Educational Measurement
ISSUES AND PRACTICE
NCME national council on measurement in education

*Educational Measurement: Issues and Practice*
*Summer 2015, Vol. 34, No. 2, pp. 8–14*

# The Problem With the Step Metaphor for Polytomous Models for Ordinal Assessments

David Andrich, *University of Western Australia*

**Keywords:** graded response model, item response theory, polytomous items, polytomous Rasch model

Penfield's (2014) "Instructional Module on Polytomous Item Response Theory Models" begins with a review of dichotomous response models. He refers to these as *The Building Blocks of Polytomous IRT Models: The Step Function*. The mathematics of these models and their interrelationships with the polytomous models is correct. Unfortunately, the *step* characterization for dichotomous responses, which he uses to explain the two most commonly used classes of polytomous models for ordered categories, is incompatible with the mathematical structure of these models. These two classes of models are referred to in Penfield's paper as *adjacent category models* and *cumulative models.* At best, taken in the dynamic sense of *taking* a step, the step metaphor leads to a superficial understanding of the models as mere descriptions of the data; at worst it leads to a misunderstanding of the models and how they can be used to assess if the empirical ordering of the categories is consistent with the intended ordering. The purpose of this note is to explain why the step metaphor is incompatible with both models and to summarize the distinct processes for each. It is also shows, with concrete examples, how one of these models can be applied to better understand assessments in ordered categories.

## Adjacent Category Models

First consider the *adjacent category* models. They are called such because, as shown below, a simple rendition of the model involves the ratio of the probabilities of two adjacent categories. Consider, for example, the polytomous Rasch model (PRM), the simplest special case of the class of adjacent category models. It has two common parameterizations, the *partial credit* and *rating scale*, which at the level of the response process of a single person to a single item are identical. The only difference is in the parameterization: if the structure across items is the same, as in many rating response formats (such as *strongly disagree*, *disagree*, *agree*, *strongly agree*), it may be that category parameters can be taken as identical across the items to give the rating parameterization; otherwise the items may have different maximum scores and different parameter estimates to give the partial credit parameterization. Whether or not the rating formulation holds is an empirical question.

It is helpful to have the graphical presentation of the response structure for the categories. Figure 1, which is similar to the second example in Figure 2 of Penfield (2014), shows the probability of a response in each category as a function of the location of the entity being assessed. Penfield refers to these probability functions as item response functions (IRFs). Figure 1 also shows the points of intersection ($b_{i1}$, $b_{i2}$, $b_{i3}$), not shown in Penfield, of pairs of adjacent IRFs. They are the points on the trait where the probability of response in adjacent categories is identical. In the case of dichotomous items, there is only one such point, and in performance assessment it is referred to generally as an item's *difficulty*. In general terms, the point which has a 50% probability of a response in either category is referred to as a *threshold.* This is the original terminology in Andrich (1978). Unfortunately, it is these points of equal probability which are referred to as *steps*, a terminology that has been propagated despite not being compatible with the either the PRM or with its generalization, the so-called general partial credit model (Muraki, 1992).

### *The Polytomous Rasch Model*

In part because of its special features, including having the least parameters, the PRM will be used to study the parameters in Figure 1. It is stressed that the points noted here are relevant for all adjacent category models. A generalization is considered later in the section.

Using the notation in Penfield (2014), the partial credit parameterization of the PRM takes the form

$$P_{i0}(\theta) = \frac{1}{1+\sum_{r=1}^{m}[\exp(\sum_{k=1}^{r}(\theta - b_{ik})]}, \qquad (1)$$

$$P_{ij}(\theta) = \frac{\exp(\sum_{k=1}^{j}(\theta - b_{ik})}{1+\sum_{r=1}^{m}[\exp(\sum_{k=1}^{r}(\theta - b_{ik})]}, \quad j > 0, \qquad (2)$$

where $m+1$ is the number of categories for item $i$. (Note that because the focus here is on just one item, and following Penfield, the maximum score $m$, which can vary from item to item, is not subscripted by $i$. Although Equations 1 and 2 comprise a common way of expressing the model, the PRM can be simplified in such a way that it helps understand why the step metaphor is not compatible with it. If the numerator in Equation 2 is expanded and incorporated in the denominator, it takes the form

$$P_{ij}(\theta) = \frac{\exp(j\theta - \sum_{k=1}^{j} b_{ik})}{1+\sum_{r=1}^{m}[\exp(r\theta - \sum_{k=1}^{r} b_{ik})]}, \quad j > 0, \qquad (3)$$

*David Andrich, Chapple Professor, Graduate School of Education, The University of Western Australia, Crawley, Western Australia 6009, Australia; david.andrich@uwa.edu.au.*

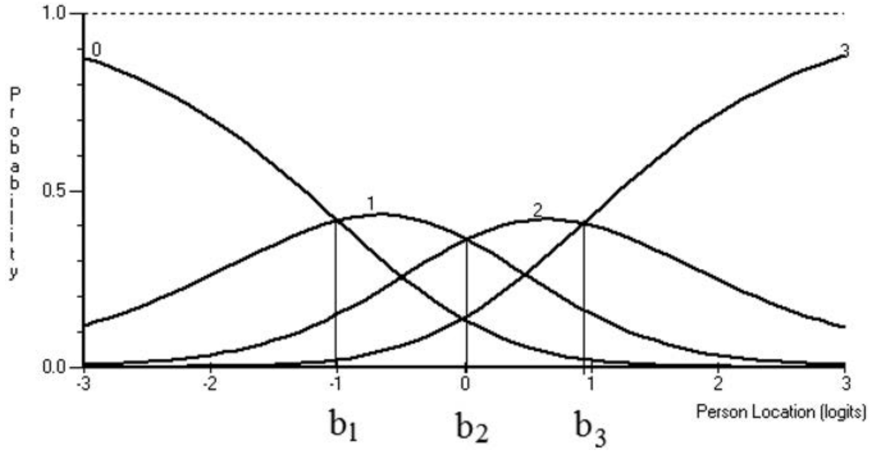


FIGURE 1. Polytomous item with four ordered categories.

where $j$ is simply the *count* of the number of thresholds $b_{ik}$ exceeded in the response, $m - j$ is the number not exceeded, and $\theta$ is the location of any person on the trait. The orientation of the trait in Figure 1 is increasing from left to right. Thus, a score of 2, in which $m = 3$, implies a response which exceeds both thresholds 1 and 2 but fails threshold 3. In addition, by defining $\sum_{k=0}^{0} b_{ik} \equiv 0$, Equation 3 also characterizes Equation 1, the response in the first category, in which no thresholds are exceeded. This gives the general, single equation

$$P_{ij}(\theta)=\frac{\exp(j\theta-\sum_{k=0}^{j}b_{ik})}{\sum_{r=0}^{m}\left[\exp(r\theta-\sum_{k=0}^{r}b_{ik})\right]},\ j=0,\ 1,\ 2,\ ...,m. \qquad (4)$$

On simplification, the ratio of the probabilities of $P_{ij}(\theta)$and $P_{i(j-1)}(\theta)$ takes the form

$$P_{ij}(\theta)/P_{i(j-1)}(\theta)\ =\ \exp(\theta - b_{ij}),\ \ \forall j,\ j=1,2...,m. \qquad (5)$$

Further,

$$P_{ij}(\theta)/(P_{i(j-1)}(\theta)+P_{ij}(\theta))\ =\ \exp(\theta - b_{ij}) \\ /(1+\exp(\theta - b_{ij}),\ \ \forall j,\ j=1,2...,m, \qquad (6)$$

which is the dichotomous Rasch model at threshold $b_{ij}$. Equation 6 can be used to derive Equation 4, and hence the terminology of *adjacent category* for this class of models. Equation 6, the conditional probability of a response in the intended higher of two adjacent categories, is a *latent* dichotomous response between the categories. It is stressed that even though it is the *building block* of adjacent category models, this response is never observed.

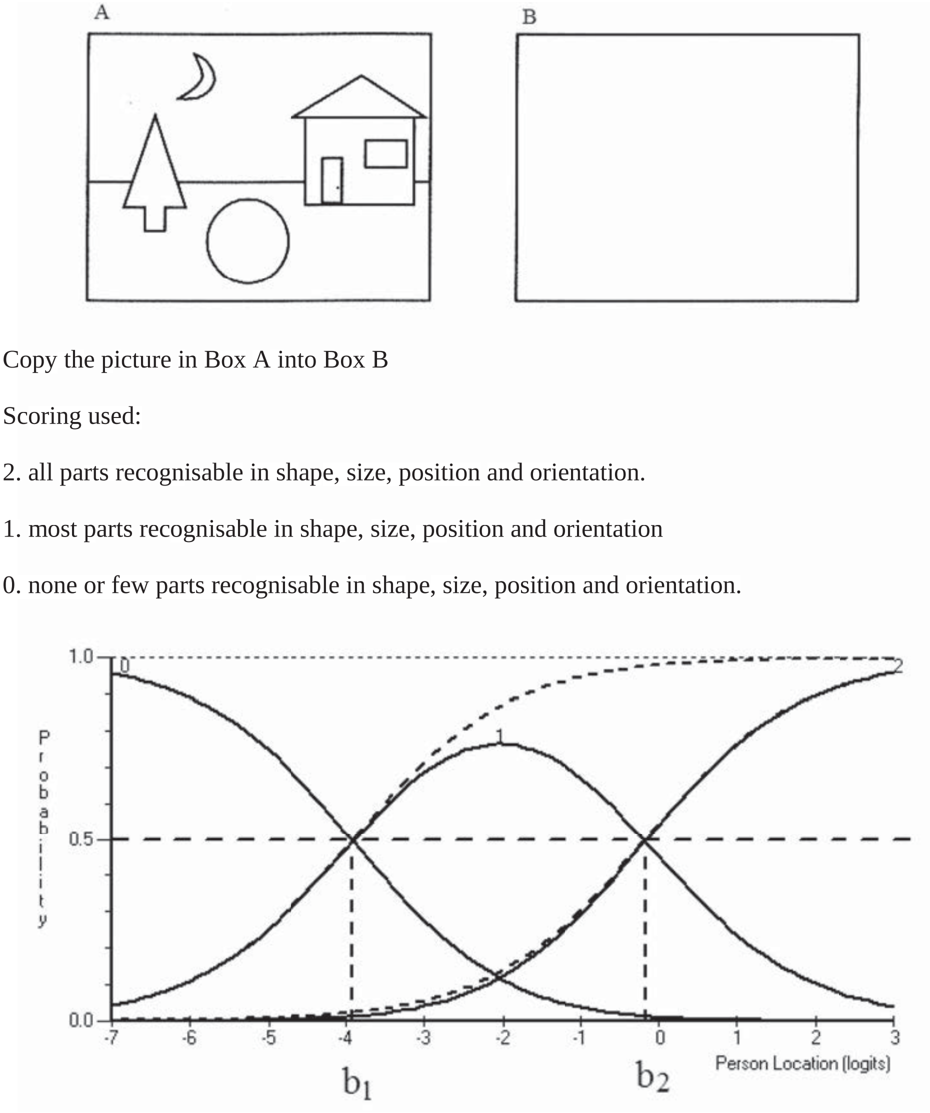
The characterization by Penfield for the step metaphor is that of *taking* a step across the parameter $b_{ij}$. Thus, for response functions as shown in Figure 1, and acknowledging Masters (1982) and Muraki (1992), Penfield writes:

> One can conceptualize the score an examinee receives as being determined by the success that she has had in transitioning, or stepping, to successfully higher score categories. (Penfield, 2014, p. 39)

This conceptualization implies a dynamic, sequential process of some kind. However, Equation 4 *does not characterize the process* by which the person being assessed reaches a location on the continuum. It does not characterize it, either in the case of self-assessment as in attitude assessment, or in the case of performance assessment. Instead, because it is a static model, and for a single location $\theta$, the PRM characterizes only the *probability* of a response in each category as a function of the values of the thresholds. That it cannot be a sequential step process can be clarified in a number of ways.

First, although Penfield uses the dichotomous response at the thresholds to extend to the polytomous response, the language of *transitioning* or *stepping* across the difficulty parameter in the dichotomous case is never used. For example, in a test of achievement where the response categories may be *incorrect* and *correct*, the person does not *transit* from incorrect to correct—the person simply responds, and the response is *classified* as incorrect or correct. In the vast literature on dichotomous response items and response models for them, the difficulty parameter, which locates the point of equal probability of the two responses, has never been characterized as a transition from incorrect to correct.

The correct generalization from the dichotomous to the polytomous responses is that, rather than a performance being classified in one of two ordered categories ($m = 1$), it is *classified* in one of $m + 1$ ordered categories ($m > 0$). For example, if a performance, say in acting, is to be classified according to some protocol into four ordered categories, the actor is not considered to go through a poor performance, and then an improved performance, to finally a performance in the highest category—indeed an actor might perform superbly at the beginning of the play, and then tail off at the end. Then, if the entire performance is to be assessed by a single rating, the judge has to make an on balance judgment according to the assessment protocol. The assessment is that of the placing the person's performance in one of the categories on the trait, not how the person transitioned in getting to that category. In their studies the acting students might improve on the trait, just as students in schools improve their standing on the curriculum that is taught to them, but the model does not characterize this improvement over time—it is simply an assessment of where the person is at time of assessment, and not how the person reached there.



Copy the picture in Box A into Box B

Scoring used:

2. all parts recognisable in shape, size, position and orientation.

1. most parts recognisable in shape, size, position and orientation

0. none or few parts recognisable in shape, size, position and orientation.

FIGURE 2. An item where the second threshold adds substantively to the first threshold.

Second, for this class of models, the probability of a response in any category is a function of *all thresholds*. This can be seen by observation that the denominator of Equation 4 is a function of all thresholds—it is a sum of all the numerators. Because of this feature, Thissen and Steinberg (1986) also referred to this class of models as *divide by total* models. Because the probability of a response in any category is a function of *all thresholds*, it means that if the third threshold is changed in Figure 1, then the probability of a response in the first category is also changed. This feature is incompatible with the idea that the model characterizes transitioning from one category to the next—if it did, the probability of transitioning across the first threshold from the first to the second category would not depend on the location of the last threshold.

Third, all threshold parameters in Equation 4 are referenced to the same origin and therefore cannot be conceptualized as successive steps. The thresholds are, for example, like the markings of a weight scale—one does not describe the point of 80 kg as a step which is smaller than the step of 81 kg. Moreover, in classifying a person according to weight, one does not characterize how the person reached that weight—a young person might have grown from a smaller weight to the present one; an older person might have shrunk from an earlier weight to the current one. By analogy to these types of measurement, the PRM model simply characterizes the probability of being classified in one of the categories, defined by the thresholds, as a function of the location $\theta$ of a person on the trait—it does no more and no less.

In arguing that it was incompatible with the PRM, Molenaar (1983) called the step metaphor *seductive*; Ostini and Nering (2006) note explicitly that this metaphor is not compatible with the model. Penfield also refers to Tutz (1990) as providing a motivation for the step metaphor. However, the model that Tutz develops is a genuine sequential processing model and is not one of the adjacent category models such as the PRM. Placing Tutz's example with the other models in the use of *step* confuses the processes behind the different models.

Given that the item parameters (thresholds) are not steps, how can they be interpreted? They are interpreted exactly as they are in the dichotomous case—as difficulty parameters. This is elaborated briefly below. There are, as Penfield indicates, dichotomous responses at the thresholds, but, as already noted above, they are latent, never observed. In addition, because there is one observed response in one category only, these latent responses are not independent—they are constrained. The constraint is an implied Guttman structure on the latent responses at the thresholds. Thus, if the assessment in the example of Figure 1 is in category 2 (a score of 1), the implication is having exceeded threshold 1 but neither threshold 2 nor threshold 3; likewise, if an assessment is in category 3 (a score of 2), the implication is having exceeded both thresholds 1 and 2, but not threshold 3. And elegantly, if the response is in the first category (a score of 0), the number of thresholds exceeded is 0; if the response is in category 4 (score of 3), all three thresholds have been exceeded. This characterization reflects the severe constraint that order places on the latent, dichotomous, responses at the thresholds.

Because the successive categories are intended to be ordered in the sense that they successively reflect more of the trait, the implication is that the thresholds which define the boundaries of the categories are not only expected, but *required* to increase in difficulty. However, and importantly in the PRM, the empirical evidence can be to the contrary—that is, the threshold *estimates* can be reversed. This reversal is a property of the assessment data, and is relevant to understanding whether the rating or partial credit scoring—whichever is relevant—is working as intended. Figures 2 and 3 below show two items with partial credit in the assessment of numeracy, the first with ordered and the second with reversed threshold estimates. The figures also show the latent, dichotomous response probability curves at the thresholds. These are real items that were used in the assessment of elementary school children in Western Australia and the responses were analyzed using the PRM (van Wyke, 2003). The parameters of the two items are on the same scale.

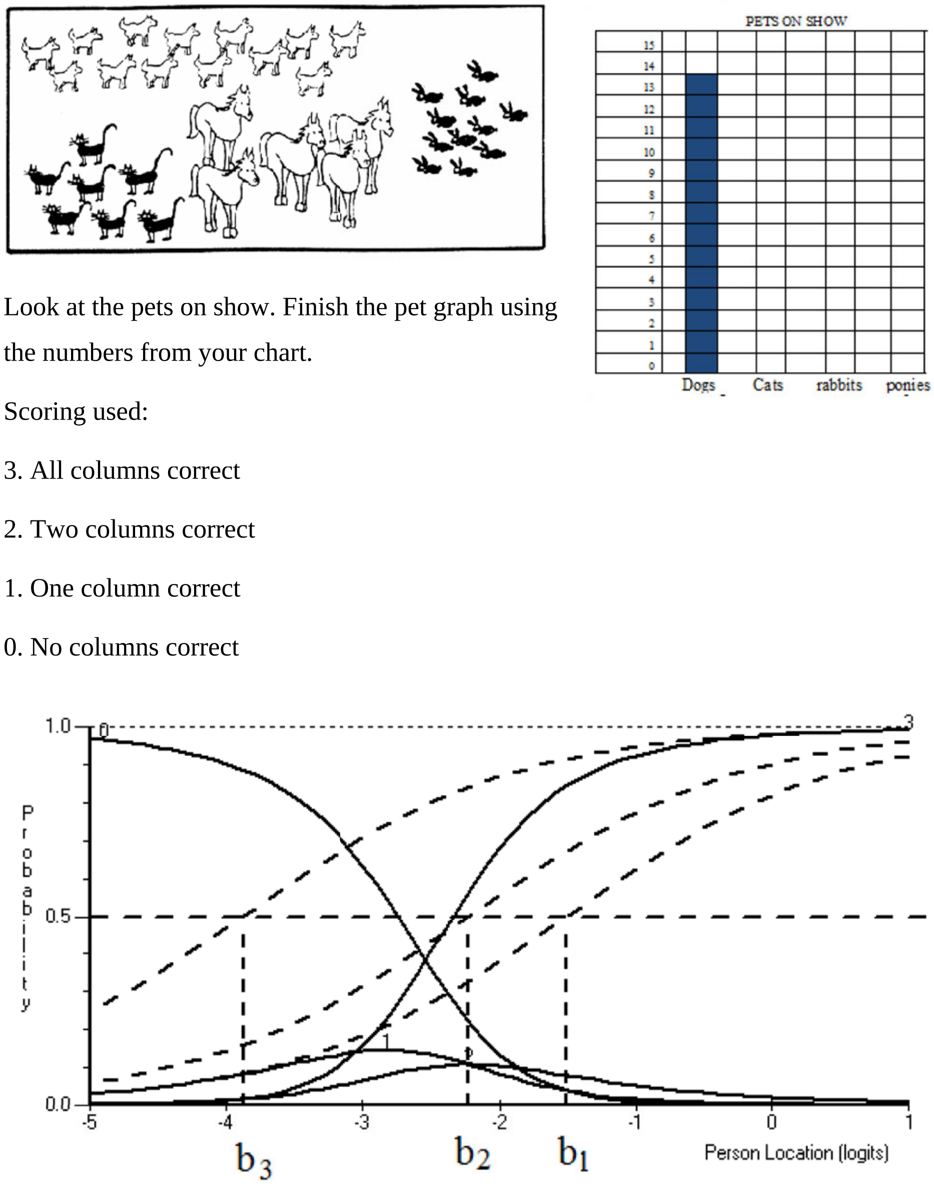
Consider now the items and their scoring structure more closely. In Figure 2, a drawing with *all parts recognizable* ... (score of 2) reflects greater understanding on the variable than *most parts recognizable* ..., (score of 1) and reflecting this property, the second threshold estimate is more difficult than the first one. On the other hand, in Figure 3, the achievement of the score of 3 does not represent a greater value on the trait than scores of 2 or 1. It is evident that if a student draws one column correctly, then the student will almost certainly draw all correctly—the successive scores reflecting more columns drawn correctly do not reflect additional standing on the variable—the additional scores are for an equivalent performance for all responses and because of the constraint on responses, the successive thresholds appear easier than the first one. At best, having students complete more than one column is a waste of their time in the testing situation. In both cases the student responding does not go through any steps on an implied continuum in achieving the final performance. In the step conceptualization, it would be said that taking step 3 is easier than taking step 1; however, it is clear that these so called steps are essentially of equal difficulty and further reinforces that the step conceptualization is not compatible with the model.

Two points are stressed regarding Figures 1–3. First, they represent the probability distribution of the response of a *single person* at any location on the trait, and *not of the distribution of persons in the sample* among the categories. Thus, Figure 2 shows that a student located between approximately −4 and 0 logits, is most likely to obtain a score of 2, a genuine partially correct response. On the other hand, Figure 3 shows that, for a student located between approximately −3.9 and −1.5 logits, in which scores of 1 and 2 have *their own* highest probabilities, *simultaneously both* 0 and 3 have a higher probability. Effectively, the response is dichotomous. In general, reversed threshold estimates provide evidence of some kind of problem with the empirical ordering of the categories which in turn provides an opportunity to improve this important feature of an item. It is stressed that the source of the problem is not provided by the estimates themselves, nor simply by the frequencies in the data—for example, that some frequencies are very small. In some cases, small frequencies, which give large standard errors of the threshold estimates, may be an explanation, but it is not necessarily an explanation. In general, it needs to be considered whether there is a problem with the definition, operation or some other aspect of the categories, and this can only be done by considering the empirical set-up.

Second, a typical test of fit which compares expected and observed frequencies of correct responses is irrelevant in deciding that reversed thresholds reveal a problem with the empirical ordering of the categories. It is irrelevant because, for such a purpose, the argument is circular: the frequencies in the data are used to estimate the parameters which in this case lead to reversed threshold estimates, and then these same reversed threshold estimates are used to obtain the expected frequencies. It may be that reversed threshold estimates are part of a general problem with an item which leads to misfit; however, fit does not imply that reversed thresholds are not a problem in the empirical ordering of the categories.

Penfield's Figure 5B shows a figure similar to Figure 3, yet the reversed thresholds (in his terminology, steps) go unremarked.

The same interpretation, that the threshold parameters are not steps, can be made with the generalized partial credit parameterization (Muraki, 1992). On the other hand, there is a major difference between the PRM (partial credit or rating scale parameterization) and the generalized parameterization. Unlike the PRM, the generalized model has different discrimination parameters at the thresholds for different items. Penfield refers to models with different discriminations as *more flexible*, implying that such models are better than those less flexible. However, that view is not universal. That flexibility is better comes from the statistical paradigm of modeling data, where the task is to find a model that fits the data, not from a measurement paradigm which requires the data to meet specified conditions (Andrich, 2004, 2013b). In interpreting the parameters of the PRM, Andrich (1978) began with different discriminations for the latent responses at the thresholds, and showed that if they were made equal then the model became a Rasch model in the sense that it had sufficient statistics for its parameters. Muraki's reintroduction of different threshold discriminations destroys sufficiency. The sufficient statistic for each person is simply the total score across items, which can be seen from the coefficient $j$ of $\theta$ in Equation 3. An advantage of the model with sufficient statistics is that the item parameters can be estimated independently of the person parameters, and vice versa, and there is a compelling *measurement* perspective that responses should not only be substantively valid, but also fit the relevant member of the Rasch class of models (Rasch, 1961; Wright, 1997). Then the item parameter estimates, within



Look at the pets on show. Finish the pet graph using the numbers from your chart.

Scoring used:

3. All columns correct

2. Two columns correct

1. One column correct

0. No columns correct

FIGURE 3. An item where the second and third thresholds do not add substantively to the first threshold.

the relevant frame of reference, can be taken to characterize the items independently of the person distribution. On the other hand, in the models with different discriminations at the thresholds, the item parameter estimates are entangled with the estimated distribution of person parameters.

## Cumulative Models

Although Penfield's algebraic relationships for cumulative models are also correct, not only is the step concept incompatible with these models, but the cumulative and the adjacent categories models characterize incompatible processes. A feature of adjacent category models is that it is not possible to combine a pair of adjacent categories without changing the model and the relationships among all categories. Thus if two or more categories are combined by removing a threshold, it changes the probability of responses in other categories. Moreover, the sum of the probabilities of adjacent categories calculated *post hoc* is not the same as if the categories were combined *a priori* and responses made in one of the smaller number of categories. Anderson (1984) considers that this property of adjacent category models is compatible with the way assessments work in the social sciences.

On the other hand, in cumulative models combining adjacent categories, either *a priori* or *post hoc,* does not change the probabilities of outcomes in all categories. To achieve this feature, there is only a single latent response process for the distribution among all categories, and this distribution is divided into categories *post hoc*. Such a distribution is shown in Figure 4. The process described by Anderson that is relevant to this model is *group continuous*, for example, income in dollars: 0–2,000, 2,001–3,000, and so on (Anderson, 1984, p. 2). Anderson considers that this property of cumulative models is *not* compatible with the way assessments work in social sciences. Figure 4 shows a single, continuous latent distribution around the difference between the person location $\theta$ and the item location $\delta$ $(\theta - \delta)$. The figure also shows three

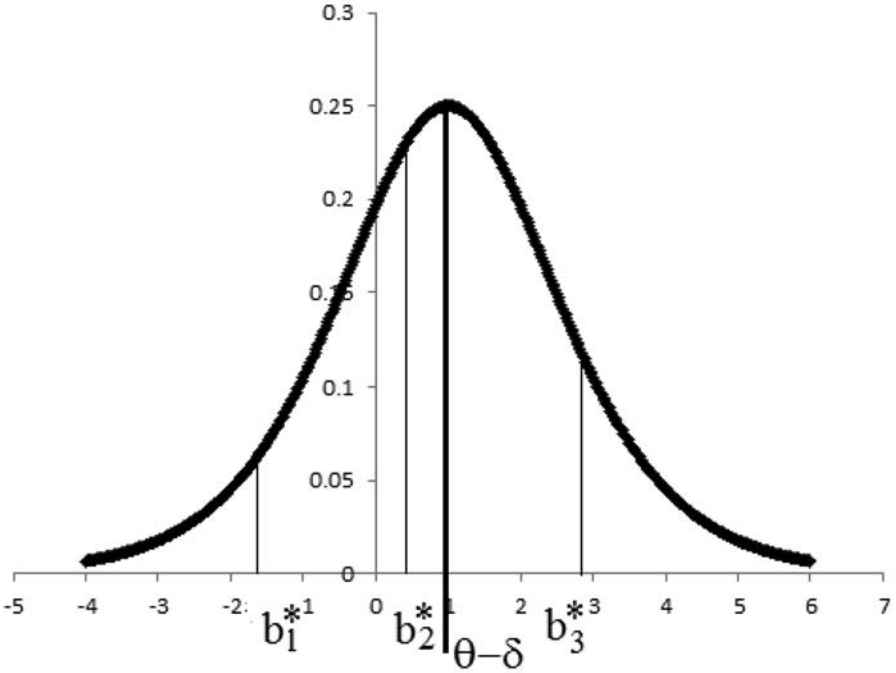


FIGURE 4. The response process corresponding to cumulative models.

thresholds, superscripted with an asterisk (*) to distinguish them from the thresholds in the adjacent category models. Then if $P_{ij}^*(\theta)$ is the area under the curve from $b_{ij}^*$ to $\infty$, a cumulative probability, and the total area under the curve is given by$P_{i0}^*(\theta) = 1$, then the probability of a response in any category $j = 0,\ 1,\ 2...,m$ is given by the successive differences

$$P_{ij}(\theta) \;=\; P_{ij}^*(\theta) - P_{i(j+1)}^*(\theta);\; P_{im}(\theta) \;=\; P_{im}^*(\theta). \quad (7)$$

This class of models, based on Thurstone and Chave (1929), was developed further by Samejima in the psychometric context (1969, 1997). Because of the structure of Equation 7, Thissen and Steinberg (1986) referred to this class of models as *difference* models. Clearly, in this model, its thresholds are always ordered even if there is evidence from the PRM that the empirical ordering of the categories is malfunctioning.

There is no concept of step in this model. The use of a cumulative model for items such as those in Figures 2 and 3, and in all assessments, must be merely descriptive without concern for any processes behind the responses. To be explicit, the model specifies the probability that a person will be classified above any threshold, that is, assessed in a category or in any category above it, and not in a specific category. This does not seem consistent with performance assessment—judges locate a performance in one of the categories, not in and beyond any particular category. For example, in the item in Figure 2, the model specifies the probability that a person drew figures in which all parts were recognizable (score 2) and then the sum of the probabilities where all parts were recognizable and where most parts are recognizable (score 1), and so on. In Figure 3, it models the probability that a person had all three columns correct (score 3), the sum of the probabilities that the person had three and two columns correct (score 2), and so on. The characterization of these cumulative probabilities does not seem compatible with how a judge makes a decision.

In the first two stages of the development of the PRM for items with ordered categories, which were based on the requirement of sufficient statistics, the item parameters had no substantive meaning (Rasch, 1961; Andersen, 1977). In a third stage, and beginning with the rating parameterization, meaning was given to them in terms of the familiar thresholds as difficulties on the continuum and discriminations at the thresholds (Andrich, 1978). In a fourth stage, the parameterization was generalized so that different items could have different thresholds (Masters, 1982). At the time it seemed difficult to appreciate the PRM's two distinctive features broached above: first that adjacent categories could not be combined arbitrarily; second that it was not only possible to have reversed threshold estimates, but that they implied that the empirical ordering of the categories violated the intended ordering. Both these features contrasted explicitly with those of the established cumulative model, seemed counterintuitive, and generated debate (Adams, Wu, & Wilson, 2012; Andrich, 1995a, 2013a; Jansen & Roskam, 1986). It seems that, in the context of these contrasts between the established model and the radically different properties of the new model, the irrelevant but *seductive* step metaphor was imagined.

Penfield's use of the step metaphor has provided the opportunities, first to explain in detail why this metaphor is incompatible with the adjacent and cumulative types of models; second to contrast the processes implied by models which are articulated further in Andrich (1995b), third to indicate that the process behind the adjacent category models is the one consistent with assessments in ordered categories, and finally to demonstrate that when the adjacent category model, the PRM, is used, the empirical ordering of the categories can be studied, understood, and where necessary improved to be consistent with the all-important intended ordering.

## References

Adams, R. J., Wu, M. L. & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*, 547–573.

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

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, *46*, 1–30.

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

Andrich, D. (1995a). Models for measurement, precision and the nondichotomization of graded responses. *Psychometrika*, *60*, 7–26.

Andrich, D. (1995b). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measuremen*t, *19*, 101–119.

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

Andrich, D. (2013a). An expanded derivation of the threshold structure of the polytomous Rasch rating model which dispels any "threshold disorder controversy". *Educational and Psychological Measurement*, *73*, 78–124.

Andrich, D. (2013b). The legacies of R. A. Fisher and K. Pearson in the application of the polytomous Rasch model for assessing the empirical ordering of categories. *Educational and Psychological Measurement*, *3*, 553–580.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Molenaar, I. W. (1983). *Item steps* (Heymans Bulletin HB-83-63—EX). Groningen, The Netherlands: University of Groningen.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36–48.

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* (pp. 321–334). Berkeley: University of California Press. Reprinted in D. J. Bartholomew (Ed.), *Measurement* (Volume I, 319–334). London: Sage Publications.

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

Samejima, F. (1997). Graded response model. In W. Van derLinden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York: Springer.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Tutz, G. (1990). Sequential item response models and ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55.

Van Wyke, J. F. (2003). *Constructing and interpreting achievement scales using polytomously scored items: A comparison between the Rasch and Thurstone models* (Unpublished doctoral dissertation). Murdoch University, Murdoch, Western Australia.

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