Understanding the response structure and process in the polytomous Rasch model

David Andrich

Murdoch University, Western Australia

Mailing address
David Andrich
Murdoch University
Murdoch 6150
Western Australia

Acknowledgements

Recent discussions with Guanzhong Luo helped in the articulation of the model for this Report. The research reported in this Report was supported in part by an Australian Research Council Linkage 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 (IIEP), and the Australian Council for Educational Research (ACER) as Industry Partners*. This Report 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. 10 ARC Linkage Grant LPo454080: Maintaining Invariant Scales in State, National and International Level Assessments. D Andrich and G Luo Chief Investigators.

The Polytomous Unidimensional Rasch Model: Understanding its response structure and process

1. Introduction

The Rasch model for ordered response categories in standard formats was derived from a sequence of theoretical propositions requiring invariance of comparisons among item and among person parameter estimates. A model with sufficient statistics is the consequence. The model was not derived to describe any particular data. (Rasch, 1961; Andersen, 1977; Andrich, 1978a).

Standard formats involve only one response in one of the categories deemed *a-priori* to reflect *increasing* levels of the property and are common in quantification of performance, status and attitude in the social sciences. The advantage of an item with more than two ordered categories is that it gives more information than a dichotomous item. Table 1 shows three common examples. Figure 1 shows a graphical counterpart using the first example in Table 1.

The ordered categories on the hypothesised continuum in Figure 1 are contiguous. They are separated on the continuum by successive points termed *thresholds*. This is analogous to mapping a location of an object on a line partitioned into equal units to obtain a physical measurement. There the objects of measurement may be located anywhere on the continuum, and the size of the categories is the unit which reflects the precision of the measurement. The measurement then is the nearest integer count of the number of units or categories from an origin. Because we do not have a fixed origin, there is no end point on the latent continuum of Figure 1 for the extreme categories, only partitions of the continuum into four contiguous categories which requires three thresholds.

In elementary analyses of data obtained from formats such as those in Table 1, and by analogy to physical measurement, successive integers are assigned to the categories and simply summed. In more advanced analyses, a probabilistic model that accounts for the finite number of categories and for sizes of the categories are applied. The

Rasch model is one such model. Nevertheless, the measurement conceptualisation of locations of objects on a continuum partitioned into contiguous categories is retained.

Table 1

Standard response formats for the Rasch model

Fail	<	Pass	<	Credit	<	Distinction
Never	<	Sometimes	<	Often	<	Always
Strongly Disagree	<	Disagree	<	Agree	<	Strongly Agree

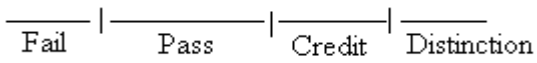


Figure 1. Graphical representation of ordered categories

1.1 Identity of rating and partial credit models

Before proceeding, a note is made on the terminology of the Rasch model. In the examples in Table 1 it might be considered that if the same format is used across all items, that the sizes of the categories will also be the same across all items. However, that is an empirical question, and it is possible that there is an interaction between the content of the item and the response format, so that the sizes of the categories are different for different items. It may also be the case that different items have different formats that are natural to the item with different numbers of categories, as for example when there are different items in an achievement test with different maximum scores. The Rasch model which has the same format across all items *and* has the same sized categories is referred to sometimes as the *rating scale model*. The model with different sized categories or with different numbers of categories is

referred to sometimes as the *partial credit model*. The difference, as will be seen, is only a matter of parameterization, and at the level of a single person responding to a single item, the models are identical. The focus of this Report is on the response of one person to one item which covers both parameterizations. Therefore, the model will be referred to simply as the polytomous Rasch model (PRM). The dichotomous model is simply a special case of the PRM, but where it is necessary to distinguish it as a dichotomous model in the exposition, it will be referred to explicitly as the *dichotomous* RM.

1.2 Distinctive properties of the PRM

The PRM has two properties that, when first disclosed, were considered somewhat counterintuitive: first, combining adjacent categories by summing the probabilities of responses in the categories and in the related sense of summing their frequencies to form a single category, can only be done under very restricted circumstances (Rasch, 1966; Andersen, 1977; Andrich, 1978a, Jansen and Roskam, 1986); second, the thresholds analogous to those in Figure 1 which define the boundaries of the successive categories may take on *values* that are not in their natural order. In part because these properties are exactly opposite to those of the then prevailing model for ordered categories, that based on the work of Thurstone (Thurstone and Chave, 1929), they have been ignored, denied, circumvented or generated debate and misunderstandings in the literature (Andrich, 2002). This earlier model is briefly summarised and contrasted with the PRM later in the Report.

1.3 Criterion that ordered categories should satisfy

One observation from these reactions to the model's properties is that in the development of response models for ordered category formats, there is no *a-priori* articulation of any criterion that data in ordered categories should satisfy – it seems it is simply assumed that if categories are deemed to be ordered, that they will necessarily operate that way. One factor that immediately comes to mind as possibly violating the required order is respondents not being able to distinguish between two adjacent categories. This has been observed in data in Andrich (1979). There may be

any number of reasons why respondents cannot distinguish between adjacent categories.

The theme of this Report is that it is an empirical hypothesis whether or not ordered categories work as intended. The Report sets up a criterion that must be met by data in an item response theory framework for it to be evident empirically that the categories are working as intended, and shows how the PRM makes a unique contribution to providing the empirical evidence. Meeting this requirement empirically is necessary because, if the intended ordering of the categories does not reflect successively more of the property, then it puts into question the very understanding of what it means to have more of the property and of any subsequent interpretations from the data. The Report is not concerned with issues of estimation and the tests of fit, which are well covered in the literature, but in better understanding the distinctive properties of the model itself, and the opportunities it provides for the empirical study of ordered polytomous response formats.

1.4 Ordering as a property of the data and not the model of analysis

An integral part of the theme of the Report is that the criterion for ordered categories are working as intended *pertains to the data*, and not to response models themselves *irrespective of the data*. The importance of distinguishing between the properties of data from procedures of models of analysis for ordered categories was recognised by R. A. Fisher who had a method for analysing data intended to be in ordered categories, and upon obtaining results for a particular data set noted: *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).* Any note from Fisher is worthy of substantial consideration and study (Rasch, 1979).

This Report demonstrates that the properties of the PRM are compatible with treating the operation of the ordered categories as an empirical hypothesis. In particular, it is demonstrated that the model has the remarkable property that from a set of structurally dependent responses in an ordered category format, *it recovers information that would arise from compatible, experimentally independent formats*.

This permits the inference regarding the empirical ordering of categories. Thus the Report does not merely describe the Rasch model for ordered categories from the perspective of modelling data and for providing invariant comparisons, but presents a case that it is the ideal model for characterising the intended response process and for testing empirically whether ordered categories are operating as required.

The Report is organised as follows. Section 2 develops an axiom for ordered response categories. The axiom arises from the conceptualisation of an experiment with independent responses at thresholds that could be devised to assess the empirical ordering of the categories. Section 3 analyses in detail response outcome spaces, something to which item response theory generally pays little attention, and shows that three different outcome spaces need to be distinguished and their relationship understood. Section 4 derives the Rasch model and relates it to the axiom for ordered categories, illustrating relationships with analyses of simulated data. Section 5 demonstrates why the probabilities and frequencies in adjacent categories cannot be summed in the PRM except in special circumstances, and characterises the model from a response process at each threshold. This permits a clearer contrast of the PRM to the other most commonly used model for ordered categories mentioned earlier, that based on the work of Thurstone which has been developed further by Bock (1975, Ch.8), Samejima (1969, 1996, 1997), and McCullagh.(1980). In the psychometric literature, this model is generally known as the *graded response model* (GRM). This model is derived briefly, and for completeness, contrasted with the Rasch model in Section 6. Section 7 is a summary which includes a suggestion as to why over the long history of the development and application of models for data in ordered categories, and despite the lead from Fisher, no previous criteria have been articulated that ordered categories must meet.

As will be seen, in understanding the Rasch model for ordered response categories and its implications, subtle and unusual distinctions between *real* and *implied*, *latent* and manifest response spaces, the *values* and *structure* of response probabilities, and between *compatible* and *determined* relationships, have to be made.

2. Criteria for data in ordered response categories

In preparation for developing and specifying an Axiom for the empirical ordering of categories, we consider some relationships between models and data. These relationships are generally taken for granted, but they are made explicit here because of their specific roles in relation to the PRM and the theme of the Report.

2.1 A comment on the uses of models

One use of models is simply to summarise and describe data. Models *describe* data in terms of a number of parameters which are generally substantially smaller than the number of data points. It is of course necessary to check the fit between the data and the model to be satisfied that the model does describe the data.

A second use of models is to characterise the *process* by which data are generated. For example, the Poisson distribution arises from, among many other circumstances, the "cumulative effect of many improbable events" (Feller, 1968, p.282), that is, where the probability of each event occurring is low, but where the opportunity for the event to occur is high. The model above is derived *a-priori* to the data in characterising a response process. If the data do not fit the model, then a question might be asked about its characterisation of the process. However, the fit of the data to the model is only a necessary, not sufficient condition, to confirm that the model characterises the process in those data.

A third and much less conventional use of models is to express *a-priori* conditions that data are required to follow if they are to subscribe to some principles. As indicated above, this is the case with the PRM. Following a series of studies, Rasch articulated conditions of invariance of comparisons that data should have if they are to be useful in making quantitative general statements. Specifically,

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison;
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; (Rasch, 1961 p.332).

These conditions of invariance were not unique to Rasch –Virtually identical conditions were articulated by Thurstone (1928) and Guttman (1950) before him. However, the distinctive contribution of Rasch, beyond those of Thurstone and Guttman, was that Rasch expressed these conditions in terms of a probabilistic model. Rasch wrote his conditions for invariance in a general equation, which, in the probabilistic case for dichotomous responses, takes the form

$$\Pr\{(Y_{ni} = y_{ni}, Y_{nj} = y_{nj}); \beta_n, \delta_i, \delta_j \mid f(y_{ni}, y_{nj})\} = \vartheta(y_{ni}, y_{nj}, \delta_i, \delta_j), \quad (1)$$

where Y_{ni}, Y_{nj} are random variables whose responses (y_{ni}, y_{nj}) take the values $\{0,1\}$, β_n and δ_i, δ_j are location parameters of person *n* and item *i* and item *j* respectively, and the right side of Eq. (1) is independent of the person parameter β_n. As indicated already, this leads to a class of models with sufficient statistics for the parameters, which generalises to the PRM.

The key advantage of specifying the conditions in terms of a model is that mathematical consequences, some of which might be initially counterintuitive, can be derived. This is the case with the Rasch model. However, when the consequences follow mathematically from the specification as compelling as that of making relatively invariant comparisons, then because they can provide genuinely new insights that might not be apparent immediately intuitively, they should be understood. Another distinctive consequence of this use of a model is that no amount of data analysis and demonstration of fit to the model or otherwise is relevant to the *case for the model*.

Ideally, and where relevant, all three conditions, characterising data, characterising a process, and characterising a requirement, would be met. Again, it will be shown that in circumstances relevant to social measurement where the response to an item is in ordered categories, the PRM satisfies the last two conditions, it always being an empirical question whether or not data are described by a model.

2.2 Definition of a threshold in the presence of experimental independence

Suppose that it is intended to assess the relative location of persons on some construct that can be mapped on a linear continuum, for example an achievement test. Items of would be landmarks of achievement (Thurstone, 1925) with items of successively increasing difficulty reflecting greater achievement on the continuum and requiring successively increasing ability for success.

Suppose the responses of persons to items are scored dichotomously, for a *successful* and *unsuccessful* response respectively. From such responses, and in an arbitrary unit, the dichotomous RM (Rasch, 1960, 1961; Wright & Panchapakesan, 1969; Fischer and Molennar, 1995; Wright, 1997) can be used to estimate the relative location of items on the continuum. This model takes the form

$$\Pr\{Y_{ni} = y\} = \frac{e^{y(\beta_n - \delta_i)}}{1 + e^{\beta_n - \delta_i}} \qquad (2)$$

where the variables are identical to those of Eq. (1). The response function of Eq.(2) for $y = 1$ is known as the item characteristic curve (ICC). Three ICCs for the dichotomous RM are illustrated in Figure 2. The data giving rise to estimates in Figure 2 were simulated for 5000 persons responding to six items independently (two sets of three items) with locations of $-1.1, -0.1, 1.2$ respectively in the first set. Only the responses of the first set of three items are shown in Figure 1. These data, together with those of the second set, are used to illustrate other points later in the

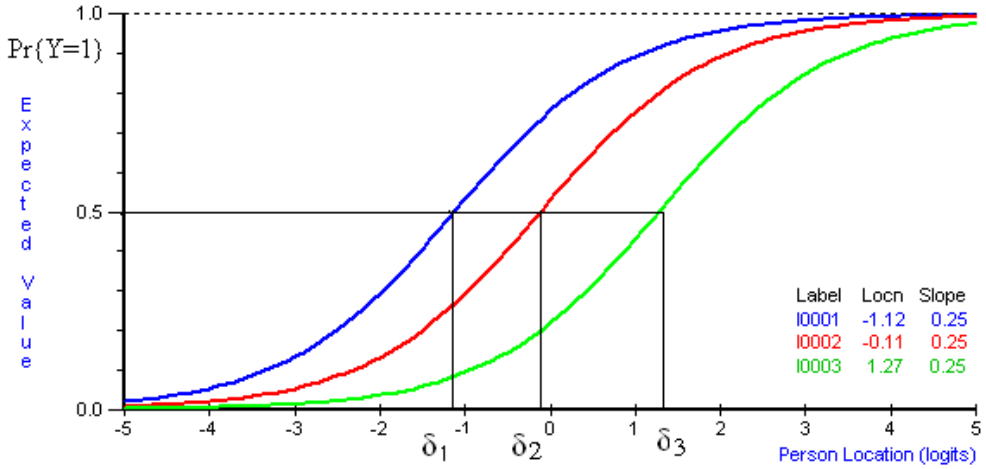


Figure 2: ICCs and location estimates for three items

The responses $\{0,1\}$ are ordered; response $y = 1$ is deemed *successful*, the response $y = 0$ *unsuccessful*. In achievement testing δ_i is referred to as the difficulty of the item. In general terms, and following psychophysics (Bock and Jones, 1968), it is termed a *threshold* – it is the point at which the person with same location $\beta_n = \delta_i$ has equal probability of being successful and unsuccessful -

$\Pr\{Y_{ni} = 1\} = \Pr\{Y_{ni} = 0\} = 0.5$.

In the Rasch model, the ICCs are parallel, with many implications (Wright, 1997), most of which are exploited in this Report. We use the dichotomous RM to construct the PRM. However, to better understand the PRM for more than two ordered categories, the two parameter logistic model (2PLM),

$$\Pr\{Y_{ni} = y\} = \frac{e^{y\alpha_I(\beta_n - \delta_i)}}{1 + e^{\alpha_I(\beta_n - \delta_i)}}, \qquad (3)$$

where α_i, known as the discrimination parameter, characterises the slope of the ICC, is encountered (Birnbaum, 1968).

2.3 The Guttman structure and the dichotomous Rasch model

Another more explicit view of the location of items of increasing difficulty, equivalent to Thurstone's notion of landmarks, is that of Guttman (1950) who enunciated an idealised *deterministic* response structure for unidimensional items. The Guttman structure is central in understanding the PRM.

For I dichotomous items responded to independently, there are 2^I possible response patterns. These are shown in Table 2 for the case of three items. The top part of Table 2 shows the subset of *patterns* of responses according to the Guttman structure. The number of these patterns is *I+1*.

Table 2
The Guttman Structure with I =3 dichotomous items in threshold order

Items	1	2	3	Total Score x	$\Pr\{(y_{n1}, y_{n2}, y_{n3}) \mid x\}$
I+1=4 Guttman response patterns					
	0	0	0	0	1
	1	0	0	1	0.667
	1	1	0	2	0.678
	1	1	1	3	1
2^I– I–1 =4 Non-Guttman response patterns					
	0	1	0	1	0.248
	0	0	1	1	0.085
	1	0	1	2	0.235
	0	1	1	2	0.087

The rationale for the Guttman structure in Table 2 (Guttman, 1950) is that for unidimensional responses across items, if a person succeeds on an item, then the person should succeed on all items that are easier than that item, and that if a person fails on an item, then the person should fail on all items more difficult than that item. The content of the items with different difficulties operationalises the continuum.

With experimentally independent items, it is possible that a deterministic Guttman structure will not be observed in data. Then to locate items on a continuum, the dichotomous RM may used. The dichotomous RM is a probabilistic counterpart of

the Guttman structure which is a deterministic limiting case (Andrich, 1985). Specifically, for any person, the *probability* of success on an easier item will always be greater than the *probability* of success on a more difficult item. This statement is evident from the parallel ICC curves in Figure 2.

In the Guttman structure, as is evident in Table 2, the total score, $x, x = \sum_{i=1}^{I} y_i$, completely characterises the response pattern. In the dichotomous RM, the total score plays a similar role, though probabilistically; it is a sufficient statistic for the person parameter (Rasch, 1961; Andersen, 1977). Then if item thresholds are ordered in difficulty and for a given total score, the Guttman pattern has the greatest *probability* of occurring.

Furthermore, because of sufficiency, the probability of any pattern, given the total score x, is independent of the person's ability. Thus the probabilities of the patterns of responses for total scores of 1 and 2, shown in Table 2, are given by

$$\Pr\{(y_{n1}, y_{n2}, y_{n3}) | x = 1\} = \frac{e^{-y_{n1}\delta_1 - y_{n2}\delta_2 - y_{n3}\delta_3}}{e^{-\delta_1} + e^{-\delta_2} + e^{-\delta_3}}, \quad (4)$$

$$\Pr\{(y_{n1}, y_{n2}, y_{n3}) | x = 2\} = \frac{e^{-y_{n1}\delta_1 - y_{n2}\delta_2 - y_{n3}\delta_3}}{e^{-\delta_1 - \delta_2} + e^{-\delta_2 - \delta_3} + e^{-\delta_1 - \delta_3}}, \quad (5)$$

respectively, both of which are independent of the person ability β_n and are special cases of Eq (1). These equations are the basis of conditional estimation of the item parameters independently of the person parameters (Andersen,1973).

2.4 Design of an experiment to assess the empirical ordering of categories

Section 2.3 summarised the relevant features of the Guttman structure and the dichotomous RM which are used later in the Report to build the PRM for more than two ordered categories. We now consider the design of an experiment in which the empirical ordering of the categories can be investigated. The key feature of this experiment is that of empirical *response independence* among the judgements.

To make the case concrete, consider the ordered category descriptors shown in Table 3 that were used in assessing the abilities of students to write a narrative in relation to a particular criterion. The responses among the categories are not independent in the sense that if a response is made in one category, it is not made in any other category. The task is to construct a design compatible with Table 3 in which independence of judgements prevails.

Table 3
Operational definitions of ordered classes for judging essays and a response structure compatible with the Rasch model

Fail (F)	*Inadequate setting* Insufficient or irrelevant information given for the story. Or, sufficient elements may be given, but they are simply listed from the task statement, and not linked or logically organised.
Pass (P)	*Discrete setting* Discrete setting as an introduction, with some details which also show some linkage and organisation. May have an additional element to those listed which is relevant to the story.
Credit (C)	*Integrated setting* There is a setting which, rather than simply being at the beginning, is introduced throughout the story.
Distinction (D)	*Integrated and manipulated setting:* In addition to the setting being introduced throughout the story, pertinent information is woven or integrated so that this integration contributes to the story.

Reprinted with permission from Harris, 1991, p.49. The labels of Fail, Pass, Credit, Distinction have been added in this Table for the purposes of this Report.

Clearly from the descriptors for each category, there is an intended ordering in the quality of performance with respect to the feature of *setting*.

We take it that the category descriptors operationalise the writing variable to be measured and describe the qualities that reflect successively better writing on this continuum. We further note that the first and least demanding category is a complement to the second category, and that the other categories show increasing quality of writing with respect to *setting*. We shall see how this complementarity of the first category to the others plays out.

The experimental design involves taking the descriptors in Table 3 and constructing independent dichotomous responses at thresholds which are of increasing difficulty.

Instead of one judge assigning an essay into one of four categories, consider a design with three judges. Each judge only declares whether each essay is successful or not in achieving the standard at one of Pass, Credit or Distinction. Thus we have three *independent* dichotomous random variables. Although there are four categories, there are only three independent responses. The F descriptor helps in understanding the variable in the region of Fail/Pass, and helps the judge decide on the success or otherwise of the essay at this standard. This is the role of the F descriptor in this design.

We now consider this experimental design, summarised in Table 4, more closely.

Table 4: Experimental design giving independent responses

	Inadequate setting F	*Discrete setting* P	*Integrated setting* C	*Integrated and manipulated setting* D
Judgement 1	**Not P**	**P**		
Judgement 2		**Not C**	**C**	
Judgement 3			**Not D**	**D**

The descriptors, as already indicated, describe the variable and what it takes to reflect more of its property. The better the essay in terms of these characteristics the greater the probability that it will be deemed successful at each level.

Two further specific cases may be highlighted in order to make clear the operation of the experimental design.

First, suppose the judge considers that the essay does satisfy the P descriptor, but observes that it does not meet the C or D descriptors. Then the judge should classify

the essay as a success. Second, suppose that the judge, still with respect to success at P, considers that an essay satisfies even the qualities described in C or D or some combination of these. Because of the implication of the structure of the descriptors as ordered categories, which implies that C reflects more of the property to be measured than P, and D even more of the property than C, the judge must classify it as a success at P. The more of the properties of C and D an essay has, the greater the probability of it being classified successful at the P level.

Similar interpretations follow for decisions at each of the other categories. Instructions would need to be clear to the judges to ensure their understanding of this experimental task.

2.5 Requirements of data from the experimental design

In such an experimental design it would be *required* that the success rate at P is greater than that at C, and that the success rate at C is turn be greater than at D. That is, it is required that it is more difficult to be successful at D than at C, which in turn is more difficult than being successful at P.

If that were not the case, for example, the success rate at C was the same as at D for the same essays, then it would be inferred that the judges do not distinguish between the two levels consistently. This could arise, for example, if the judge at C were harsher than intended and the judge at D was more lenient than intended. Thus it may be that the experiment did not work, and it would need to be studied further to understand why this is the case. But such evidence is central to treating the ordering of the categories as an empirical hypothesis to be tested in the data.

Not only would we require that the categories increase in difficulty with the a-priori ordering of the descriptors, but we would want this to be the case *irrespective* of the location of the essays on the continuum. It seems untenable that the empirical ordering of P, C and D is in one direction for some of essays in one part of the continuum and in a different direction for essays in another region. This uniformity of the relationships between these levels in probability is guaranteed if the success rate curves at different levels of quality of the essays are parallel, that is, if the dichotomous responses at that the corresponding thresholds follow the dichotomous RM. This is the essential justification for applying the dichotomous RM to such data.

In summary, if $\delta_P, \delta_C, \delta_D$ are the difficulties of the thresholds at the criteria of Pass, Credit and Distinction respectively, then we require the data to fit the dichotomous RM and that $\hat{\delta}_P < \hat{\delta}_C < \hat{\delta}_D$.

We stress that there is no artificial constraint in the design, or in the model of analysis, that would ensure the apparent correct ordering of the thresholds no matter the properties of the data. As required by Fisher, the ordering of thresholds and fit to the dichotomous RM would be a property of the data, not a constraint on the design or the procedure of analysis.

The suggested empirical design of Table 4 above may not be constructed in any particular data collection. However, it is argued here that any design that answers whether or not the ordering of the categories is working as intended *should be equivalent to this design*.

2.6 An axiom for the empirical ordering of response categories

The requirement for evidence of the empirical ordering of response categories which is not a property of the model of analysis, is called here an Axiom. It is called an Axiom because it seems fundamental to the analysis of response formats with ordered categories. It is also called an Axiom in order to draw attention to the surprising feature that no such requirement seems ever to have been articulated. We return to this point in Section 7.

Axiom: *In a design in which the objects of measurement are judged experimentally independently as successful or unsuccessful in meeting the requirements of successive, contiguous categories on a continuum, the categories will be said to be empirically ordered correctly if the relative difficulties of achieving a success increases with the intended ordering of the categories.*

3. Construction and analyses of outcome spaces

We now formalise some concepts and notation, and derive formal relationships concerned with response spaces. This is not typical in psychometrics but is essential in understanding the PRM. For purposes of exposition, we proceed first from the case of experimentally independent responses at the successive categories as in Table 4 to one in which the responses in categories are not independent as in Table 3. We then reverse the direction of derivations from the design of Table 3 in which a response are dependent, to infer experimentally independent responses as in Table 4.

3.1 The experimentally independent outcome space Ω

Let $\{Y_{nix} = y\}, y \in \{0,1\}$, be m_i experimentally independent random variables of the response for person *n* with respect to $x = 1,2...m_i$ thresholds for item *i*. For example, we assume that an item such as that in Table 3 has been decomposed into three independent judgements at successive categories beginning with the second category as in the design of Table 4. As noted earlier, with four categories, we can have only three independent responses.

Let $\Pr\{Y_{nix} = 1\} \equiv P_{nix}; 1 - P_{nix} \equiv Q_{nix}$. (6)

Let $\Omega \equiv \{(y_{ni1}, y_{ni2}, ...y_{nix}, ...y_{nim_i})\}$ be the outcome space for all 2^{m_i} possible response patterns across the *x* thresholds. For example, in the case of three thresholds as in Tables 2 and 4, there are $2^3 = 8$ possible response patterns. These, and the corresponding probabilities, are shown in Table 5. With experimental independence, the probability of any pattern $(y_{ni1}, y_{ni2}, ...y_{nix}, ...y_{nim_i})$ is given by

$$\Pr\{(y_{ni1}, y_{ni2}, ...y_{nix}, ...y_{nim_i})\} = \prod_{x=1}^{m_i} P_{nxi}^{y_{nxi}} Q_{nxi}^{1-y_{nxi}} . \quad (7)$$

In the case of the response pattern $(1,1,0)$ for three thresholds, for example,

$\Pr\{(1,10)\} = P_{ni1}P_{ni2}Q_{ni3}$. Clearly,

$$\sum_{\Omega} \Pr\{(y_{ni1}, y_{ni2}, \ldots y_{nix}, \ldots y_{nim_i})\} = \sum_{\Omega} \prod_{x=1}^{m_i} P_{nix}^{y_{nixi}} Q_{nix}^{1-y_{nix}} = 1. \quad (8)$$

The outcome space Ω and probabilities of all outcomes are also shown in Table 5.

Because it proves critical to keep a track of the outcome space of a response, we formalise Eq. (6) as

$$\Pr\{Y_{nix} = 1 \mid \Omega\} \equiv P_{nix}; \Pr\{Y_{nix} = 0 \mid \Omega\} = 1 - P_{nix} \equiv Q_{nix} \quad (9)$$

Table 5

The experimentally independent outcome space Ω and the Guttman subspace Ω'

y_{n1i}	y_{n2i}	y_{n3i}	
0	0	0	
Q_{ni1}	Q_{ni2}	Q_{ni3}	
1	0	0	
P_{ni1}	Q_{ni2}	Q_{ni3}	
1	1	0	
P_{ni1}	P_{ni2}	Q_{ni3}	Ω'
1	1	1	
P_{ni1}	P_{ni2}	P_{ni3}	
0	1	0	
Q_{ni1}	P_{ni2}	Q_{ni3}	
0	0	1	
Q_{ni1}	Q_{ni2}	P_{ni3}	Ω
1	0	1	
P_{ni1}	Q_{ni2}	P_{ni3}	
0	1	1	
Q_{ni1}	P_{ni2}	P_{ni3}	
$\sum_{\Omega} \Pr\{(y_{n1i}, y_{n2i}, \ldots y_{nki}, \ldots y_{nmi})\} = \sum_{\Omega} \sum_{k} {}^{3}\prod_{k=1}^{m_i} P_{nki}^{y_{nki}} Q_{nki}^{1-y_{nki}} = 1$			

In the case where the thresholds are ordered in difficulty, let $\{y_{ni1}, y_{ni2}, \ldots y_{nix}, \ldots y_{nim_i}\} = \{\underbrace{1,1,1,\ldots 1}_{x}, \underbrace{0,0,0\ldots,0}_{m_i - x}\}$ be a sequence of responses in which the first x are successes, followed by $m_i - x$ failures. That is, the response is a Guttman pattern.

3.2 The restricted outcome space of the Guttman structure Ω'

Let Ω' be the outcome subspace of all possible Guttman patterns. The number of patterns is $m_i + 1$. In the case $m_i = 3$, the number of response patterns is 4; the outcome space Ω' in this case is also shown in Table 5.

The next step is distinctive in understanding the implications of the PRM. Suppose there are a large number of essays that have been judged independently at each level. Although it would not be done in practice, for the purpose of the exposition, suppose that only those essays whose judgements conformed to the Guttman pattern were retained. This may also seem wasteful of data, but as will be seen when the process is reversed, the implied Guttman structure is the only form of data available.

Moreover, and perhaps remarkably, it will be shown that the threshold estimates obtained from applying the PRM to data of the Guttman subspace are statistically equivalent to those obtained applying the dichotomous RM to data of the whole space.

Let $X_{ni} = x = \sum_{k=1}^{m_i} y_{nik}; \quad x \in \{0,1,2,...,m_i\}$ be the random variable defined by the sum of the successes across the thresholds for item *i*. Within a Guttman subspace, in which the thresholds are ordered, the total score recovers each pattern completely. The possible values of this variable are shown in Table 6, which shows another outcome space $\Omega'_{0,1}$ to be defined shortly.

An important advantage of taking responses only in the Guttman structure is that a response in a category can be inferred. For example, the response pattern of $x = 0 \equiv (0,0,0)$ implies being unsuccessful at P, and at C and at D. Therefore the response pattern implies the category of F. The response pattern $x = 1 \equiv (1,0,0)$ implies being successful at P, but unsuccessful at C and at D. Therefore this response pattern implies the category of P, and so on. The inferred response category from the Guttman pattern and each total score are shown in Table 6. We return to this inference when

completing the reverse direction in the derivation, that from the dependent response data as in Table 3 to the inferred independent response data in Table 4.

Table 6

Inferred category and the total score $X_{ni} = x$	y_{ni1}	y_{ni2}	y_{ni3}	
0 Fail	0	0	0	Ω'
1 Pass	1	0	0	$\Omega'_{0,1}$
2 Credit	1	1	0	
3 Distinction	1	1	1	

The sum of the probabilities of the Guttman patterns within the full outcome space Ω is not 1. It is given by

$$D = Q_{ni1}Q_{ni2}Q_{ni3}...Q_{nim_i} + P_{ni1}Q_{ni2}Q_{ni3}...Q_{nim_i} + P_{ni1}P_{ni2}Q_{ni3}...Q_{nim_i} + ...P_{ni1}P_{ni2}P_{ni3}...P_{nim_i} \quad (10)$$

Then the probability of each of the Guttman patterns, *conditional on the outcome space* Ω', is given by simply normalising the responses within Ω' so that their probabilities do sum to 1, that is

$$\Pr\{X_{ni} = x \mid \Omega'\} = P_{ni1}P_{ni2}P_{ni3}....P_{nix}Q_{nix+1}Q_{nix+2}...Q_{nim_i} / D, \quad (11)$$

where D is the normalising factor ensuring that $\sum_{x=0}^{m_i} \Pr\{X_{ni} = x \mid \Omega'\} = 1$. Although constructed simply as a normalising factor, because it contains all the terms of the numerator, it is important to the operation of the PRM. It ensures that the probability of a response in any category is a function of the probabilities of all of the other categories. Thus through the constraint of the Guttman structure, we have a probability structure that is compatible with that of Table 3 in which the response in any category cannot be independent of the response of any other category, there being only one response in one of the categories.

This line of reasoning, beginning with an independent outcome space at thresholds, and then restricting the outcome space to the Guttman structure, I used (Andrich, 1978a) in resolving the coefficients of the PRM presented in Rasch (1961) and Andersen (1977), and applied it to a case to demonstrate an interpretation of reversed thresholds in the

context of contingency tables (Andrich, 1979). In this Report we take the understanding of that interpretation a stage further.

In anticipation of this stage, we summarise the development so far: we have (i) reduced an outcome space of m_i experimentally independent dichotomous items to a subspace which has the Guttman structure, (ii) observed that the responses in this space parallel the responses in one of just $m_i + 1$ ordered categories as in the general formats of Table 1 and the specific format of Table 3, and (iii) recognised that the probability of a response in any category is a function of the probability of the response in all other categories.

3.3 The doubly conditioned outcome space $\Omega'_{x-1,x}$

Consider now an even smaller outcome space, $\Omega'_{x,x-1}$, a subspace of the Guttman space Ω'. It is the response space corresponding to two successive total scores within the Guttman subspace. This outcome space is shown in Table 6 for $x = 1$.

We take the ratio of the probabilities of the outcomes within this doubly condition outcome space $\Omega'_{x-1,x}$:

$$\frac{\Pr\{X_{ni} = x \mid \Omega'\}}{\Pr\{X_{ni} = x-1 \mid \Omega'\}} = \frac{P_{ni1}P_{ni2}P_{ni3}....P_{nix}Q_{nix+1}Q_{nix+2}...Q_{nim_i}/D}{P_{ni1}P_{ni2}P_{ni3}....P_{nix-1}Q_{nix}Q_{nix+1}Q_{nix+2}...Q_{nim_i}/D} = \frac{P_{nix}}{Q_{nix}}. \qquad (12)$$

This ratio is simply the odds of a response x rather than x-1, giving the probability of a response in category x within the space $\Omega'_{x-1,x}$ as

$$\frac{\Pr\{X_{ni} = x \mid \Omega'\}}{\Pr\{X_{ni} = x-1 \mid \Omega'\} + \Pr\{X_{ni} = x \mid \Omega'\}} = \frac{P_{nix}}{P_{nix} + Q_{nix}} = P_{nix} \qquad (13)$$

Notating explicitly the outcome space $\Omega'_{x-1,x}$, Eq. (13) may be expressed as

$$P_{nix} = \Pr\{X_{ni} = x \mid \Omega'_{x-1,x}\} = \frac{\Pr\{X_{ni} = x \mid \Omega'\}}{\Pr\{X_{ni} = x-1 \mid \Omega'\} + \Pr\{X_{ni} = x \mid \Omega'\}} \quad (14)$$

The probability of Eq. (14) implies a dichotomous response. Clearly, there is no observed experimentally independent dichotomous response within the Guttman subspace. Therefore this response is *implied* or *latent*. Being latent, however, does not make it any less important. Furthermore, being located in the higher of two ordered categories can be considered the implied successful response relative to the two categories, and the implied response in the lower of the two categories an unsuccessful response. Then the probability of Eq. (14) is the probability of being successful in adjacent categories within the Guttman subspace Ω'.

The remarkable result in Eq. (14) is that the *value* P_{nix} in the doubly conditioned sample space $\Omega'_{x-1,x}$, which is the probability of the successful responses between the two adjacent categories, is exactly the probability of *success* at the corresponding category in the experimentally independent outcome space Ω, that is,

$$\Pr\{X_{ni} = x \mid \Omega'_{x-1,x}\} = \Pr\{Y_{nix} = 1 \mid \Omega\} = P_{nix}. \quad (15)$$

In Section 4 where we show the results of simulation studies according to the above theory, we will demonstrate that the estimates from the independent outcome space Ω and the outcome space Ω' when the corresponding model to each outcome space is applied, are indeed statistically equivalent. That is, data from a subspace, when analysed by the model that corresponds to the structure of that subspace, gives statistically equivalent estimates for the thresholds to those obtained from the whole space when analysed according to the model for independent responses.

It is the result of Eq. (15) which provides unique possibilities for studying the empirical ordering of categories from a response structure that has experimental dependence to one in which experimental independence is implied. To do so unequivocally, we need now to proceed in the reverse direction from that in which we have proceeded so far in this Section. That is, since the responses in a format system such as that of Tables 1

and 3 are not from experimentally independent dichotomous outcomes, but have only one response in one of only m_i categories, we need to proceed from this response format. However, we are given the cue as to how to proceed from the above derivation in proceeding from the experimentally independent response format to a restricted dependent one.

3.4 Identification of a outcome space from the definition of success between two adjacent categories in a non-experimental design

The notation we use in proceeding in the reverse direction is identical to that used in the above subsections. However, it is stressed that these derivations begin mathematically in their own right.

Thus let $X_{ni} = x,\ x \in \{0,1,2,...,m_i\}$ be the random variable that denotes the response in one of $m_i + 1$ successive categories. This integer variable denotes only order, with successive integers corresponding to successive empirical categories, there being no implication that the sizes of the categories are in any sense the same.

Further, let $\Pr\{X_{ni} = x\}$ be the probability of the response x: $\sum_{x=0}^{m_i} \Pr\{X_{ni} = x\} = 1$.

Now define

$$\frac{\Pr\{X_{ni} = x\}}{\Pr\{X_{ni} = x-1\} + \Pr\{X_{ni} = x\}} = P_{nix} \qquad (16)$$

and $Q_{nix} = 1 - P_{nix}$.

Clearly the response *implied* by Eq. (16) is dichotomous relative to the two categories $x-1$ and x. The response is implied or latent because there is no observed response of the kind - there is only one response in one of the $m_i + 1$ categories. The outcome space of this latent response is denoted $\Omega'_{x-1,x}$ so that Eq. (16) can be written more explicitly as

$$\Pr\{X_{ni} = x \mid \Omega'_{x,x-1}\} = P_{nix}. \quad (17)$$

Because there is only one response in one of the categories, and because the ordering of the categories with x reflecting the higher of the two categories, P_{nix} can be considered the conditional probability of an implied *successful response* between category *x* and $x-1$. Then $Q_{nix} = 1 - P_{nix} = \Pr\{X_{ni} = x-1 \mid \Omega'_{x,x-1}\}$ is the probability of an *unsuccessful* response between the same two categories.

Wright and Masters (1982) began there derivation of the PRM from Eq. (16), but did not make explicit the outcome space and introduced the dichotomous Rasch model for P_{nxi} prematurely to be able to uncover the full significance of this starting point. It also led them into errors of interpretation (Andrich, 2005).

To begin the process of modelling the observed response in one of the $m_i + 1$ ordered categories in terms of the latent response between a pair of categories, we need to obtain a statement for $\Pr\{X_{ni} = x\}$ and understand how it might hold. These implied responses, as indicated earlier cannot be independent, and so cannot be assumed to hold in an experimentally independent sample space which would have an outcome space of 2^{m_i} elements.

In particular, we need to *infer* an outcome space, if any, for $X_{ni} = x, x \in \{0,1,..m_i\}$. Let this outcome space be Ω'. Making it explicit in Eq. (16) gives Eq. (18)below.

That is, let $P_{nix} = \dfrac{\Pr\{X_{ni} = x \mid \Omega'\}}{\Pr\{X_{ni} = x-1 \mid \Omega'\} + \Pr\{X_{ni} = x \mid \Omega'\}}$ (18)

from which $Q_{nix} = 1 - P_{nix}$.

To stress, we begin with Eqs. (16) and (17) in which a subspace $\Omega'_{x,x-1}$ is defined which contains two adjacent response categories. Now we need to infer a subspace for the response $X_{ni} = x$ which we denote Ω'.

Let $\pi_x \equiv \Pr\{X_{ni} = x \mid \Omega'\}$ for convenience in the derivation in which the subscripts *n* and *i* are dropped, it being understood that the response is that of a single person to a single item. Similarly, we drop the same subscripts in the probability statements for convenience and let $P_{nix} = P_x$ and $Q_{nix} = Q_x$.

Then $P_x = \dfrac{\pi_x}{\pi_{x-1} + \pi_x}$. (29)

From Eq. (29), we need to derive $\pi_x \equiv \Pr\{X_{ni} = x \mid \Omega'\}$ in terms of P_{nix} and Q_{nix}.

From Eq. (29),

$P_x(\pi_{x-1} + \pi_x) = \pi_x$, $\pi_x(1 - P_x) = \pi_{x-1}P_{nix}$, that is $\pi_x Q_{nix} = \pi_{x-1}P_{nix}$, giving

$$\pi_x = \pi_{x-1}\frac{P_x}{Q_x}. \tag{20}$$

Beginning with $\pi_x, x = 1$, the recursive relationship

$$\pi_x = \pi_0 \frac{P_1}{Q_1}\frac{P_2}{Q_2}\frac{P_3}{Q_3}\cdots\frac{P_x}{Q_x} = \pi_0 \prod_{k=1}^{x}\frac{P_k}{Q_k} \tag{21}$$

follows. However $\sum_{x=0}^{m_i} \pi_x = 1$; therefore $\pi_0 = \dfrac{1}{1 + \sum_{x=1}^{m_i}\prod_{k=1}^{x}\frac{P_k}{Q_k}}$, and substituting for π_0 in Eq. (21)

gives $\pi_x = \dfrac{\prod_{k=1}^{x}\frac{P_k}{Q_k}}{1 + \sum_{x=1}^{m}\prod_{k=1}^{x}\frac{P_k}{Q_k}}$. (22)

That is, in full,

$$\pi_x = \frac{\frac{P_1}{Q_1}\frac{P_2}{Q_2}\frac{P_3}{Q_3}\cdots\frac{P_x}{Q_x}}{1+\frac{P_1}{Q_1}+\frac{P_1}{Q_1}\frac{P_2}{Q_2}+\frac{P_1}{Q_1}\frac{P_2}{Q_2}\frac{P_3}{Q_3}+\ldots\ldots\frac{P_1}{Q_1}\frac{P_2}{Q_2}\frac{P_3}{Q_3}\ldots\frac{P_m}{Q_m}}, \quad (23)$$

which on simplification

gives $\pi_x = P_1P_2P_3....P_xQ_{x+1}Q_{x+2}...Q_m / D$, where

$$D = Q_1Q_2Q_3...Q_m + P_1Q_2Q_3...Q_m + P_1P_2Q_3....Q_m + ...P_1P_2P_3...P_m, \quad (24)$$

that is , in full

$$\Pr\{X_{ni} = x \mid \Omega'\} = P_{ni1}P_{ni2}P_{ni3}....P_{nix}Q_{nix+1}Q_{nix+2}...Q_{nim_i} / D. \quad (25)$$

Eq. (25), which gives the probability of a response in any category x, implies a successful or an unsuccessful response between *every* pair of successive categories (and not just, for example, the first *x* categories). Further, it implies a particular structure of successful and unsuccessful responses for each response $X_{ni} = x$. This structure, which should be obvious, is formalised for completeness below.

Let $Y_{nix}, x = 1,2,...m_i$ be a sequence of dichotomous random variables, $Y_{nix}, y; y \in \{0,1\}$ which correspond to the successes and failures implied in the right side of Eq. (25), and let $\Pr\{Y_{nix} = 1\} = P_{nix}; Q_{nix} = 1 - P_{nix}$. At this point we specify nothing about the independence or dependence of these responses.

We show now that if Eq. (25) holds, where $D > 0$ is a real valued constant, then the response vector of the sequence of the random variables $Y_{nix}, x = 1,2,...m_i$ is

compatible with the form $\{Y_{ni1}, Y_{ni2}, Y_{ni3},....Y_{nix}, Y_{nix+1}Y_{nix+2},....Y_{nim_i}\} = \{\underbrace{1,1,1,...1}_{x},\underbrace{0,0,0...,0}_{m_i - x}\}$

in which a sequence of *x* 1s is followed by a sequence of $(m_i - x)$ 0s.

If $\Pr\{X_{ni} = x \mid \Omega'\} = P_{ni1}P_{ni2}P_{ni3}....P_{nix}Q_{nix+1}Q_{nix+2}.....Q_{nim_i} / D$,

then according to the definition of the random variable $Y_{nix} = y, y \in \{0,1\}$ above,

$$\Pr\{X_{ni} = x \mid \Omega'\} =$$
$$[\Pr\{Y_{ni1} = 1\}\Pr\{Y_{ni2} = 1\}...\Pr\{Y_{nix} = 1\}\Pr\{Y_{nix+1} = 0\}\Pr\{Y_{nix+2} = 0\}...\Pr\{Y_{nim_i} = 0\} \mid \Omega']/D \ .$$

Thus $X_{ni} = x$, which arises with the pattern of probabilities of success and failure in Eq. (25), is consistent with the vector of responses

$$(Y_{ni1}, Y_{ni2}, Y_{ni3}, ...Y_{nix}, Y_{nix+1}, Y_{nix+2}, ...Y_{nim_i}) = \{\underbrace{1,1,1,...1}_{x}, \underbrace{0,0,0...,0}_{m_i - x}\} \ .$$

This implies that the outcome space Ω' which was to be inferred from Eq. (18) is the Guttman structure. Only $m_i + 1$ such patterns are possible, which is exactly the number of response categories. It must be stressed that this Guttman outcome space Ω' is *latent*, not a manifest space. There is no observed response between any adjacent category. It is also inferred as the outcome space starting with Eq. (18).

3.5 Inferring an experimentally independent outcome space Ω

Given the Guttman space Ω', we infer the existence of a complete space Ω of which Ω' is a subspace. In this complete space we can infer experimentally independent responses. Thus from the construction of Eq. (25), it is clear that

$$\sum_{x=0}^{m_i} \Pr\{X_{ni} = x \mid \Omega'\} = \sum_{x=0}^{m_i} P_{ni1}P_{ni2}P_{ni3}....P_{nix}Q_{nix+1}Q_{nix+2}.....Q_{nim_i} / D = 1, \qquad (26)$$

and therefore that

$$\sum_{x=0}^{m_i} P_{ni1}P_{ni2}P_{ni3}....P_{nix}Q_{nix+1}Q_{nix+2}.....Q_{nim_i} = D \neq 1 \ . \qquad (27)$$

Now consider all 2^{m_i} patterns of responses, $(Y_{ni1}, Y_{ni2}, Y_{ni3}, ...Y_{nix}, Y_{nix+1}Y_{nix+2}, ...Y_{nim_i})$ in which $Y_{nix} = y \in \{0,1\}$ of which the $m_i + 1$ Guttman patterns $\{\underbrace{1,1,1,...1}_{x},\underbrace{0,0,0...,0}_{m_i - x}\}$ are a subspace. Let this outcome space be denoted Ω. Then by the definition of Y_{nix} and Eq. (27), it can be shown readily that the probabilities of the set of all possible 2^{m_i} patterns
$\{(y_{ni1}, y_{ni2}, ...y_{nix}, ...y_{nim_i}) \mid \Omega\}$ sum to 1, that is

$$\sum_{\Omega} \Pr\{(y_{ni1}, y_{ni2}, ...y_{nix}, ...y_{nim_i})\} = \sum_{\Omega} \prod_{x=1}^{m_i} P_{nix}^{y_{nixi}} Q_{nix}^{1-y_{nix}} = 1. \quad (28)$$

However, Eq. (28) implies experimental independence of responses in the sense that

$$\Pr\{(y_{ni1}, y_{ni2}, ...y_{nix}, ...y_{nim_i})\} = \Pr\{y_{ni1}\}\Pr\{y_{ni2}\}....\Pr\{y_{nim_i}\}. \quad (29)$$

That is, we reason that given that there is a Guttman outcome space Ω' with the $m_i + 1$ response patterns, and with the probability structure of Eq. (25), it is the *subspace* of an experimentally independent outcome space Ω with 2^{m_i} response patterns.

In summary, the *values* of the probabilities of successes $P_{nix}, x = 1,2,3...m_i$, which hold in the outcomes space $\Omega'_{x-1,x}$ for two successive categories within the outcome space Ω', holds in the inferred experimentally independent and complete space Ω. This means that if the thresholds can be estimated from the Guttman sample space Ω' according to Eq. (25), then the values of the thresholds can be inferred to be statistically equivalent to those in an experimentally independent outcome space Ω *compatible* with Ω'.

In this inferred experimentally independent outcome space, as we argued in the previous section, we would require our axiom for ordered categories to hold. That is, we would require that difficulties of the thresholds are in the natural order of the categories.

It is stressed that if one set of data was collected experimentally according to the Design of Table 4, and another set was collected according to the ordered response categories in the usual design of Table 3 for the same essays and the same definition of the categories, nothing in the above analysis would guarantee that the thresholds would be the same from the two data collections. Whether or not they give equivalent results in any particular assessment is an empirical question. It might be interesting to conduct such experiments. However, the analysis implies that there is a hypothetical experimentally independent outcome space Ω of dichotomous random variables with inferred probabilities of which the Guttman outcome subspace Ω' is a subspace. In any complete space Ω, the ordering of the threshold difficulties needs to be in the natural order of the categories if the categories are working empirically as intended.

4. Construction and interpretation of the PRM

The above analysis of outcome spaces has not rested on the dichotomous RM. However, our earlier specification in Section 2 was that the latent thresholds should have the same relative difficulties irrespective of the location of essays on the continuum, and with the higher ordered category reflecting a greater location of the essays on the continuum. This, we concluded, implied the responses should conform to the dichotomous RM. Now we bring the dichotomous RM and the analysis of outcome spaces in the previous section together.

4.1 Specifying the dichotomous RM at the latent dichotomous response between successive categories

Because the latent Guttman structure arises as the outcome space irrespective of the starting point of the derivation, we begin with Eq. (25), that is,

$$\Pr\{X_{ni} = x \mid \Omega'\} = P_{ni1}P_{ni2}P_{ni3}....P_{nix}Q_{nix+1}Q_{nix+2}...Q_{nim_i} / D . \quad (30)$$

Recall this is the conditional implied dichotomous response between categories $x-1$ and x, where the latter response is the implied success.

$$\text{Let } P_{nix} = \frac{e^{\beta_n - \delta_{ix}}}{1 + e^{\beta_n - \delta_{ix}}} \tag{31}$$

and $Q_{nix} = 1 - \frac{e^{\beta_n - \delta_{ix}}}{1 + e^{\beta_n - \delta_{ix}}} = \frac{1}{1 + e^{\beta_n - \delta_{ix}}}$.

According to the dichotomous RM of Eq. (2) with the threshold δ_{ix} made explicit.

Inserting Eq. (31) in Eq. (30) gives

$$\Pr\{X_{ni} = x \mid \Omega'\} = \frac{e^{\beta_n - \delta_{i1}}}{1 + e^{\beta_n - \delta_{i1}}} \frac{e^{\beta_n - \delta_{i2}}}{1 + e^{\beta_n - \delta_{i2}}} \cdots \frac{e^{\beta_n - \delta_{ix}}}{1 + e^{\beta_n - \delta_{ix}}} \frac{1}{1 + e^{\beta_n - \delta_{ix+1}}} \cdots \frac{1}{1 + e^{\beta_n - \delta_{im_i}}} / D$$

which, on making D explicit, simplifies to

$$\Pr\{X_{ni} = x \mid \Omega'\} = e^{x\beta_n - \sum_{k=0}^{x} \delta_{ik}} / \gamma_{ni} \tag{32}$$

where $\delta_{k0} \equiv 0$ and is used for notational convenience, and

$$\gamma_{ni} = \sum_{x=0}^{m_i} e^{x\beta_n - \sum_{k=0}^{x} \delta_{ik}} \tag{33}$$

is the normalising factor in Eq. (32) ensuring that the sum of its probabilities is 1.

Eq. (32) is a general form of the PRM.

4.2 Other parameterisations

Because the thresholds are identified with each item, it is often convenient to consider the thresholds of an item as deviations from its overall location. Let $\delta_{ik} = \delta_i + \tau_{ik}$ where $\sum_{k=0}^{m_i} \tau_{ik} = 0$. Then δ_i is the mean of the thresholds δ_{ik} and τ_{ik} are thresholds which are deviations from δ_i. This gives Eq. (32) in the form

$$\Pr\{X_{ni} = x \mid \Omega'\} = e^{x\beta_n - x\delta_i - \sum_{k=0}^{x} \tau_{ik}} / \gamma_{ni} = e^{x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_{ik}} / \gamma_{ni}. \quad (34)$$

This is also a convenient form of the model for estimation of the threshold parameters which can be reparameterised into principal components (Andrich and Luo, 2003) of which δ_i is the first principal component. Further, this reparameterisation permits estimates of thresholds to be obtained even when some categories have 0 frequencies.

Because the total score $r_n = \sum_{i=1}^{I} x_{ni}$ is the sufficient statistic for the person parameter β_n (Andersen, 1977; Andrich, 1978a), this estimation can be carried out conditionally, eliminating the person parameters. As indicated earlier, we do not deal in this Report with estimation and tests of fit, and concentrate on the hypothesis of the empirical ordering of the categories as manifested in the threshold estimates, and the understanding of the operation of the model in relation to these threshold estimates. However, we note that the thresholds are estimated independently of the person parameters, and therefore independently of any distribution of these parameters. This means that the threshold estimates *reflect a structural relationships amongst each other* and not the distribution of the person parameters.

Eq. (34) is also convenient if it is hypothesised that the distances between thresholds might be the same across items, with the only difference among the item parameters being the location of the item determined by the mean of the thresholds. The thresholds are then not subscripted by the item parameter giving

$$\Pr\{X_{ni} = x \mid \Omega'\} = e^{x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_k} / \gamma_{ni}.$$

(35)

It is this parameterisation which is sometimes referred to as the *rating scale model*, while it is parameterisation of Eq. (32) which is sometimes referred to as *partial credit model*. It should be clear that the response model is the same in Eqs. (35) and (32), and that the only difference is in the parameterisation. Whether the thresholds are

equidistant among items is again an empirical question, not a response process question. We return to the question of the process later in the Report.

An important observation from Eq. (32), which reflects the dependence among the responses, is that the probability of a response in any category $\{X_{ni} = x \mid \Omega'\}$ is a function of all the thresholds, and not just of an adjacent pair defining a category. This is easily seen from the denominator $\gamma_{ni} = \sum_{x=0}^{m_i} e^{x\beta_n - \sum_{k=0}^{x} \delta_{ik}}$ in Eq. (32). The denominator contains all thresholds, so a change of the value of any one threshold will change the probability of a response in every category. In particular, the probability of a response in the first category is a function of the last threshold. This confirms the structural dependence in the responses among categories.

4.3 Identity of the dichotomous RM in the full Ω space and the Guttman space Ω'

For completeness, we now summarise the relationships among the parameters of the PRM according to Eq. (15). Inserting the dichotomous RM for P_{nix}, we have

$$\Pr\{X_{ni} = x \mid \Omega'_{x-1,x}\} = \Pr\{Y_{nix} = 1 \mid \Omega\} = \frac{e^{\beta_n - \delta_{ix}}}{1 + e^{\beta_n - \delta_{ix}}}. \qquad (36)$$

Thus using the example of grades awarded in Figure 1 and Table 4, the explicit relationships between the responses in the outcome spaces and the parameters are shown in Table 7. In particular, the last row of Table 7 shows that *the threshold δ_{ix}, at which the probability of a successful response is 0.5 at category $x, x > 0$ in the experimentally independent outcome space Ω, is identical to the threshold $\delta_{ix-1,x}$ at which the probability of a successful response $x > 0$ relative to the adjacent category $x - 1$ is 0.5 in the constrained outcome space $\Omega'_{x-1,x}$.*

Table 7 Equivalences of corresponding thresholds in the spaces Ω and $\Omega'_{x-1,x}$

Outcome space	Response	$x = P$	$x = C$	$x = D$
$\Pr\{Y_{nix} = 1 \mid \Omega\}$		$\frac{e^{\beta_n - \delta_{iP}}}{1 + e^{\beta_n - \delta_{iP}}}$	$\frac{e^{\beta_n - \delta_{iC}}}{1 + e^{\beta_n - \delta_{iC}}}$	$\frac{e^{\beta_n - \delta_{iD}}}{1 + e^{\beta_n - \delta_{iD}}}$
	=			
$\Pr\{X_{ni} = x \mid \Omega'_{x-1,x}\}$		$\frac{e^{\beta_n - \delta_{iFP}}}{1 + e^{\beta_n - \delta_{iFP}}}$	$\frac{e^{\beta_n - \delta_{iPC}}}{1 + e^{\beta_n - \delta_{iPC}}}$	$\frac{e^{\beta_n - \delta_{iCD}}}{1 + e^{\beta_n - \delta_{iCD}}}$
		$\delta_{iP} = \delta_{iFP}$	$\delta_{iC} = \delta_{iPC}$	$\delta_{iD} = \delta_{iCD}$
		$\{\delta_{ix} \mid \Omega\} = \{\delta_{ix-1,x} \mid \Omega'\}$		

4.4 Analysis of data in the different sample spaces according to corresponding models.

In order to consolidate the theoretical analysis of the outcome spaces and the way the PRM plays itself out, this section shows analyses of data in the different outcome spaces according to the corresponding model. The data are simulated.

Simulation 1

Table 8 summarises the first simulated data set. There are 6 independent dichotomous items simulated according the dichotomous RM. The number of persons simulated from a normal distribution, $\beta \approx N(0, 2)$, was 5000. The items are conceptualised as an experimental design such as that in Table 4 arising from a response format in four ordered categories as in Table 3. There are 6 independent judges where, for example, the first set of three judges might be novices and the second set of three judges might be experts. Each judge makes an independent dichotomous decision at each threshold on the same objects of measurement independently of any other judge. Rates of success at the successive threshold decrease; this is reflected in the increasing level of difficulty of the generated location of successive items in each set.

Table 8

Simulation 1 estimates of independent dichotomous ordered items using the full space and the dichotomous RM and the Guttman subspace and the PRM

Judges	Item number	Generated locations	Estimates from the dichotomous RM in a full space	Estimates from the PRM in a Guttman space
Novice 1 (P)	1	-1.1	-1.125	-1.047
Novice 2 (C)	2	-0.1	-0.119	-0.133
Novice 3(D)	3	1.2	1.267	1.196
Expert 1(P)	4	-1.7	-1.685	-1.680
Expert 2 (C)	5	0.3	0.227	0.232
Expert 3 (D)	6	1.4	1.435	1.433
Extreme scores eliminated			1214	1214
Non-Guttman patterns eliminated				1294
Number used			3786	2492
		RMSQ	0.018	0.017
			Fit	χ^2=6.081, df=4 P <0.193

N=5000; $\beta \approx N(0, 2)$

Table 8 also shows two sets of estimates of the item parameters. The first set of estimates was obtained simply from the dichotomous RM of Eq. (2) with independent responses and provides a frame of reference for the other set of estimates. The method of pairwise conditional estimation introduced in Choppin (1968), and elaborated for the PRM in Andrich and Luo (2003), was used. This method provides consistent estimates of the item parameters (Zwinderman, 1995). In this method, as in most methods, persons with extreme total score across items provide no information on the relative difficulties of the thresholds and are discarded from the item parameter estimation. The number retained in the estimation for the dichotomous RM is shown in Table 8.

The second set of estimates was obtained by *retaining only people whose response pattern followed the Guttman structure in both of the sets of three items, that of the novices and the experts*. First, persons with all *non* Guttman patterns within each set of 3 items were eliminated; second, two items were formed by summing the responses of each of the sets of three items. This gave two polytomous items with scores

ranging from 0 to 3. Then the estimates were obtained by applying the PRM of Eq. (32). That is, the model arising from the Guttman constraint on the independent responses, which accounts for the dependencies in the responses across the thresholds, was applied to data which arose from exactly that subset of responses. In this case two sets of persons were eliminated from the estimation, the first set was composed of those people who did not conform to the Guttman structure in either set of 3 dichotomous responses, and the second was composed of those who had extreme scores. These and the numbers retained in the estimation are shown in Table 8. This construction implies that the two sets of estimates were not obtained from exactly the same data - the data in the second analysis is a subset of the data in the first analysis. Table 8 also shows the root mean square deviation between the generating and the estimated values in each data set and corresponding model - $RMSD = \sqrt{\sum_{i=1}^{I^1} (\delta_i - \hat{\delta}_i)^2}$.

It is evident that these values for the two estimates are virtually identical, even though as expected the estimates themselves are not identical. The Table also shows a χ^2 test of fit based on the observed and expected frequencies, in each possible pair of responses (x_i, x_j) conditional on the total score for the analysis with the PRM. The degrees of freedom are $(m_i - 1)(m_j - 1) = 4$. Clearly, the data fit the model. There is no need to show fit for the independent dichotomous RM estimates as these fit the model by definition.

This simulated data set is not intended to be an exhaustive study of the quality of the estimates. Instead, it is intended to illustrate and make concrete the operation of the theoretical analysis of sample spaces and the application of the Rasch model when proceeding from an experimentally independent empirical design in an outcome space in which all data are available, to a subset of this outcome space which conforms to the Guttman structure subspace. It is evident from Table 8 that, as concluded from the mathematical analysis, the estimates using the two designs and corresponding Rasch models gives statistically equivalent *values* for the thresholds.

Thus the analysis from the data formed from the Guttman subspace as just two items with 4 ordered categories recovers the values of the experimentally independent

thresholds. This confirms that if only the scores from the two polytomous items were available to begin with, and the PRM was applied, then the threshold estimates could be inferred as arising from a compatible experimental design in which the responses were independent.

Figure 3 shows the ICCs and the estimates of the thresholds for the six items as analysed with the dichotomous RM. They are presented as two sets of three items. The first panel is identical to Figure 1. Figure 4 shows the estimates of the thresholds with the data conforming the Guttman subspace analysed with the PRM. It shows the probability of the response in each category $\Pr\{X_{mi} = x \mid \Omega'\}$, referred to as a category characteristic curves (CCCs) as well as the latent dichotomous ICCs of the dichotomous RM at the thresholds.

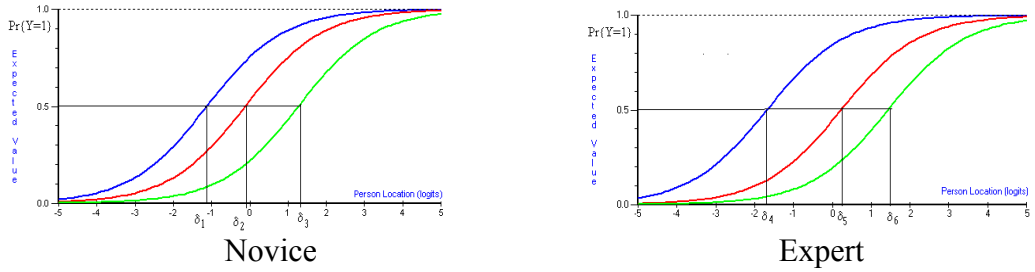


Figure 3. ICCs at threshold estimates for Simulation 1 using the dichotomous RM

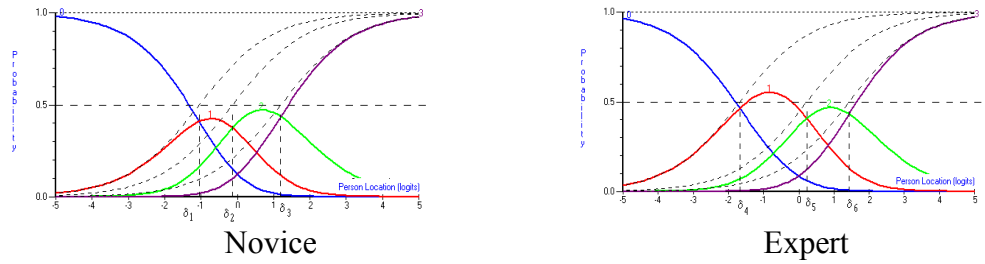


Figure 4 Category characteristic curves for Simulation 1 and latent ICCs at thresholds using the PRM.

Figure 4 shows that the CCCs for the successive categories intersect exactly at the successive thresholds. The latent ICCs at the thresholds, shown in dotted lines, correspond to the ICCs in Figure 3. The thresholds in Figure 4 can be taken to correspond exactly to the thresholds depicted in Figure 1, where the continuum is shown to be partitioned into contiguous categories. The ICCs of the latent dichotomous responses reflect *both* the probability of a successful response in relation to the adjacent pairs of categories within the set of four categories (the Guttman subspace), *and* the probability of a successful response at that threshold in the complete sample space. At the thresholds, $\beta_n = \delta_{ix}$, the probability of being located in one of the two adjacent categories, $x-1$ and x, is the same, 0.5. This is exactly the probability of being successful in the full independent space at the corresponding threshold. Further, for any person with a location value between two successive thresholds, $\delta_{x-1} < \beta_n < \delta_x$, the probability of being located in that category is higher than the probability of being located in any other category.

Consistent with this interpretation, from a response in category $X_{ni} = x$, it can be inferred that the latent response structure at the thresholds is

$(Y_{ni1}, Y_{ni2}, Y_{ni3}, \ldots Y_{nix}, Y_{nix+1} Y_{nix+2}, \ldots Y_{nim_i}) = \{\underbrace{1,1,1,\ldots 1}_{x}, \underbrace{0,0,0\ldots,0}_{m_i - x}\}$ and therefore that a response in category x implies a success at all thresholds up to δ_{ix} and a failure at all thresholds at δ_{ix+1} and beyond. This is exactly analogous to measurement and reflects the powerful constraint of order of categories.

Finally, one more observation is relevant. For any person located between a pair of adjacent thresholds, the distribution of responses among categories is unimodal. For example, consider a person located at $\beta_n = 0.5$ for the Expert group. The probabilities of responses in the successive categories from Eq. (32) are respectively, $\Pr\{X_{mi} = x \mid \Omega'\}, x = 0,1,2,3$, 0.04, 0.34, 0.45 and 0.17. Clearly, $\Pr\{X_{mi} = 2 \mid \Omega'\}$ has the highest probability among the categories, and the other categories have successively lower probabilities on either side. The generality of this result can be

proved readily from Eq. (32). It seems necessary that the responses that are to be characterised by a single location, β_n, should be unimodal. A bimodal distribution for the responses in supposedly ordered categories would suggest that a characterisation by a single number β_n would be misleading.

Simulation 2

We now turn to the case of reversed thresholds, that is, where in the experimentally independent response design as in Table 4, and for some reason or other, the thresholds are reversed relative to the natural order of the categories.

Simulation 2 begins with exactly the same data as Simulation 1, but the responses are changed in their order. For the Novice responses in Table 8, items 2 and 3 are exchanged, and for the experts, items 1 and 2 are exchanged. Table 9 shows the details analogous to those in Table 8 in the recovery of the estimates.

Table 9
Simulation 2 estimates of independent dichotomous disordered items using the full space and the dichotomous RM and the Guttman subspace and the PRM

Judges	Item number	Generated locations	Estimates from the dichotomous RM in a full space	Estimates from the PRM in a Guttman space
Novice 1 (P)	1	-1.1	-1.125	-1.132
Novice 3 (C)	2	1.2	1.267	1.452
Novice 2(D)	3	-0.1	-0.119	-0.169
Expert 2(P)	4	0.3	0.227	0.122
Expert 1 (C)	5	-1.7	-1.685	-1.651
Expert 3 (D)	6	1.4	1.435	1.379
Extreme scores eliminated			1214	1214
Non-Guttman patterns eliminated				2659
Number used			3786	1129
		RMSQ	0.018	0.054
			Fit	χ^2=0.932, df=4 P <0.920

N=5000; $\beta \approx N(0, 2)$

The estimates for the independent responses using the PRM in Tables 7 and 8 are clearly the same, except that the ordering of the items within each set is different in the two tables. However, in Simulation 2, the Guttman patterns arise from a hypothetical ordering of the items different from that in Table 8. This has a manifestation that many less persons are available for analysis when the Guttman patterns only are retained, 1129 rather than 2492. The reason for this reduction is that when the thresholds are ordered, as illustrated in Table 2, the Guttman pattern is the most likely pattern for any given total score; and when the items are not ordered then the pattern taken as the Guttman pattern is not the most likely among the patterns with the same total score. Nevertheless, the estimates from the Guttman subspace according to the PRM again gives relatively accurate estimates of the reversed thresholds in their reversed order, though not quite as accurately as in the one simulation in the ordered case as shown by the RMSD. Of course, the sample size is less than half that available in the Simulation 1 and so less accurate estimates are to be expected and incorrect inferences based on the estimates would be made. Once again, these threshold estimates can be inferred to arise from responses which are experimentally independent. Therefore, when they are disordered, it can be inferred that they would be disordered in the compatible outcome space in which the responses are experimentally independent. The Table also shows the same χ^2 test of fit as in Table 8 for the analysis with the PRM. Clearly, the data fit the model, even over fit the model according to this criterion of predicting the observed frequencies given the parameters. We return to this observation in the Summary. In anticipation, we note that this kind of fit is not relevant to the evidence of thresholds order. Again, there is no need to show fit for the independent dichotomous RM estimates as these fit the model by definition.

Despite the mathematical logic of Eqs (15) and (32) that predicts the estimation would work as it does, it seems almost startling that the model recovers the thresholds estimates even in the case when the thresholds are reversed and when the Guttman subspace is not based on the actual threshold order but on a supposed threshold order which may or may not be correct. The reason it does this is that the model requires a Guttman structure, and it estimates thresholds in such a way that the implied Guttman structure in the data is recovered (Luo, 2005). Thus in Simulation 2, the estimates tells

that the implied Guttman order of locations for the Novice judge is threshold 1, then 3 then 2, and for the Expert judge, it is threshold 2, then 1 then 3. This tells, for example, that a score of 2 for the Novice judge arises from latent successes at thresholds 1 and 3, with a failure at 2, that is, (1,0,1); and that a score of 1 for the Expert judge arises from implied successes at threshold 2 and failures at the other two thresholds, that is (0,1,0). This is clearly contradictory to the interpretation of the manifest category order. However, although contradictory in terms of the intended ordering, and to stress the theme of the Report, the threshold estimates are a property of the data as revealed by the PRM.

The results have consistent interpretations in the respective CCCs. Figures 5 and 6 with estimates from Table 9 parallel Figures 3 and 4 with estimates from Table 8. The only difference between Figures 3 and 5 is that the ordering of the items is different. Figure 5 is presented for completeness and because of its relationship to Figure 6.

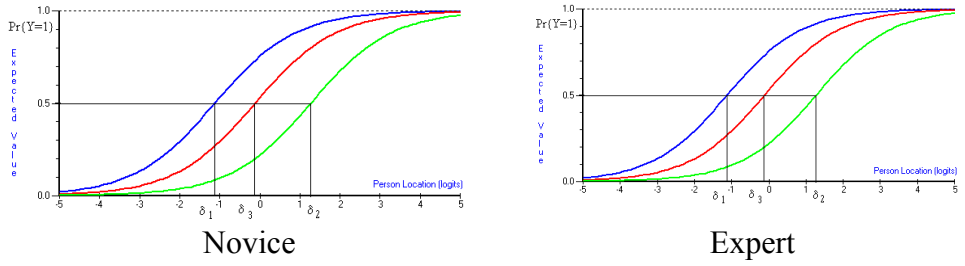


Figure 5. ICCs at threshold estimates for Simulation 2 using the dichotomous RM

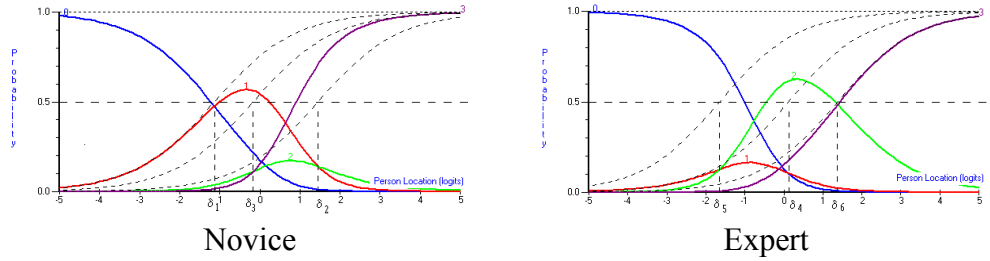


Figure 6 Category characteristic curves for Simulation 2 and latent ICCs at thresholds using the PRM.

The CCCs have a very special configuration in the presence of reversed thresholds, as shown in Figure 6. First, it is evident that the ICCs of the latent dichotomous

responses at the thresholds are equivalent in Figures 5 and 6, reflecting the disorder in the thresholds. Second, in the first panel, where thresholds 2 and 3 are reversed, $\Pr\{X_{mi} = 2 | \Omega'\}$ is never greater than $\Pr\{X_{mi} = 1 | \Omega'\}$ and $\Pr\{X_{mi} = 3 | \Omega'\}$. Indeed, even when $\Pr\{X_{mi} = 2 | \Omega'\}$ is at *its* maximum, simultaneously, $\Pr\{X_{mi} = 1 | \Omega'\}$ and $\Pr\{X_{mi} = 3 | \Omega'\}$ are greater. This seems very strange in a system of ordered categories. In the second panel, where thresholds 4 and 5 are reversed, $\Pr\{X_{mi} = 1 | \Omega'\}$ is never greater than $\Pr\{X_{mi} = 0 | \Omega'\}$ and $\Pr\{X_{mi} = 2 | \Omega'\}$. In this case, even when $\Pr\{X_{mi} = 1 | \Omega'\}$ is at *its* maximum, simultaneously, $\Pr\{X_{mi} = 0 | \Omega'\}$ and $\Pr\{X_{mi} = 2 | \Omega'\}$ are greater. Again, this seems very strange in a system of ordered categories.

To be more concrete, suppose that the scores $0,1,2,3$ correspond to the classifications of Fail, Pass, Credit and Distinction as shown in Table 4. Panel 1 of Figure 6 implies that there is no region on the continuum where the rating of Credit is most likely. In addition, the person located at the threshold δ_3, which implies having a probability of 0.5 of being successful at Distinction in the unconditional outcome space, and a probability of 0.5 of being successful between Distinction and Credit in the Guttman subspace, has simultaneously a higher probability of obtaining a score of 0(Fail) and a score of 1(Pass). This seems untenable in an ordered category system. Similarly in Panel 2 of Figure 6, there is no region of the continuum where a Pass is most likely. In addition, the person located at the threshold δ_5, which implies having a probability of 0.5 of being successful at Credit in the unconditional outcome space, and a probability of 0.5 of being successful between Pass and Credit in the Guttman outcome space, has simultaneously a higher probability of obtaining a score of 0(Fail). Again, this seems untenable in an ordered category system.

In the case of naturally ordered thresholds, we noted that the response distribution among the categories for a person between two thresholds is unimodal. In the case of disordered thresholds this is not the case. For example, consider a person located at $\beta_n = 0.77$ for the Novice group. The probabilities of responses in the successive

categories from Eq. (32) are respectively, $\Pr\{X_{mi} = x \mid \Omega'\}, x = 0,1,2,3$, 0.05, 0.34, 0.17 and 0.44. This is a bimodal distribution suggesting that a characterisation by a single number β_n would be misleading. For this person, the Expected value $E[X_{mi}] = \sum_{x=0}^{3} X_{ni} = 2$. Thus although the expected value is 2, which corresponds to Credit, the person has a higher probability of both a Pass and a Distinction. Again, this seems untenable in a successfully operating ordered category system

Finally, in Figure 1 and in the two panels of Figure 4, the distance between the successive thresholds is positive, $\delta_{x+1} - \delta_x > 0$ for all x. These values define formally the categories depicted in Figure 1. In both panels of Figure 6, this is not the case. Essentially, a category, category Credit in Panel 1 and category Pass in Panel 2, are not defined as regions on the continuum. The thresholds in Figure 6 do not define regions on the continuum such as those in Figure 1. Again, this is a property of the data. It is not, as inferred by Masters and Wright (1997) a problem with the PRM or the graphs in Figure 6. They reproduce the simulated data accurately.

Taking these observations of Figure 6 in reverse, (i) the undefined region on the continuum, (ii) the bimodal distribution of responses of a person located between a pair of adjacent but reversed thresholds values, (iii) there being no region on the continuum in which a category has a higher probability than any other category, and (iv) that a person who has 0.50 probability of being successful between two adjacent categories has nevertheless a higher probability of being classified in a lower category, are all dysfunctional symptoms of reversed thresholds in a system in which categories are intended to be ordered. They all point to a property of the data which is inconsistent with the intended ordering of the categories. Clearly the parameters are recovered accurately in both cases, even better this time in the reversed threshold case according to the RMSQ criterion.

Simulation 3

For completeness, two data sets are simulated directly according to the PRM of Eq. (32) and estimates recovered according to the same model. Tables 9 and 10 show the

cases of ordered and disordered thresholds respectively. Clearly in these cases we do not have data to estimate in the experimentally independent space Ω .

Table 10
Simulation 3 estimates of ordered latent dichotomous thresholds from the PRM

Judges	Item number	Generated locations	Estimates from the PRM in a Guttman space
Novice 1 (P)	1	-1.1	-1.286
Novice 2 (C)	2	-0.1	0.071
Novice 3(D)	3	1.2	1.242
Expert 1(P)	4	-1.7	-1.768
Expert 2 (C)	5	0.3	0.33
Expert 3 (D)	6	1.4	1.412
Extreme scores eliminated			1485
Number used			3515
		RMSQ	0.109
		Fit	$\chi^2 = 3.385$, df=4 P <0.496

N=5000; $\beta \approx N(0, 2)$

Table 11
Simulation 3 estimates of disordered latent dichotomous thresholds from the PRM

Judges	Item number	Generated locations	Estimates from the PRM in a Guttman space
Novice 1 (P)	1	-1.1	-1.247
Novice 2 (C)	2	1.2	1.364
Novice 3(D)	3	-0.1	-0.108
Expert 1(P)	4	0.3	0.294
Expert 2 (C)	5	-1.7	-1.734
Expert 3 (D)	6	1.4	1.430
Extreme scores eliminated			2018
Number used			2982
		RMSQ	0.092
		Fit	$\chi^2 = 1.734$, df=4 P <0.785

N=5000; $\beta \approx N(0, 2)$

Two further features of the analyses in Tables 9 and 10 are important to appreciate. First, as in all of our examples, the model essentially accounts for the data even though when the thresholds are reversed, the implied Guttman patterns are not the presumed ones (Luo, 2005). This again is reflected in the number of responses available for the estimation, 3515 when the thresholds are ordered and 2982 when they are disordered. In this case, disordered thresholds give more extreme scores which are excluded in the estimation. The point here is that accounting for the data by a model is not sufficient in understanding the data and whether or not the data satisfies important criteria. Sometimes misfit is found in conjunction with disordered thresholds, but with varying powers of the tests of fit in conjunction with different numbers of items, numbers of persons and the relationship between the locations of the item and person parameters, this is not always the case. In the case of intended ordered response categories, and even if it accounts for the data, the PRM can reveal whether or not the data satisfy an empirical ordering of the categories. We return to the point later in the Report.

Second, the recovery of the thresholds within the Guttman subspace Ω' and the complete space Ω operates successfully when the dichotomous responses are modelled by the dichotomous RM. If the responses at the thresholds do not follow the dichotomous RM, then it is not easy to predict the outcome. We return to this point also later in the Report.

4.5 Examples of similar items with ordered and reversed thresholds

Two make concrete the interpretations that can be made from a PRM analysis described above, two items, labelled S121 and S004 taken from the *Space* mathematics strand of the 1992 and 1996 Western Australian Literacy and Numeracy testing programs, are shown in Figure 7. These items have a similar structure, in terms of both the task demand, and in the format of the marking key. Figure 7 shows the results of the analysis according to the PRM.

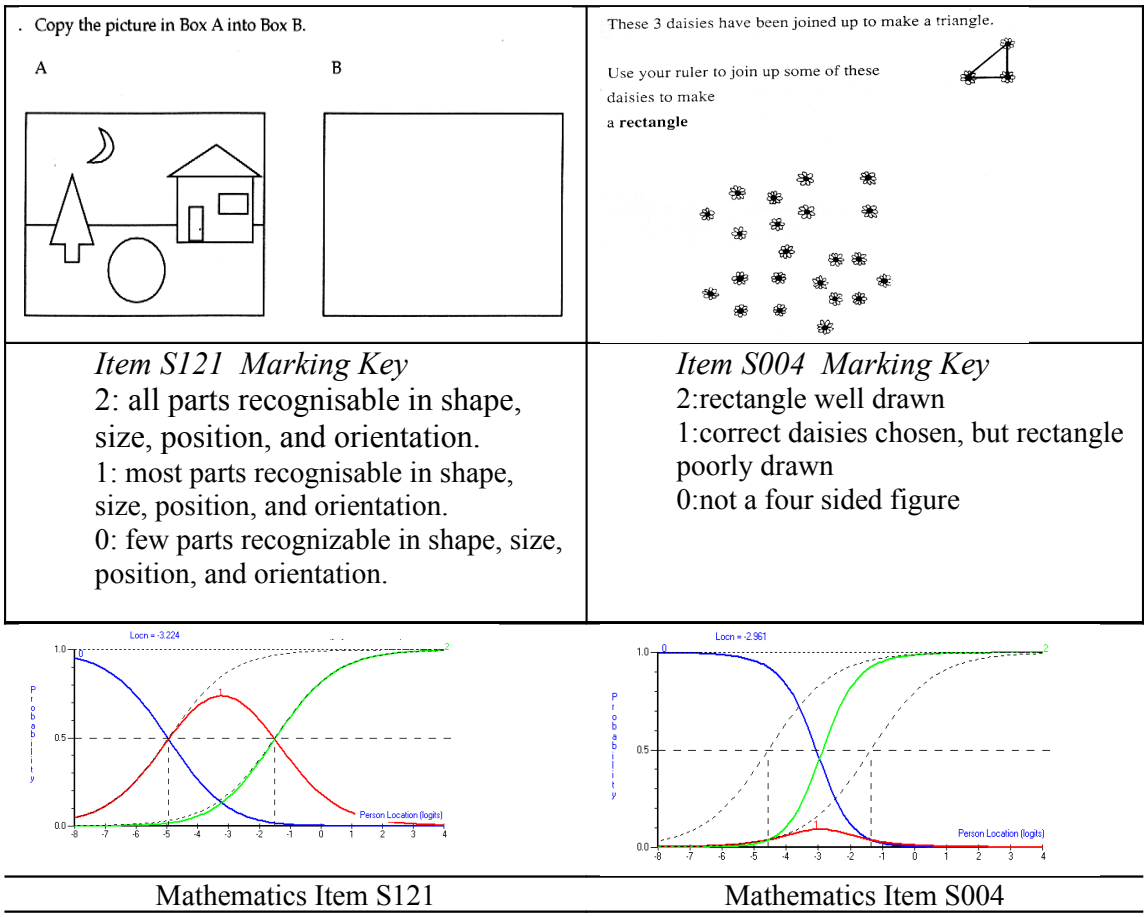


Figure 7 Two mathematics items with ordered and disordered thresholds. Example provided from Van Wyke (2003)

However, the thresholds of S121 are ordered, those of S004 are not. We examine the items now more closely.

The similarity of the structure of the items is that they both require a drawing to be completed, which is marked according to the degree of accuracy with which it is completed. There is, however, an important difference. In S121, no marks are awarded for a diagram that is mostly incorrect, one mark for a diagram which is mostly correct, and two marks for a diagram which is completely correct. For the target student group, in this case Year 3 students, both latter 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 to get some aspects wrong, and so score only 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 the categories are working as intended.

In an apparently similar way, in S004 no marks are awarded for figure mostly incorrect, 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 that a mark of 1 will occur for anyone. 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 nature of a rectangle. Support is provided for drawing the rectangle in the form of the daisies and the students are allowed to use a ruler to connect the daisies. If students know the shape of a rectangle, they will select the set of daisies which form that shape, and join them together with their rulers. Because the vertices are provided, and they can use a ruler, it is unlikely that they will not draw (reasonably) straight sides, or not join the corners. 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, they will join daisies that do not form a rectangle and they will get a mark of zero. As a result, students who, based on their performance on the test as a whole, would be expected to obtain a mark of 1 on this item, are likely to get the item wholly correct, and score two marks, or get it wholly incorrect, and score zero marks. This is the likely explanation for the reversed thresholds.
However, once this feature of the marking scheme is pointed out, even if it were pointed out by an insightful marker and not by the reversed threshold estimates using the PRM, it seems obvious to have a flaw in the ordering of the categories. The marking key is not identifying a continuum of performance in drawing which may have three ordered categories - the item is really about identifying rectangles, and evidently, students either can or cannot identify rectangles. As a result, $\hat{\delta}_{i2} < \hat{\delta}_{i1}$ indicating that making a well drawn rectangle (2 marks) is less difficult than choosing the correct daisies and drawing a poor rectangle (1 mark). The difficulty in the task,

as indicated above, is choosing the correct daisies that form a rectangle, not in the quality of the drawing, especially given that the students can use a ruler.

It is important to appreciate that the reversed thresholds do not arise from their being a low frequency of the score 1 in an absolute or population sense. It arises from persons who would be expected to obtain a score of 1 not obtaining a score of 1. Even if there is a category with zero frequency on an item, and if this is a result of the distribution of persons and not a structural problem with the item, then the thresholds will be in their natural order (Andrich, 2005).

We return to this example of Figure 7 briefly after summarising the GRM.

5. Constructing the PRM beginning with the 2PLM

In order to better understand the PRM and its properties, it is now derived beginning with the 2PLM of Eq. (3) rather than the dichotomous RM in Eq. (2). This is the construction presented in detail in Andrich (1978a). It also helps make more explicit its incompatibility with the GRM.

5.1 Derivation of the PRM using the 2PLM

Inserting the 2PLM of Eq. (3) into Eq. (25) gives

$$\Pr\{X_{ni} = x \mid \Omega'\} = \frac{e^{\alpha_{i1}(\beta_n-\delta_{i1})}}{1+e^{\alpha_{i1}(\beta_n-\delta_{i1})}}\frac{e^{\alpha_{i2}(\beta_n-\delta_{i2})}}{1+e^{\alpha_{i2}(\beta_n-\delta_{i2})}}\cdots\frac{e^{\alpha_{ix}(\beta_n-\delta_{ix})}}{1+e^{\alpha_{ix}(\beta_n-\delta_{ix})}}\frac{1}{1+e^{\alpha_{ix+1}(\beta_n-\delta_{ix+1})}}\cdots\frac{1}{1+e^{\alpha_{im_i}(\beta_n-\delta_{im_i})}}/D. \quad (37)$$

On simplification of the numerator and summarising

$$\Pr\{X_{ni} = x \mid \Omega'\} = \frac{e^{(\alpha_{i1}+\alpha_{i2}+\ldots\alpha_{ix})\beta_n - \sum_{k=1}^{x}\alpha_{ix}\delta_{ix}\alpha_{i1}\delta_{i1}-\alpha_{i2}\delta_{i2}-\ldots\ldots-\alpha_{ix}\delta_{ix}}}{\prod_{k=1}^{m_i}(1+e^{\alpha_{ix}(\beta_n-\delta_{ix})})}/D$$
$$= \frac{e^{\varphi_{ix}\beta_n+\kappa_{ix}}}{\prod_{k=1}^{m_i}(1+e^{\alpha_{ix}(\beta_n-\delta_{ix})})}/D \qquad (38)$$

where $\varphi_{ix} = \alpha_{i1} + \alpha_{i2} + \ldots\alpha_{ix}$, (39)

and $\kappa_{ix} = -\alpha_{i1}\delta_{i1} - \alpha_{i2}\delta_{i2} - \ldots\alpha_{ix}\delta_{ix}$. (40)

The term φ_{ix} was termed by Rasch (1961), who derived the model generically from a vector form of the model, as the scoring function of category x. Andersen (1977) showed that for sufficiency to hold in a unidimensional model, these coefficients had to have the constraint

$$\varphi_{ix} - \varphi_{ix-1} = \varphi_{ix+1} - \varphi_{ix}. \qquad (41)$$

It is evident that if the discriminations at the thresholds are made equal, and defined to be 1 arbitrarily and conveniently, that is, $\alpha_{i1} = \alpha_{i2} = \ldots\alpha_{ix} = 1$, then $\varphi_{ix} = x_i$ is the integer scoring function satisfying Eq. (41) and further that $\kappa_{ix} = -\delta_{i1} - \delta_{i2}\ldots - \delta_{ix}$, giving the PRM of Eq. (32). Thus the integer scoring function $\varphi_{ix} = x_i$ of the successive categories arises from the *identity of discrimination* at the thresholds and not from the equality of distances between thresholds, which of course are estimated. This integer scoring gives sufficiency, with the dichotomous RM a special case.

However, the construction beginning with the 2PLM is instructive regarding combining categories. Andersen (1977) also clarified Rasch's (1966) observation that the probabilities of responses in adjacent categories could not be summed while retaining the structure of the model of Eq. (38) by showing that the probabilities of two adjacent categories $x-1$ and x could only be combined if $\varphi_{ix-1} = \varphi_{ix}$. This abstract conclusion can be understood from Eq. (39) readily, but distinctively, by noting that if

$$\varphi_{ix-1} = \alpha_{i1} + \alpha_{i2} + \ldots \alpha_{ix-1} = \varphi_{ix} = \alpha_{i1} + \alpha_{i2} + \ldots \alpha_{ix-1} + \alpha_{ix} \quad (42)$$

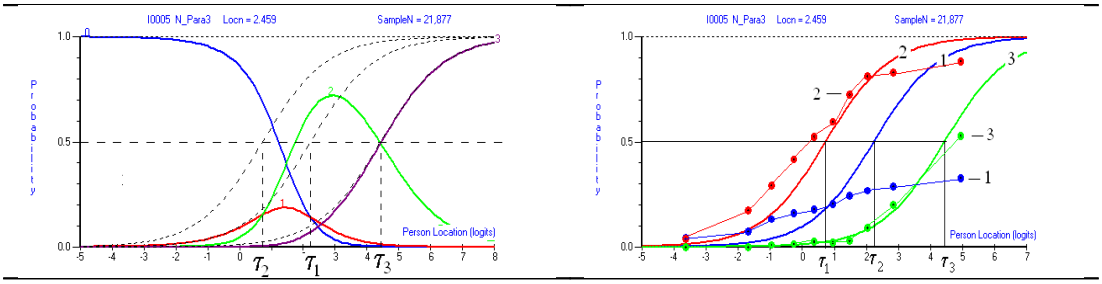
then $\alpha_{ix} = 0$. That is, the probabilities of two adjacent categories can be summed *only if the discrimination between the two categories is 0*, that is, if the responses between the two categories are random for all values of β. This seems eminently sensible. This property of the model has been discussed in detail in Jansen and Roskam (1986) and Andrich (1995).

That is, summing probabilities in adjacent categories is not consistent with the Rasch model, and correspondingly, pooling frequencies of adjacent categories and treating them as a single category, when the data follow the Rasch model, is not consistent with the model. Of course, when data do not fit the model or when thresholds are reversed, then exploratory analysis which might involve pooling categories in such a way may be instructive in understanding the misfit or the reversed thresholds.

Furthermore, it is evident from Eq. (40) that if the discriminations at the thresholds are *different in the data* so that the responses are effectively governed by products of the thresholds and discriminations at the thresholds, that is, $\kappa_{ix} = -\alpha_{i1}\delta_{i1} - \alpha_{i2}\delta_{i2} - \ldots \alpha_{ix}\delta_{ix}$, and if the data are analysed according to the PRM in which the discriminations are identical, that is $\kappa_{ix} = -\delta_{i1} - \delta_{i2} - \ldots \delta_{ix}$, then the threshold parameter estimates will be affected by the different discriminations in the data. In particular, this effect on the threshold locations will manifest itself when the discrimination between a pair of adjacent categories is 0, that is, the responses between two adjacent categories is random. In some cases where such disturbance of the model is present, reversed threshold estimates can appear.

Figure 8 shows such an example in which students' writing was assessed in one of four categories. It is evident from the first panel of the Figure that thresholds 1 and 2 are reversed. It is evident from the conditional proportions of successes between adjacent categories from 10 class intervals in the second panel, that the empirical discrimination of threshold 1 is much flatter than the discriminations at the other two

thresholds. This invites an exploratory collapsing of categories 0 and 1 and more importantly a study of the marking key and its interpretation by the judges.



Reversed CCC
Need to add the threshold dotted curves
In black

● Observed conditional proportions of successes in adjacent categories:
$proportion\ \{x_{ni} \mid (x-1)_{ni}\ or\ x_{ni}\}$

Figure 8 The Category Characteristic Curves (CCC) and the threshold characteristic curves for an item in the assessment of writing. Example from

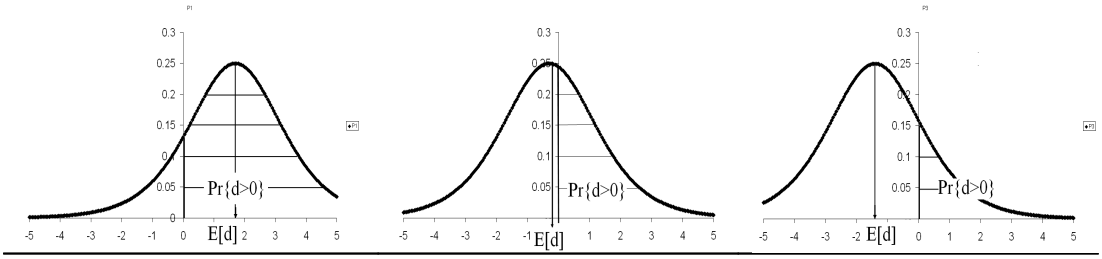
5.2 A response process at each thresholds

In order to help better contrast the PRM with the GRM, we now consider an interpretation based on a continuous dichotomised response process at each threshold. The interpretation follows that of Thurstone (1928), but with the logistic distribution function substituted for the Gaussian normal (Bock and Jones,1968) and with a person location fixed rather than being selected randomly from some population (Andrich, 1978b).

Let $d_x = \beta_n - \delta_{ix} + e$ be a continuous random variable characterising a continuous latent process in the engagement between person n and item i at threshold x. The variable e is an additive random process, an error, and for the Rasch model it is not subscripted by either the identities of the person or item – it is homogenous across persons and items. Then if $d_x > 0$, the response is taken as successful, otherwise it is unsuccessful. Substituting the logistic density function for the normal, the density $f(dn)$ of d_x is given by

$$f(d_x) = \frac{\exp(d_x - (\beta_n - \delta_{ix}))}{[1 + \exp(d_x - (\beta_n - \delta_{ix}))]^2}. \qquad (43)$$

Without specifically including a parameter for the variance of this distribution, given that it is taken to be homogenous across persons and items and therefore leaving it arbitrary, this distribution has a variance of $\pi^2/3$ (Gumbel, 1961). Figures 9a-9c show this process for three values of the thresholds δ_{ix}: -1.7, 0.3, 1.4 used in the second set of three items in the simulation studies. The shaded region shows the probability of $d_x > 0$. Clearly, the probability of a success decreases as the difficulty of the threshold increases.



9a: $E[d] = \beta_n - \delta_{ik}$, $\beta_n = 0, \delta_{ik} = -1.7$

9b: $E[d] = \beta_n - \delta_{ik}$, $\beta_n = 0, \delta_{ik} = 0.3$

9c: $E[d] = \beta_n - \delta_{ik}$, $\beta_n = 0, \delta_{ik} = 1.4$

Figure 9. Response processes at three thresholds for the PRM

Formally, $E[d_x] = \beta_n - \delta_{ix}$, which shows that the response density is centred at the location of the difference between the person and the threshold. It is this location that gives rise to different probabilities of success. Then, with the simplification immediately carried out, the probability

$$\Pr\{d_x > 0\} = \int_0^\infty \frac{\exp(d_x - (\beta_n - \delta_{ix}))}{[1 + \exp(d_x - (\beta_n - \delta_{ix}))]^2} dd_x = \frac{\exp(\beta_n - \delta_{ix})}{1 + \exp(\beta_n - \delta_{ix})}, \qquad (44)$$

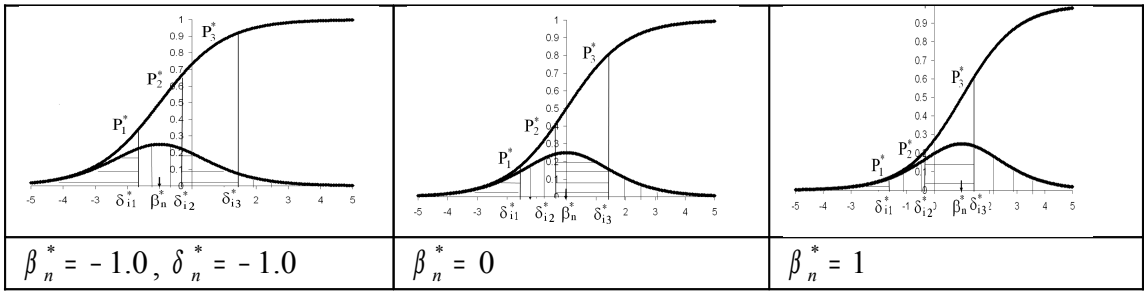
which is the dichotomous RM of Eq. (2) with the threshold δ_{ix} made explicit.

6 The Graded Response Model

The GRM can be motivated analogously to the processes in Figure 9, but there is a key difference (Bock, 1975). Rather than there being a response process at each threshold, there is only one response process for a person for an item with fixed thresholds.

6.1 The response processes of the GRM

Let $d_{ni} = \beta^*_n + e$ be a continuous random variable characterising a process of engagement between person n and item i. The location of person n is asterisked to distinguish it from the location in the PRM. $E[d_{ni}] = \beta^*_n$. Note that the thresholds are not identified in this process. Here we take e to be homogenous across persons and items and to be distributed again as the logistic density. This is an equivalent parameterisation to the PRM. More parameters have been used in the PRM. Figure 10 shows the process implied by the GRM. Let $X_{ni} = x$ be the random variable for the location of person n to item i in category $x \in (0,1,2...m_i)$. If $\delta^*_{ix} < d_{xi} < \delta^*_{ix+1}$, then $X_{ni} = x$, where $\delta^*_{ix}, \delta^*_{ix+1}$ are successive thresholds, also notated by an asterisk to distinguish them from the thresholds in the PRM.



$\beta^*_n = -1.0, \delta^*_n = -1.0$	$\beta^*_n = 0$	$\beta^*_n = 1$

Figure 10. Implied response processes for the GRM

It is this person location that gives rise to different probabilities of success. Then, with the simplification immediately carried out, the probability

$$\Pr\{\delta^*_{ix} < d_{xi} > \delta^*_{ix+1}\} = \int_{\delta^*_{ix}}^{\delta^*_{ix+1}} \frac{\exp(d_{ni} - \beta^*_n)}{[1+\exp(d_x - \beta^*_n)]^2} dd_{ni}$$

$$= \frac{\exp(\delta^*_{ix+1} - \beta^*_n)}{1+\exp(\delta^*_{ix+1} - \beta^*_n)} - \frac{\exp(\delta^*_{ix} - \beta^*_n)}{1+\exp(\delta^*_{ix} - \beta^*_n)} \quad . \qquad (45)$$

$$= \frac{\exp(\beta^*_n - \delta^*_{ix})}{1+\exp(\beta^*_n - \delta^*_{ix})} - \frac{\exp(\beta^*_n - \delta^*_{ix+1})}{1+\exp(\beta^*_n - \delta^*_{ix+1})}$$

Consider the components of Eq (45). They can be considered the probability of a success in a dichotomous response at thresholds δ^*_{ix} and δ^*_{ix+1} respectively.

Let $$P^*_{nix} = \frac{\exp(\beta^*_n - \delta^*_{ix})}{1 + \exp(\beta^*_n - \delta^*_{ix})}, \quad (46)$$

and let the probability of a response in category $X_{ni} = x$ be

$$\Pr\{X_{ni} = x\} = P_{nix}. \quad (47)$$

Then with a little rearrangement of the probabilities it can be shown that

$$\begin{aligned} P^*_{ni1} &= P_{ni1} + P_{ni2} + P_{ni3} \\ P^*_{ni2} &= P_{ni2} + P_{ni3} \\ P^*_{ni3} &= P_{ni3} \end{aligned} \quad . \quad (48)$$

It is evident that as in the development of the PRM, the dichotomous responses at the thresholds are implied, there being only one response in one of the categories. Second, they cannot be independent.

In addition, the person parameters cannot be eliminated in the estimation, and generally, for purposes of estimation, it is necessary to assume that the persons arise as a random sample from some population. We continue focussing on one person responding to one item for the purpose of drawing essential comparisons and contrasts between the PRM and the GRM.

From Eq. (48), it can be seen that the success of a dichotomous response at any threshold δ^*_{ix} is made of a sequence of implied dichotomisations by summing the probabilities in the categories on either side of the threshold. In data, this corresponds to adding the frequencies in adjacent categories. We can see immediately that this is in contrast to the PRM, in which such addition of probabilities and frequencies cannot be carried out unless the discrimination at a threshold is 0.

Finally, and most relevant to the theme of this Report, it is evident from Eq. (48) that if $P_{nix} > 0, \forall x$, then $P^*_{nix} > P^*_{nix+1}$ and therefore from Eq. (46) that $\delta^*_{ix+1} > \delta^*_{ix}$. Unlike in the PRM, and unlike the position of Fisher which is that the empirical ordering of categories is an hypothesis about the data, the ordering of the thresholds $\delta^*_{ix+1} > \delta^*_{ix}$ is a property of the GRM, irrespective of any properties of the data.

The difference in the construction of the PRM and the GRM which leads to this contrast in the interpretation of the respective thresholds is in the constraint that is imposed at the implied thresholds that takes account of the dependence among the implied dichotomous responses. In the PRM, the constraint is on the pattern of the responses themselves, in particular the Guttman structure is imposed; in the PRM it is on the pattern of probabilities. In the PRM, the relevant implied response probabilities (at the thresholds) are not constrained amongst each other, in the GRM they are.

It seems that the construction of the PRM involves a mixture of the locations of the thresholds and the locations of the persons. The higher responses of persons, which implies higher locations, seems to imply that necessarily the thresholds must also be in the same order.

6.2 Comparing the PRM and GRM thresholds in an example

The PRM and the GRM are models that cannot be transformed into each other. However, a kind of comparison can be made by deriving thresholds of the GRM from the thresholds of the PRM. This is done in order to further contrast the interpretations of the PRM and the GRM, specifically using the case of mathematics item S004 with reversed thresholds in Figure 7. It is also used to make a comment on the calculation of thresholds of the PRM to the thresholds of the GRM by Masters and Wright (1997).

Table 12 shows the construction of the thresholds of the GRM from the thresholds, $\hat{\delta}_{i1}$ $\hat{\delta}_{i2}$ for item S004 of the PRM. Because the parameters of the model cannot be

mapped onto one another other in a one to one correspondence, this transformation is a heuristic one. The calculation involves finding the location values β in the metric of the PRM for $P^*_{ni1} = 0.5$ and $P^*_{ni2} = 0.5$ where P^*_{ni1} and P^*_{ni2} are constructed from Eq. (48) given P^*_{nix} from the PRM. It must be stressed that these are the locations of the two GRM thresholds on the PRM metric, and that person locations obtained in general in the GRM would not be correlated perfectly with the locations for the PRM. This in itself makes it problematic to compute the GRM thresholds from those of the PRM and then to make interpretations. Nevertheless, from the perspective of the description of the data, with the same number of parameters, the descriptions will be virtually identical in terms of general tests of fit.

For illustrative purposes in item S004 of Figure 7, we now see from Table 12 that the GRM thresholds are close, but unlike the case of the PRM thresholds, they are in their natural order. We can now see that with such thresholds, there would be nothing in the values of the thresholds that would draw attention to what is a problem with the scoring key. Of course, in our simulated data with revised thresholds too, the thresholds of the PRM are in the correct order.

Table 12 Construction of the thresholds of the GRM from the PRM

β	δ_{i1}	δ_{i2}	P_{ni1}	P_{ni2}	$P^*_{ni1} = P_{ni1} + P_{ni2}$	$P^*_{ni2} = P_{ni2}$
$\delta^*_{i1} =$ -3.062	-1.360	-4.562	0.091	0.409	**0.500**	0.409
$\delta^*_{i2} =$ -2.860	-1.360	-4.562	0.091	0.500	0.591	**0.500**

I strongly advise against the construction of GRM thresholds from those of the PRM for purposes of interpretation just because it hides information revealed by the PRM thresholds. This position contrasts with that of Masters and Wright (1997) who consider that the graphs of the PRM with reversed thresholds, such as those in Figures 6 and 7, are a problem, and not that there is a problem with the item as reflected in the data and by the PRM thresholds. They then advocate calculating the GRM thresholds and interpreting them. However, because the graph is only a rendition of the model, Masters and Wright effectively advocate abandoning the PRM. They might as well

have started with the GRM. Essentially, they dismiss the empirical evidence that the categories are not working as intended by shifting to a model that cannot bear evidence on this question. Incidentally, this kind of shifting from one model to an incompatible one for interpretation is in itself a rather superficial use of statistical modelling.

The PRM provides a unique opportunity to study and understand the very basis of the data collection in formats with ordered response categories, and to check that the operational definition of increasing levels of the property across the categories is working empirically. It seems a shame to me to use the GRM in these circumstances. It seems an even greater shame to begin with the PRM, presumably because it is more tractable, and then to abandon it in favour of interpretations as though the GRM was used in the first place.

6.3 Neither the PRM nor the GRM characterise a stepwise process

Remarkably and unfortunately, both the PRM (Masters, 1982; Adams, Wilson and Wang, 1997) and the GRM (Samjima, 1996) have been promoted as a model that characterises data in which the respondent moves successively in *steps*, and then having failed a step does not proceed any further. Neither model can characterise such a process. As was shown in detail in Section 4, the implied response in the PRM for a score of x is a success at the first x thresholds and an unsuccessful response at all of the succeeding ones, not a no attempt at those following x. As noted, the probability of response in any category, including the first, is a function of all the probabilities of all the thresholds, including the last.

In the PRM, the probabilities of a response in any pair of categories is only a function of the threshold between them. However, every category has a probability of an outcome. Thus even if a person has a response in the first of four categories, the model provides the probability of a response in the last category – The model does not have an undefined probability of a response in the last category as would be required if the person, having failed at the first threshold (step), does not proceed to engage in the subsequent steps.

Interpreting either model as a model for a step process where (i) a response occurs at a step only if there is a success on the previous step, and the probability is only a function of the difficulty of that step, seems to arise from a lack of close study of the process behind the data generation for the PRM and the GRM, and the kind of model that might characterise a step process. The step process in which

$$\Pr\{X_{nix} = 1 \mid X_{nix-1} = 1\} = \Pr\{X_{nx} = 1\} = P_{nix},$$

$$\Pr\{X_{nix} = 1 \mid X_{nix-1} = 0\} \text{ is undefined}$$

because an unsuccessful response at step x-1 results in no attempt at step x, has the probability structure and sample space shown in Table 13. This is shown for a case of just 3 steps where $\Pr\{X_{nix} = 0 \mid X_{nix-1} = 1\} = \Pr\{X_{nx} = 1\} = 1 - P_{nix} = Q_{nix}$.

Table 13 Outcome space and probabilities in a stepwise process

Step			
1	2	3	Joint probability
Q_{ni1}	Undefined	Undefined	Q_{ni1}
P_{ni1}	Q_{ni2}	Undefined	$P_{ni1}\,Q_{ni2}$
P_{ni1}	P_{ni2}	Q_{ni3}	$P_{ni1}\,P_{ni2}\,Q_{ni3}$
P_{ni1}	P_{ni2}	P_{ni3}	$P_{ni1}\,P_{ni2}\,P_{ni3}$
	Sum of joint probabilities		1

Clearly the sum of the probabilities sums to 1, and so the outcome space is complete and responses independent. The response space is no more than that of a very rigorous tailored testing regime in which, given only one failed response, no further opportunity is given for a response. Thus it is the *opportunity* of the response that is affected by a failure at a previous response, not a change of probability. In the PRM and the GRM there is a probability of a response in every category.

Reconsider the examples of mathematics items in Figure 7. In both cases the students do not in fact take steps in their responses, and if they fail a step they do not proceed. They complete the task, and then the marker decides the mark that is to be awarded based on the performance. Neither the PRM nor the GRM is a model of transitions probabilities as in a Markov chain; they are static models with a probability of a

response in one of all categories given a *fixed person location and fixed threshold locations*. This involves a classification process.

7. Summary

The Report studies the PRM from first principles demonstrating the implications of its two distinctive properties: first, that the thresholds that partition the continuum into contiguous categories when the categories are working as intended, can be reversed in data and then not define a category; second, that the category frequencies and probabilities cannot be summed routinely to form a new category arbitrarily.

Regarding the second property, it is shown that two categories can be combined in the above way only if the discrimination between the adjacent categories is zero, that is, if the response between the two categories is random. In this way, the model is shown to be different from the alternate model for ordered categories, the GRM, which is constructed by combing categories in just this way.

Most of the space in the Report was devoted to the first of the properties, the possible reversal of threshold estimates. First, it was argued that in the case where a design could be constructed in which the responses at the thresholds were independent, it would be required that the threshold difficulties show the intended order of the successive categories. To stress its importance, this requirement was termed an Axiom. Second, it was then shown that the reversal of threshold estimates could be interpreted in a response outcome space in which responses at the thresholds were independent. Therefore, third, it was concluded that the reversed threshold estimates indicated a problem with the empirical ordering of the categories.

This argument was set in the context of a theme of the Report, given a direction from Fisher, that the ordering of the categories should be a property of the data, and not a property of the model by which they are analysed. Thus the empirical ordering of the thresholds becomes an empirical question, not simply an assertion. The property of possible reversals of the threshold estimates in the PRM were again contrasted to the corresponding property of the GRM in which the ordering of its thresholds, defined

differently from the way they are defined in the PRM, is a property of the model irrespective of the properties of the data.

In the process of the development of this argument, three possible functions of statistical models were summarised: (i) accounting for and describing data evidenced by tests of fit; (ii) characterising a process; and (iii) characterising a requirement. In the case of the PRM, it was stressed that it was not derived from an attempt to model data, but from an attempt to characterise the requirement of invariant comparisons among persons and among items. Thus the case for the model does not rest on any test of fit and whether the data fit the model or not is an empirical question. Second, it was shown that the model characterised a process of ordered classifying, and therefore that it was entirely compatible with the intended response process in item formats with ordered response categories. However, the usual tests of fit do not seem to bear on this the question of whether the classification process is working as intended. This can be inferred directly from the empirical ordering of the threshold estimates.

One of the points that the Report raises concerns that lack of any previously specified criterion other than that implied in Fisher's approach for the ordering of categories.

It is suggested here that the answer can be found in the dominant paradigm in statistical analysis, that the case for a model is that it accounts for the data at hand. If the model does not account for the data at hand, another model is sought (Andrich, 2004). It is so dominant that it is even used in the case when the model has been derived to characterise a process or when it is derived to meet requirements, such as the PRM. That is, it is seen that the case of accounting for data is both necessary and sufficient for it to be utilised. There seems no other explanation as to why the diagrams such as those in Figures 6, 7 and 8, which invite questioning the operation of the categories, have not been questioned outside the framework of the PRM, and often ignored or circumvented even within that framework. That is, it seems that the reason they have not been questioned is that tests of fit can be found in which the data seem to be accounted for by the model. A case is made as forcefully as possible in this Report that in the case where are model is derived from criteria other than for the

purpose of modelling a particular data set, but for example to meet a requirement and to characterise a process such as the PRM, statistical tests of fit are not sufficient.

One final comment is made regarding the person estimates. It might be suggested that in view of the model being able to estimate reversed thresholds which can be interpreted in an experimentally independent design, that therefore the model should simply be used to estimate person locations without concern for the ordering of the thresholds. There are two responses to this superficial suggestion. First, the reversal of the thresholds tells that there is something wrong with the intended ordering of the categories, and that signals that there is a problem with the data. The reversed thresholds are a symptom of a problem in the data which the model exposes. The person estimates contain whatever problem is in the data that leads to such estimates.

The second is that the reversed threshold estimates give an indication where the attempt at understanding what it means to have more of the property has failed. It gives an opportunity to better understand the variable and to construct more rigorous items. To use the model simply to account for the reversed thresholds in the data would again be taking the position that the fit of the data to the model is both a necessary and sufficient case for the application of the model.

References

Adams, R.J., Wilson, M., and Wang, W. (1997) The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement,* 21, 1-23.

Andersen, E.B. (1973). Conditional inference for multiple choice questionnaires.*British Journal of Mathematical and Statistical Psychology, 26,* 31-44.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-574.

Andrich, D. (1978b). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2* (2), 449-460.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. *Biometrics, 35* (2), 403-415.

Andrich, D. (l985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological Methodology,* San Francisco, Jossey-Bass. (Chapter 2, pp. 33-80.).

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.

Andrich, D. (1995). Models for measurement, precision and the non-dichotomization of graded responses. *Psychometrika, 60,* (1) 7-26.

Andrich, D. (2004) Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care,* 42, 7 – 16.

Andrich, D. and Luo, G. (2003) Conditional estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement 4,* 205 - 221.

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.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical theories of mental test scores* (pp. 397 –545). Reading, Mass.: Addison-Wesley.

Bock, R.D. (1975). Multivariate statistical methods in behavioral research. New York: McGraw-Hill.

Bock, R.D. & Jones, L.V. (1968). *The measurement and prediction of judgement and choice.* San Francisco: Holden Day.

Choppin, B. (1968). An item bank using sample-free calibration. *Nature, 219,* 870-872

Feller, W. (1968) An introduction to probability theory and its applications. Vol. 1. 3rd Edition. New York: John Wiley.

Fisher, R.A. (1958). *Statistical methods for research workers* (13th ed.). New York: Hafner.

Fischer, G. H. and Molenaar, I.W. (1995) (Eds.) Rasch models: foundations, recent developments, and applications. New York: Springer

Gumbel, E, J. (1961) Bivariate logistic distributions. Journal of the American Statistical Association, 56, 335 -349.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star and J.A. Clausen (Eds.), *Measurement and Prediction,* pp.60-90. New York: Wiley.

Harris, J. (1991) Consequences for social measurement of collapsing categories within items with three or more ordered categories. Unpublished Master of Education Disseration, Murdoch University, Murdoch, Western Australia.

Jansen P.G.W. & Roskam, E.E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51*(1), 69-91.

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.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society,* Series B, *42*(2), 109-142.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B.D. Wright, (1980). Chicago: The University of Chicago Press.

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* (pp.321-334). Berkeley CA: University of California Press.

Rasch, G. (1966). An individualistic approach to item analysis. In P.F. Lazarsfeld and N.W. Henry, (Eds.). *Readings in Mathematical Social Science (*pp.89-108). Chicago: Science Research Associates.

Rasch, G. (1979) Personal communication.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 34*(2, No.17).

Samejima, F. (1996) Graded response model. In W.J. van der Linden and R.K. Hambleton (Eds.) *Handbook of Item Response Theory*. (pp. 85 – 100). New York. Springer.

Samejima, F. (1997) Departure from normal assumptions: a promise for future psychometrics with substantive mathematical modelling. *Psychometrika*, 62, 471 - 493.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433-51.

Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology 33,* 529-54.

Thurstone, L.L. & Chave, E.J. (1929). *The Measurement of Attitude.* Chicago: University of Chicago Press, pp.415-430.

Van Wyke, J. F. (2003) Constructing and interpreting achievement scales using polytomously scored items: a comparison between the Rasch and Thurstone models. Unpublished EdD Dissertation, Murdoch University.

Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.

Wright, B.D. (1997) A history of social science measurement. Educational Measurement: Issues and Practice, 16, 4, 33-45.

Wright, B.D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Zwinderman, A. H. (1995) Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement, 19* , 4, 369 - 375.