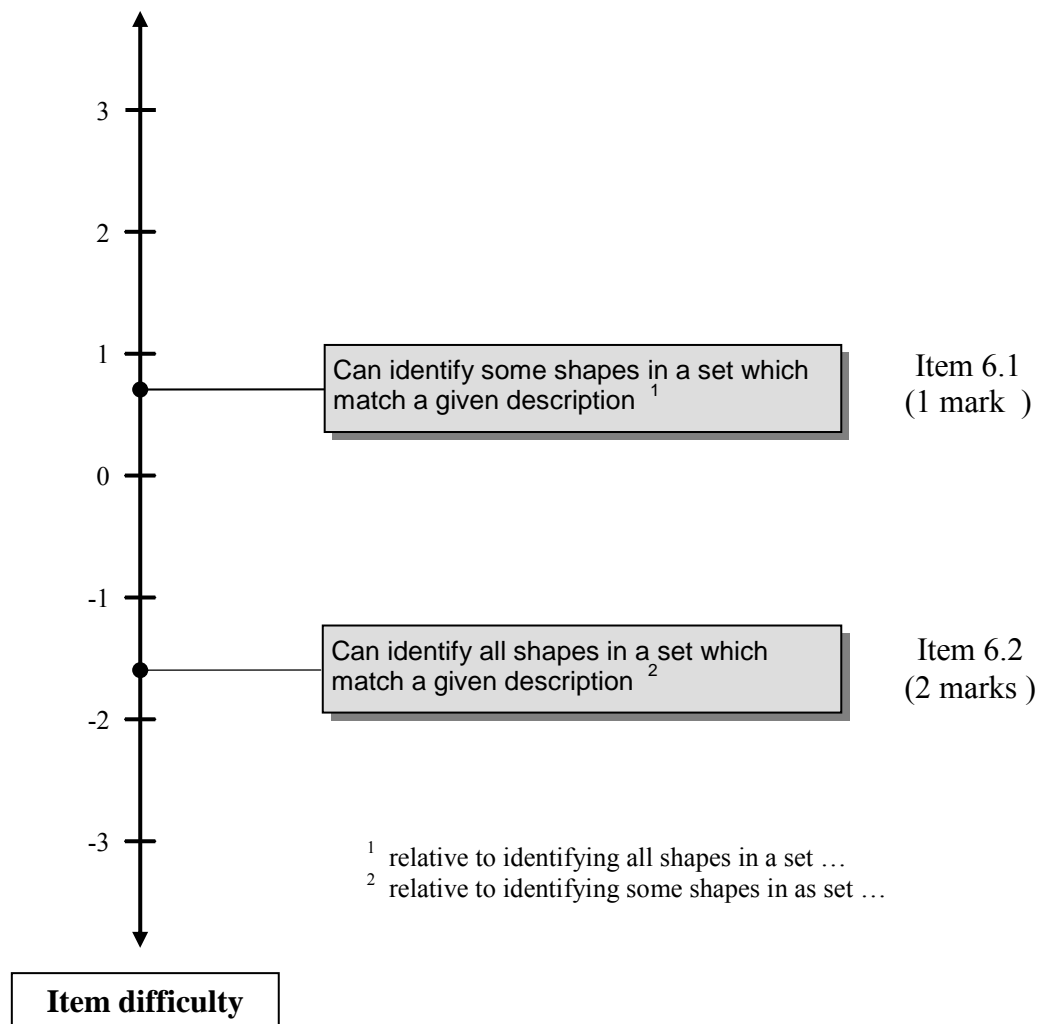


Figure 2.15: Locating Item 6 with disordered Rasch thresholds on the achievement scale

Clearly, it is not possible to use these thresholds to locate points on the achievement scale as indicators of the underlying ability continuum, in the way that was done with Item 5. A different treatment is required to resolve this problem.

In order to understand this problem more fully, the next section proposes a typology for polytomously scored items, and then examines a number of items with ordered and disordered thresholds within this typology in more detail.

The analysis is in two parts. Firstly, a general typology for polytomously scored mathematics, based on the way they are scored, items is proposed. Following this, a number of items - with ordered and disordered thresholds - are examined within the framework of this

typology in order to gain a greater understanding of the significance of disordered thresholds as an indication of item functioning, and the consequences for retaining or rejecting such items.

### 3. TYPOLOGY OF POLYTOMOUS ITEMS

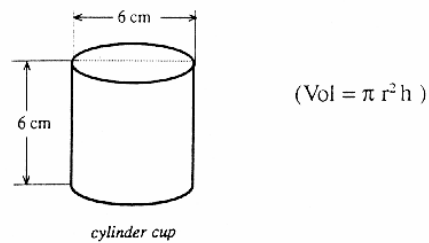
For organisational purposes, all polytomous items used in these testing programs, together with their marking keys, can be usefully characterised as one of three types. Depending on the way it is scored, each type can be categorised as follows:

1. Hierarchically scored items
2. Incrementally scored items
3. Decrementally scored items

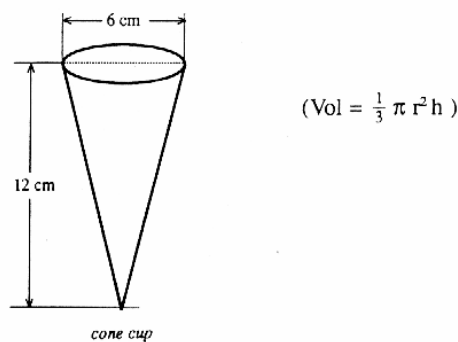
### 3.1 HIERARCHICALLY SCORED ITEMS

Item M029, shown in Figure 3.1, is an example of an hierarchically scored item.

At the kiosk peanuts are sold in cylinder-shaped cups.



A salesperson is trying to encourage the kiosk operator to change the shape of the cup. He shows the kiosk operator a cone-shaped cup.



Which one of the above two cups would hold more peanuts?

Explain your answer. \_\_\_\_\_

#### Marking Key

3 ..... cylinder: shows volume of cones is less than cylinder

eg: cone is  $113 \text{ cm}^3$  and cylinder is  $170 \text{ cm}^3$

2 ..... cylinder: because it holds more or has greater capacity; but does not show calculation of the volumes

1 ..... cylinder, with inadequate or no explanation

0 ..... inadequate response

Figure 3.1: Item M029

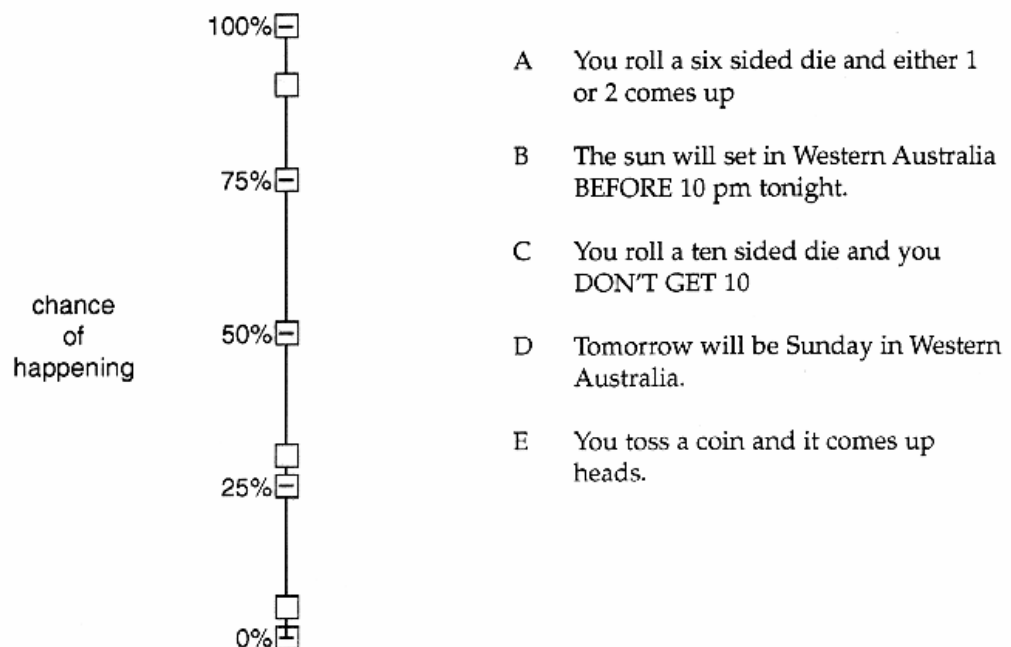


Here the marking key identifies a hierarchy of responses, and provides a description of the response at each level in the hierarchy. Scoring is based on comparing the student's work with the range of descriptions and allocating a score according to the level in the hierarchy that the response matches. Scores can range from 0 for an inadequate response through to full marks for the highest level of response, with one or more levels of partially correct response between these responses.

### 3.2 INCREMENTALLY SCORED ITEMS

Item C155, shown in Figure 3.2, is an example of an incrementally scored item.

Think about how likely it is that the following things will happen:



Match the letters to a box on a number line to show about what percentage chance you think each event has of happening.

#### Marking Key

5/4/3/2/1 .....one mark for each correct matching

0 .....no correct matchings

Figure 3.2: Item C155

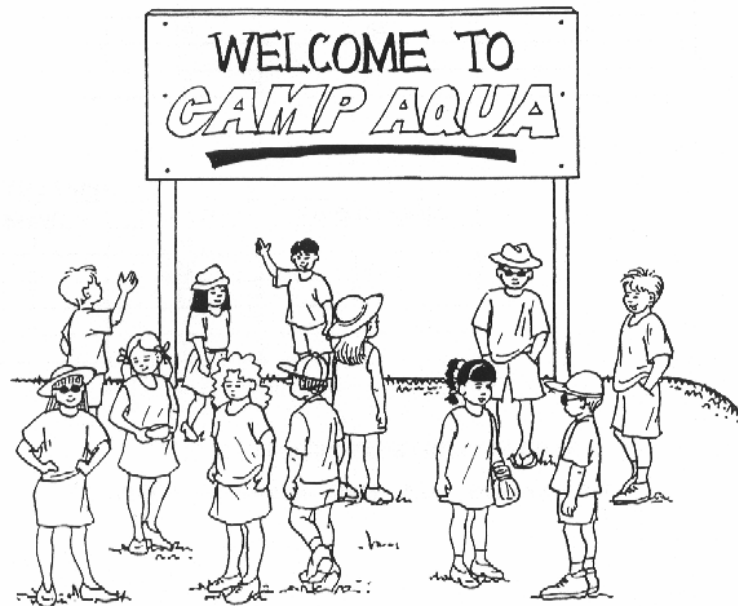
Here the marking key specifies a number of elements which go together to make up a correct response. The level of response is determined by the number of elements that the student has correct. Scoring is based on incrementing the student's score for each element that the student has correct. Depending on the number of elements, there can be any number of levels of response. Typically, there are three to five levels of response.

### 3.3 DECREMENTALLY SCORED ITEMS

Item N021, shown in Figure 3.3, is example of a decrementally scored item.

At Camp Aqua children meet at the camp entrance to start the day.

In the picture below you can see what 12 children are wearing today.



What percentage of the group is wearing both a hat and sunglasses?

#### Marking Key

2..... 25%

1..... fractional or approximate answer

eg:  $\frac{3}{12}$ ,  $\frac{1}{4}$ , 24% - 26%

0..... incorrect answer

Figure 3.3: Item N021

Here the marking key specifies the correct response together with a range of partially correct responses, for which marks are deducted. Typically, marks are deducted because the student's response, although substantially correct, contains computational errors, is inaccurate, or is incomplete. Scoring is based on decrementing the student's score for each of these omissions. Although there can be any number of levels of response, there are usually only three.

#### 4. ANALYSIS OF ITEM FUNCTIONING

Within each of these types there are examples of items which reveal ordered thresholds as well as examples of items which reveal disordered thresholds. It has been argued in this report that ordered Rasch thresholds provide confirming evidence that the ordering of categories is functioning as intended. This section shows the advantages in studying items with reversed Rasch thresholds from the perspective of the content of the items.

To understand the impact that the methodological procedures associated with these views have on the interpretation of the achievement scale, two examples of each type of item, one with ordered thresholds and one with disordered thresholds, are discussed in the following section. To support the discussion, the marking key and Category Probability Curves associated with each item have also been included.

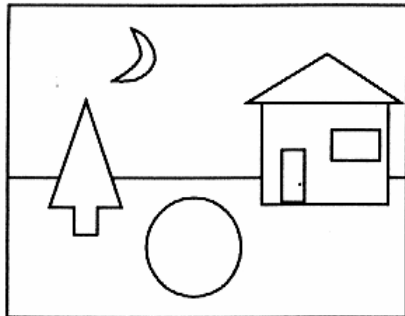
##### 4.1 HIERARCHICALLY SCORED ITEMS

Items S121 and S004, taken from the *Space* strands of the 1992 and 1996 testing programs, are examples of hierarchically scored items. These items, shown in Figures 4.1 and 4.2, have a similar structure, in terms of both the task demand, and the format of the marking key.

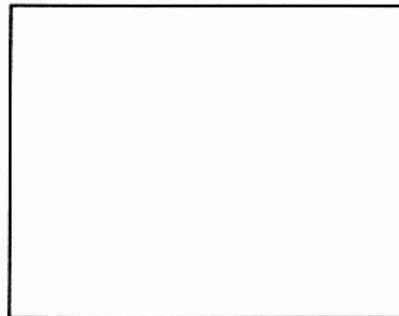
## Item S121

. Copy the picture in Box A into Box B.

A



B



## Marking Key

- 2 ..... all parts recognisable in shape, size, position, and orientation  
 1 ..... most parts recognisable in shape, size, position, and orientation  
 0 ..... inadequate response

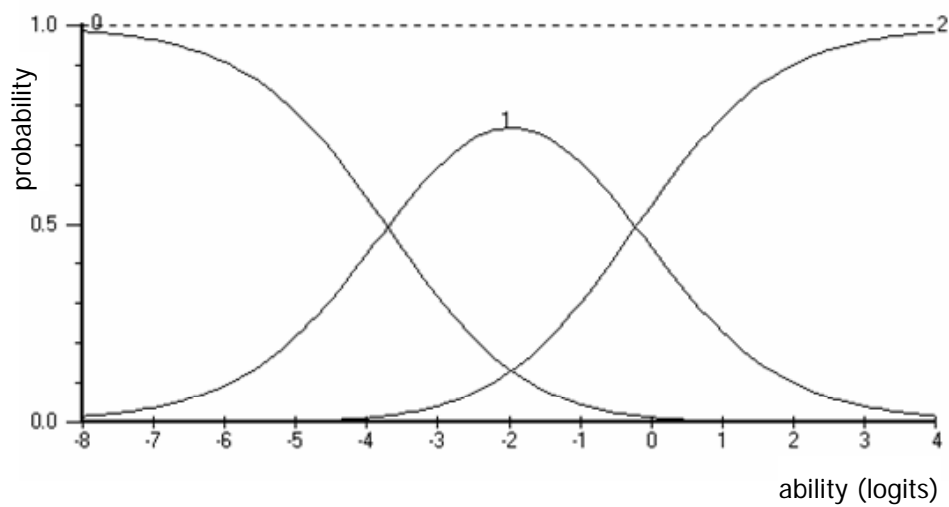
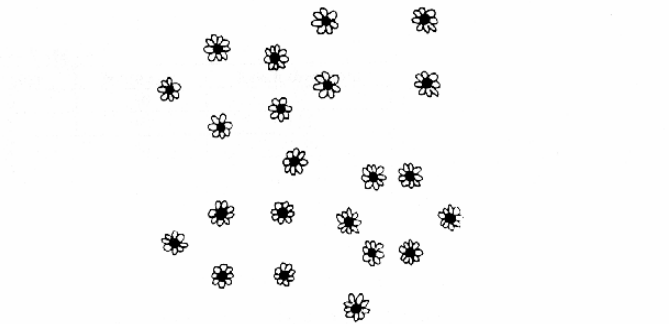


Figure 4.1: Item S121 and associated Category Probability Curves

## Item S004

These 3 daisies have been joined up to make a triangle.

Use your ruler to join up some of these  
daisies to make  
a **rectangle**



### Marking Key

2 ..... rectangle well drawn

1 ..... correct daisies chosen, but rectangle poorly drawn

eg: sides not straight, corners not joined, opposite sides not parallel

or

four sided figure well drawn but not a rectangle

0 ..... inadequate response

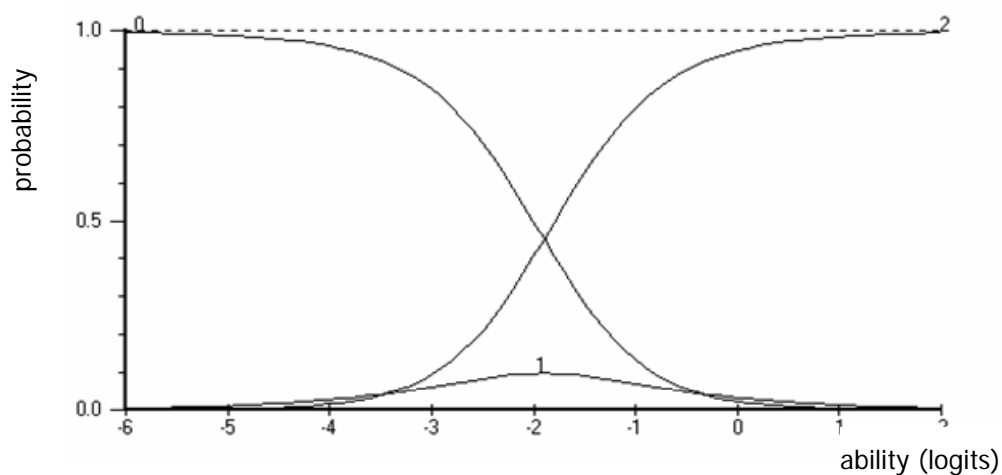


Figure 4.2: Item S004 and associated Category Probability Curves

In each case, the student is prompted to complete a drawing, which is then scored 0/1/2, according to criteria described by the marking key.

Despite the similar nature of the items, the associated Category Probability Curves reveal that Item S121 has ordered Rasch thresholds, while Item S004 has disordered Rasch thresholds. In order to understand why this has occurred, it is helpful to compare the marking keys, for it is in the functioning of the marking keys that threshold reversal occurs.

The point has already been made that the items, and their marking keys, have a similar structure. Both require a drawing to be completed, which is then marked according to the degree of accuracy with which it is completed. There is an important difference however even though three levels of response have been identified in each. In Item S121, one mark is awarded for a diagram which is mostly correct. Two marks are awarded for a diagram which is completely correct. For the target student group, in this case Year 3 students, both responses are reasonably likely to occur. There are a number of shapes in the diagram which need to be copied accurately, oriented correctly, and placed in the correct location. As a result, there are many opportunities for a student who does not have highly developed skills in drawing, orientating and locating to get something wrong, and so score one mark. It is also reasonable to assume that it is the students who do not have well developed drawing skills who are most likely to get something wrong, and so score one mark. Students who have well developed drawing skills are much more likely to be the ones who score two marks. In other words, the marking key appears to identify a developmental continuum – made up of three categories - for drawing shapes, and place students on this continuum appropriately according to their ability. The categories are working as intended.

In an apparently similar way, Item S004 identifies three levels of response, with one mark awarded for a diagram which is mostly correct, and two marks awarded for a diagram which is completely correct. Once again, the target group is Year 3 students. In this case, however, it appears to be much less likely for both responses to occur. The reason for this is that, although the question appears to be testing drawing ability, and the marking key certainly suggests this, the real demand of this item is in understanding the concept of a rectangle. Support is provided for drawing the rectangle in the form of the daisies and students can use a ruler to join them. If students know the shape of a rectangle, they will select the set of daisies which form that shape, and join them together. Because the vertices are provided, it is unlikely that they will not draw (reasonably) straight sides, or not join the corners, especially with a ruler. In other words, it is unlikely that they will draw the rectangle poorly. Of course, if they do not know the shape of a rectangle, then it *is* likely that they will draw it poorly but also not join the correct daisies. In this case, they will score no marks, rather than one mark.

As a result, even those students who, based on their performance on other items might be expected to get 1 mark, are likely to get the item wholly correct and score two marks or get it wholly wrong and score no marks. A partially correct drawing is not a likely response for any group of students. The marking key is not identifying a continuum of performance in drawing - the item is really about identifying rectangles, and generally students either can or cannot identify rectangles. Their performance in drawing the rectangles is not based on their ability to draw. Instead, it is based on their ability to identify rectangles, which evidently only has two levels of achievement – they can do it or they can't do it.

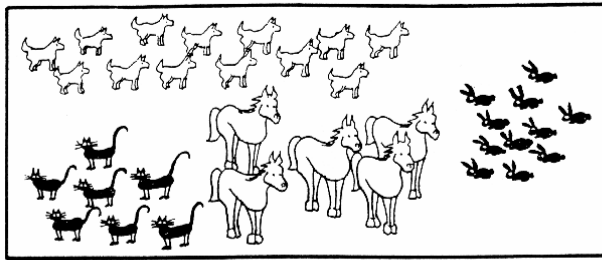
The significant point arising from this discussion is that it is the disorder in the thresholds of Item S004 which signals that the item is not functioning as intended. Expert analysis of the item itself is then required to identify the nature of the problem.

## 4.2 INCREMENTALLY SCORED ITEMS

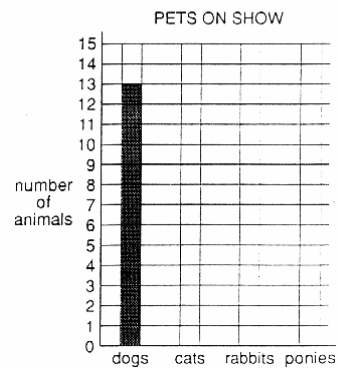
Items C011 and C155, taken from the *Chance and Data* strands of the 1992 and 1996 testing programs, are examples of incrementally scored items. These items are shown in Figures 4.3 and 4.4 respectively. Although the content of the items is quite different, their structure is the same.

## Item C011

Look at the pets on show.



Finish this Pet graph using the numbers from your chart.



### Marking Key

- 3 ..... all columns correct  
 2 ..... two columns correct  
 1 ..... one column correct  
 0 ..... no columns correct

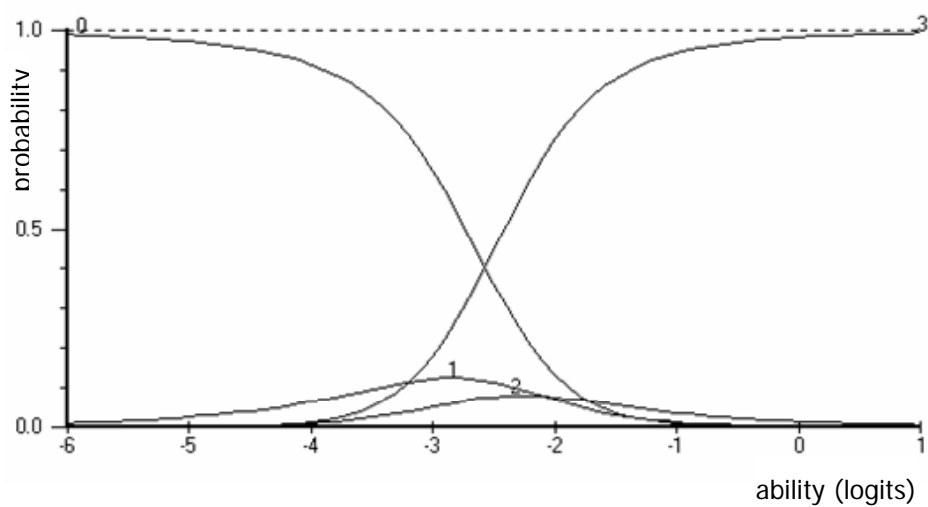
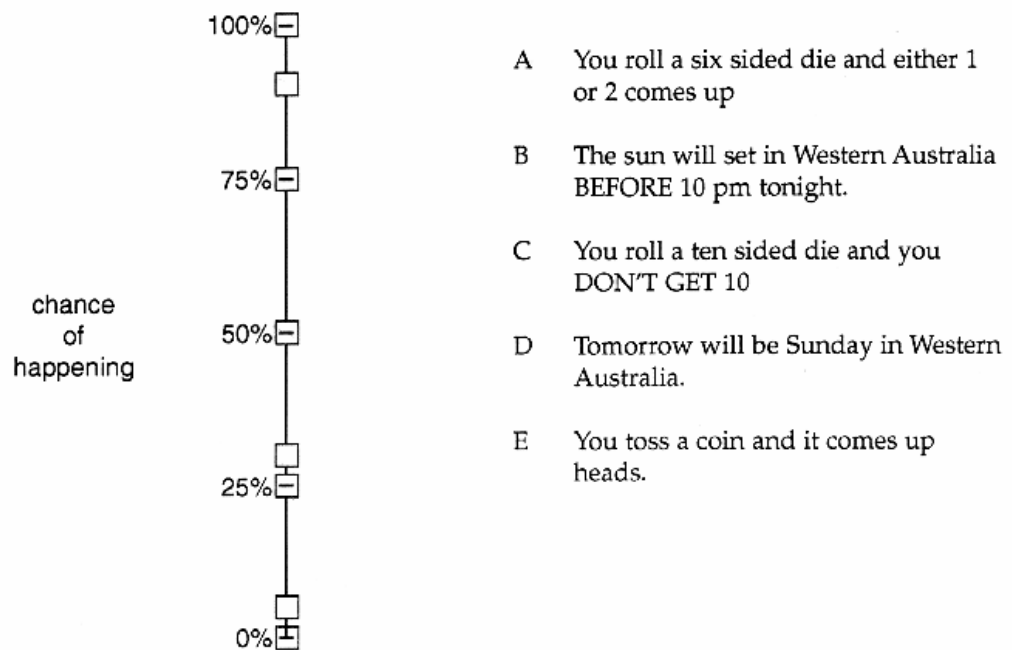


Figure 4.3: Item C011 and associated Category Probability Curves



## Item C155

Think about how likely it is that the following things will happen:



Match the letters to a box on a number line to show about what percentage chance you think each event has of happening.

### Marking Key

5/4/3/2/1 .....one mark for each correct matching

0 .....no correct matchings

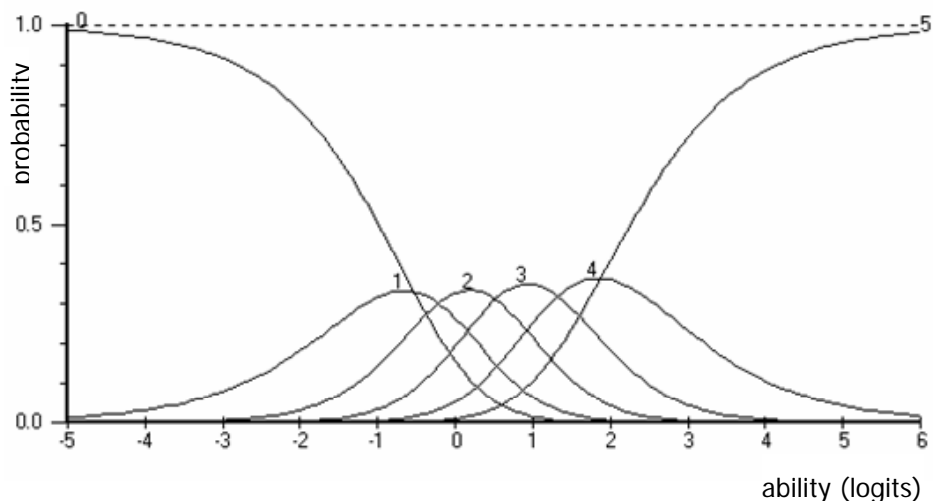


Figure 4.4: Item C155 and associated Category Probability Curves

Each is made up of a number of parts which need to be completed correctly, with the final mark built up from the number of correctly completed parts.

The Category Probability Curves of these items reveal that Item C155 has ordered Rasch thresholds, while Item C011 has disordered thresholds. In order to understand why this has occurred, it is helpful to look at the task demands of each item.

Careful analysis of Item C011 reveals the task demand in constructing each column is exactly the same. In each case, the student is required to count the number of animals – 7 cats, 11 rabbits and 5 ponies – and then draw a column on the graph to represent this number. The graph is already labelled, so there are no complicating decisions to be made about layout of the graph. The scale includes all consecutive numbers in the range required, so there is no complicating factor about some numbers not being shown on the scale. In this item, the student is simply required to complete an equivalent task three times.

The Category Probability Curve shows only two categories functioning correctly. Students are most likely to either get none of the columns correct, and score 0 marks or get all three columns correct and score 3 marks. Given that each of the three columns is equally difficult, this is not surprising. If students cannot complete one column correctly, then it follows that they will score 0 marks. If they can complete one column correctly, however, it is likely that they will repeat the process, and also complete the second and the third columns correctly. In this way, the item is functioning dichotomously. It is simply testing whether students understand how to draw a column graph. There are no levels of performance beyond the simple question of “can they do it or can’t they do it”. Importantly, it is the disorder of the thresholds that provides the signal that the item is not functioning as intended. Ignoring this signal implies the item has a measurement capability which it clearly does not possess.

The properly ordered thresholds of Item C155, in contrast, show all six categories working as intended. In this item, the student is required to match each of five events with the probability of it occurring. The key difference between this item and the previous item is that the events are not all equally familiar or equally easy. The events range from some with which all students will be quite familiar, such as the probability of a coin coming up heads and the probability that the sun will set before 10pm (although this is made a little more difficult because of the less familiar notion of 0%), through to some which are quite challenging, such as the probability that a six-sided die will come up with a 1 or 2. As a result, each successive

score on this item indicates more ability. A student who scores 2 marks does not do, twice, the same thing as a student who scores 1 mark does once. They do the same thing, and then go on and do something else which is more difficult. A student who scores 3 marks does this, and then goes on and does something even more difficult as well. And so on.

In this way, it can be seen that Item C155 is measuring student's ability in six categories, in a way which Item C011 failed to measure in any more than two categories. Once again, it is the thresholds, now ordered, which confirm that the categories are working as intended.

### 4.3 DECREMENTALLY SCORED ITEMS

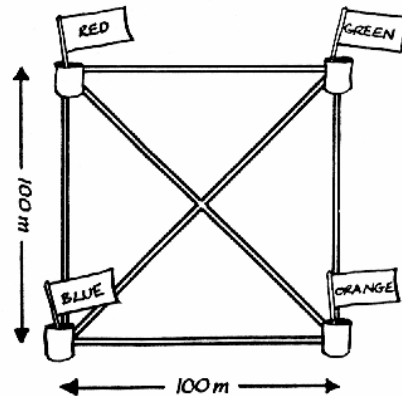
Items S019 and C016, taken from the *Space* and *Chance and Data* strands of the 1992 testing program, are examples of decrementally scored items. These items are shown in Figures 4.5 and 4.6 respectively. In each case, the marking key specifies the correct response, scored 2 marks, a partially correct response worth 1 mark, and a response that receives 0 marks.

## Item S019

Later you and your friends have a bike racing competition. Three courses are available. The turning points are marked by coloured flags. The courses are shown on the following map.

They are:

RED-GREEN-ORANGE-RED  
 RED-ORANGE-BLUE-RED  
 GREEN-ORANGE-BLUE-GREEN



- (b) If you ride the GREEN-ORANGE-BLUE-GREEN course, through what angle does your bike turn at

(ii) the BLUE mark? \_\_\_\_\_

### Marking Key

2 .....  $135^\circ$

1 .....  $45^\circ$

0 ..... incorrect response

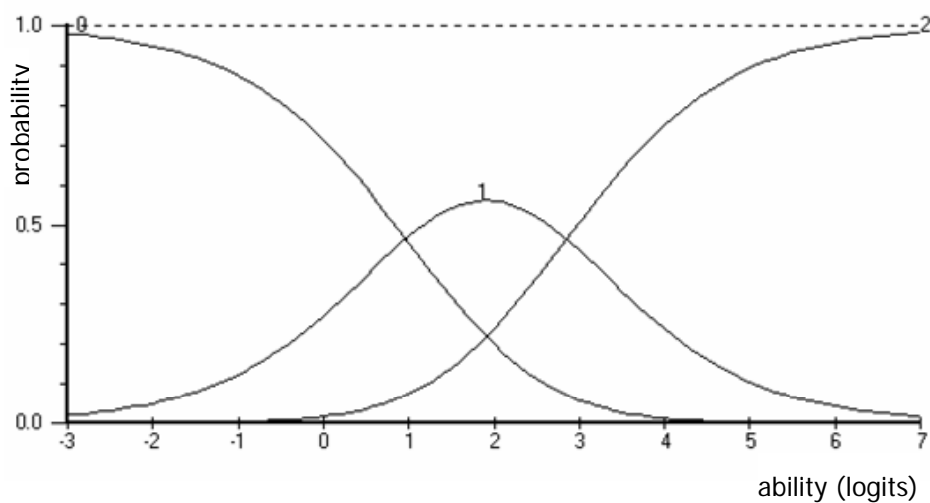
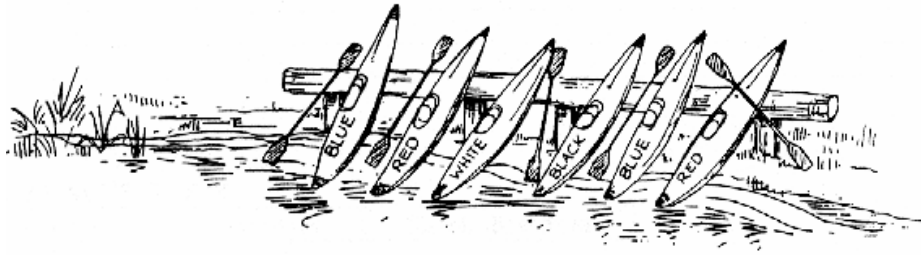


Figure 4.5: Item S019 and associated Category Probability Curves

## Item C016

Six children are going to have a kayak race. They draw coloured marbles out of a hat, each representing a different kayak.



Doug chooses first. What are the chances that he will get a blue kayak?

\_\_\_\_\_

Doug gets his blue kayak. Amanda is to choose second. She wants either a blue or a black kayak. What are the chances that she gets what she wants?

\_\_\_\_\_

### Marking Key

2 ..... 2/5

1 ..... 1/5

0 ..... incorrect response

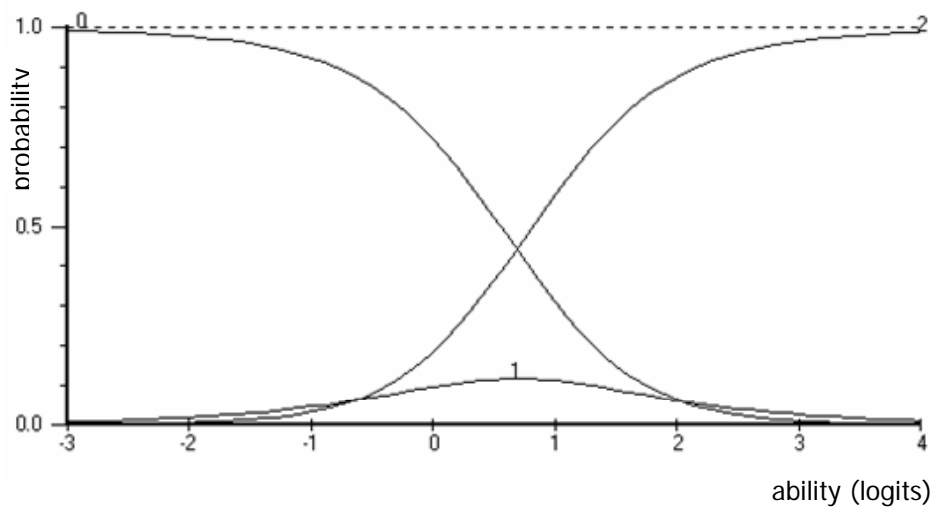


Figure 4.6: Item C016 and associated Category Probability Curves

The Category Probability Curves of these items reveal that Item S019 has ordered Rasch thresholds, while Item C016 has disordered Rasch thresholds.

Once again, in order to understand the source of the disordered thresholds in Item C016, it is helpful to analyse the task of these items.

In response to the question “What are the chances...” in Item C016, the marking key awards a score of 1 mark to a response of “one out of five”, and 2 marks to a response of “two out of five”. To arrive at the partially correct response, the student correctly recognises that there are five canoes to choose from, but incorrectly states that only one will suit Amanda. To arrive at the fully correct response, the student must recognise that two of the canoes will suit Amanda.

The problem here is that there is no logical or empirical evidence to suggest that these responses form a developmental continuum, such that it would be meaningful to place them along an achievement scale. While it is true that the response “two out of five” is the higher level response, there is nothing to suggest that a response of “one out of five” marks a position *along the way* to a response of “two out of five”. Faced with the question “What are the chances...” any student who understands that fractions will provide an answer, and that the choice and hence the denominator is “out of five”, already knows a great deal about dealing with probability. Knowing that out of the possible choices there is either one canoe or two canoes which meet the required condition is not the next developmental step in dealing with probability. This is a simpler concept. It is much more likely to have been understood *well before* the concept of “out of five”. As a result, the partially correct response identified by the marking key is not developmentally easier than the fully correct response. On a difficulty scale, both are comparable. As a result, the item really functions dichotomously. Students who have little understanding of simple probability are not likely to even get the “out of five” part right, and will score 0 marks. Students with sufficient understanding to get the “out of five” part right are also likely to get the number of chances right as well. As a result, the middle response category, with a score of 1, fails to function properly. Importantly, the threshold disorder which results is the signal that the item has failed to function as intended.

Although Item S019 and its associated marking key have a similar structure to Item C016, its Rasch thresholds are ordered, indicating that the categories are functioning as intended. The

explanation for the difference once again lies at the level of the conceptual understanding which the item attempts to measure. Item S019 shows the map of a bicycle race, which includes a turn of  $135^\circ$  at one point. Students are asked to describe this angle. The marking key awards a score of 2 marks for a response of  $135^\circ$  and 1 mark for a response of  $45^\circ$ . As the diagram reveals, the difference between the levels of these responses is quite substantial. The path which the bicycle takes, as drawn, is an angle of  $45^\circ$ . An entirely reasonable, although incorrect, answer would be to place a protractor over the corner, and read off an angle of  $45^\circ$ . It takes a significantly higher level of reasoning to see that for the bicycle to turn *back* along an angle of  $45^\circ$ , it must turn *through* an angle of  $135^\circ$ . This is most readily visualized as the angle traced by the front wheel of the bicycle as it turns through the corner, but it is not easily visualized in this way.

In contrast to Item C016, the marking key for Item S019 has identified two levels of response which do form part of a developmental continuum – understanding that the turned angle is really  $135^\circ$  is more difficult, and comes later developmentally, than recognizing that the drawn angle is  $45^\circ$ . Here there is logical and empirical evidence that it is meaningful to place these responses in this order on an achievement scale. Students who are awarded 1 mark for this item know something about angle measure, whereas students who are awarded 2 marks know this, but also know something more. The categories are working as intended, and this is evident in the ordering of the thresholds.

## 5. IMPLICATIONS OF THRESHOLD DISORDER

The three classes of polytomous items proposed in this chapter reflect important differences in the way the marking keys operate to identify a hierarchy of responses, with scores allocated to each level in the hierarchy. Nevertheless, in all three classes, there is an intended ordering in the hierarchy – a higher level response is expected to require more ability than a lower level response, and the scoring key is structured to reflect this. In all three cases, it has been seen that when the response data for each item is analysed according to the Rasch model, ordered Rasch thresholds provide confirming evidence that the hierarchy of responses identified in the marking key reflects the underlying order of the ability continuum – the marking key “gets it right”. Just as importantly, when disordered Rasch thresholds are derived from the Rasch model, it has been seen that this provides evidence that the marking key does not reflect the underlying ability continuum – in such cases, the marking key “gets it wrong”. Disorder in the thresholds is the signal that the item has failed to function as intended, and provides an opportunity to change the item, or change the marking key so that it

does reflect the hierarchy of responses that the item has elicited.



## APPENDIX

### 1.

Following Andrich's (1978) derivation of the threshold form of the Rasch model (Andrich, 1978), a second school of thought has emerged about how to understand the problem of threshold disorder, and what to do about it.

According to this interpretation, there is no requirement that threshold values be ordered.

According to this view, thresholds refer only to a local relationship between a pair of adjacent categories, and are characterised as steps. The difficulty of reaching each threshold is identified not as the difficulty of reaching that threshold in relation to all other thresholds but only in relation to the previous threshold. From this viewpoint, it is perfectly reasonable for a later threshold in a problem to be easier than an earlier threshold. It simply means that any person who passes the earlier, more difficult, threshold is highly likely to pass the later one as well. Consequently, while an item which displays such thresholds may fail to offer useful discrimination it is not dysfunctional and can still be used in the analysis.

Disordered thresholds are not taken to mean that there is a problem with the way an item is functioning, but only that there is a problem in the way the item can be displayed.

The problem of using such items to construct the achievement scale is resolved by calculating an alternative set of thresholds, called Thurstone thresholds, based on the Thurstone Cumulative Probability Model. The thresholds derived from the Thurstone model are always in a natural order, irrespective of the order of the thresholds produced by the Rasch model.

### 2.

The form of the Rasch extended logistic model underpinning this interpretation is given in Masters (1982) and Wright and Masters (1982). Instead of the threshold parameter  $\tau_{ki}$  they use the parameter  $\delta_{ki}$  and refer to this parameter as a "step". In their version, the probability of a person of ability  $\beta_n$  being classified in category  $x$  in a test item with ordered response categories scored  $0, 1, \dots, m$ , where the difficulty level of the threshold of category  $x$  is  $\delta_{xi}$  is given by:

$$\Pr\{X_{ni} = x\} = \frac{e^{\sum_{j=0}^x (\beta_n - \delta_{ij})}}{\sum_{k=0}^m e^{\sum_{j=0}^k (\beta_n - \delta_{ij})}} \quad (12.8)$$

This equation can be shown to be identical to equation (2.6) (Luo, 2005)

Following the interpretation of Wright and Masters (1982), Masters (1982) and Masters and Wright (1997), disordered thresholds merely indicate that later categories are less difficult than earlier categories. The process of moving from the more difficult to the less difficult categories is characterised as one of simply taking larger then smaller steps. In order to overcome the difficulty of displaying these categories in a meaningful way on an achievement scale, they draw on the work of Edwards and Thurstone (1952) and Samejima (1969) where the cumulative probability of passing each threshold is used to construct the scale. These Thurstone thresholds are, in the way that they are constructed, always ordered.

As a result, items with reversed Rasch thresholds are not viewed as problematic, and are routinely incorporated, together with items with ordered Rasch thresholds, in analyses and displays. Below is a summary as to why this view hides problems with items revealed by the Rasch model.

### 3.

The following analysis is an attempt to understand the effect of Wright and Masters' interpretation on understanding item functioning when thresholds are revealed to be disordered in the Rasch model. Each of these items has already been discussed extensively in Section 4 in relation to Andrich's interpretation of threshold disorder.

#### **Item S004**

Wright and Masters, following their interpretation of threshold disorder, would explain the reversed thresholds of Item S004 in terms of it being difficult to score the first mark, and then, having scored the first mark, easy to score the second mark, in completing the item. In their own terms - a large step followed by a small step. In order to display these thresholds in a meaningful way, they rescale the disordered Rasch thresholds as Thurstone thresholds. This would show the second threshold as more difficult, but only marginally so, than the first threshold.

This explanation invites the following observation.

This example provides a clear illustration that passing thresholds cannot be characterised by the taking of steps. Clearly, anyone who scores two marks for this item does not firstly gain one mark by producing a poorly drawn rectangle, one having sides not straight or corners not joined, and then go on to gain a second mark by producing a well drawn rectangle. They simply draw a rectangle. If they draw it well, they score two marks. If they draw it poorly, they score one mark. This is what is intended by this item, and this is how it is marked. There is no sequence of steps.

### **Item C011**

Once again, Wright and Masters, following their interpretation of threshold disorder, would explain the reversed thresholds of Item C011 in terms of it being relatively difficult to score the first mark, in this case by correctly completing the first column of the graph, but relatively easy to score the second mark and third marks by correctly completing the second and then third columns as well.

Of course, such an explanation misses the point that the task demand in constructing each column is exactly the same.

### **Item C016**

Once again, Wright and Masters, following their interpretation of threshold disorder, would see the reversed thresholds of Item C016 in terms of it being relatively difficult to score the first mark, in this case by producing the partially correct response, but relatively easy to score the second mark by improving on this to produce the fully correct response.

This example provides another illustration that passing thresholds cannot be characterised as the taking of steps. Students who score two marks for this item do not firstly gain one mark by providing a partially correct response, and then go on to gain a second mark by giving a fully correct response. They simply provide a response. If they do it well, they score two marks. If they do it less well, they score one mark. There is no sequence of steps that students need to take in completing this item, and furthermore, there is no suggestion in the marking key that marks are allocated according to any sequence either.

#### 4.

This analysis examined the functioning of a number of polytomous items with disordered thresholds, and compared them with similar items where threshold disorder did not occur. As a result, it became clear that there was nothing in the way these items functioned to suggest that characterising thresholds as steps was either accurate or useful. Firstly, it was not accurate, because although each marking key described a sequence of responses, from an incorrect response to a fully correct response, there was no expectation that students, in answering these questions, were required or expected to follow the same sequence. The sequence described a hierarchy, not a pathway. Secondly, it was not useful, because it then served to legitimise the use of the Thurstone model to display these thresholds. In doing so, the opportunity to see disordered Rasch thresholds as signaling anomaly in the data and hence a flaw in the item itself, and to learn from this, was missed.

Amongst the items developed in the 1992 and 1996 testing programs, some of which have been studied here, a range of problems were identified. In some cases, the problem was a simple technical matter which could easily be put right in subsequent testing. In other cases, the problem was deeper, reflecting a fundamental misunderstanding of the ability continuum that the item was trying to operationalise. In such cases, major rewriting of the item or its marking key was required in order to achieve the desired range of responses, or it was rescored and subsequently re-analysed as a dichotomous item, reflecting the way that the item was functioning in practice. In extreme cases, the item was simply removed from the testing program.

Two important points emerge from this analysis.

The first is that there is no evidence here to support the view that thresholds can be interpreted as steps, which refer only to a local relationship between categories. Interpreting thresholds as steps suggests that, as it is not unreasonable for later steps to be of lesser magnitude than earlier steps, it is also not unreasonable for later thresholds to have lower values than earlier thresholds. This suggests that, in principle, reversal of thresholds is an acceptable outcome of the Rasch Model. The evidence in this analysis suggests instead that this is not the case.

The second is that by accepting, in practice, that reversal of thresholds does not indicate an underlying problem with the way an item is functioning, the opportunity to learn from the Rasch Model is lost. When thresholds are ordered, this is confirming evidence that the

marking key is a faithful reflection of the underlying ability continuum that it operationalises. When thresholds are disordered, it is evidence that the marking key fails to do this. Both outcomes provide valuable information about the underlying ability continuum, and properly understood, both can contribute to its measurement.

### 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, 561-574.
- Andrich, D. (2002) Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. *Journal of Applied Measurement*. 3, 325 – 357. Reprinted in E. V. Smith & R. M Smith, Introduction to Rasch Measurement: Theory, Models and Application. JAM Press; Minnesota. Ch. 8, Pp 167 – 200.
- Andrich, D, (2005) The Rasch model explained. In Sivakumar Alagumalai, David D Curtis, and Njora Hungi (Eds.) *Applied Rasch Measurement: A book of Exemplars*. Springer-Kluwer. Chapter 3, 308 - 328.
- Edwards, A.L. & Thurstone, L.L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika*, 17, 169-180.
- Luo, G. (2005) The relationship between the Rating Scale and Partial Credit Models and the implication of disordered thresholds of the Rasch models for polytomous responses. *Journal of Applied Measurement*, 6(4) 443-455.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Masters, G.N. and Wright, B.D. (1997) The partial credit model. In W.J. van der Linden and R.K. Hambleton (Eds.) *Handbook of Item Response Theory*. (pp. 101– 121). New York. Springer.
- Luo, G. (2005) The relationship between the Rating Scale and Partial Credit Models and the implication of disordered thresholds of the Rasch models for polytomous responses. *Journal of Applied Measurement*, 6(4) 443-455.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, 34(2, No.17).
- Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.