Chapter 3 – Polytomous Rasch Models

As described earlier, the way of dichotomizing and combining response data used in the Rasch class of polytomous models results in a general expression for the probability of a person responding in a given item category. This expression takes the form of Equation 1.3 which, in practice, expands in the manner shown in the example of Equation 1.4. Masters and Wright (1984) demonstrate, more specifically, that virtually all Rasch polytomous models can take the form

$$P_{i_g}(\theta_n) = \frac{e^{\sum_{g=0}^{l}(\theta - b_{i_g})}}{\sum_{h=0}^{m} e^{\sum_{g=0}^{h}(\theta - b_{i_g})}} \quad , \tag{3.1}$$

where θ_n is the person trait level,

b_{i_g} is the location parameter of the category boundary function for category g of item i,

$l = 0, \ldots, g,$ (see Equation 1.4) and

$h = 0, \ldots, g, \ldots, m.$

Algebraically, the distinction between the different polytomous Rasch models can be described in terms of different expressions that can be used to represent the location parameter (b) of the category boundaries (Masters & Wright, 1984). Table 3.1 lists some of the possibilities. It should provide a useful reference as each model is described in turn.

INSERT TABLE 3.1 ABOUT HERE

Partial Credit Model

Development of the partial credit model (PCM) is associated with the work of Masters (1982; 1988a; 1988b). While it historically follows the development of the rating scale model, the PCM is a more general model and will therefore be introduced first. It is interesting to note that Rasch (1961) himself developed a polytomous model based on the measurement principles underlying his dichotomous model. The PCM however, was not developed as an elaboration of Rasch's polytomous model (Masters 1982).

The PCM is constructed in the manner described above for all ordered polytomous Rasch models. That is, it is built on the successive dichotomization of adjacent categories. In the PCM case this results in a separate location parameter (*b*) for each category boundary (*g*) of each item (*i*), hence the notation for the PCM location parameter in Table 3.1. This approach allows the number of categories to vary from item to item in a test (Masters, 1982).

Masters (1982; 1988a; Wright & Masters, 1982) describes category boundaries between adjacent categories as steps that respondents must complete. These steps are ordered within an item as a consequence of the way in which the item is constructed. The archetypal example is a mathematics item where each part of the problem must be completed in sequence and where each successfully completed part of the problem (step) earns the respondent some credit.

The model is not however restricted to use with items that have response categories defined by component analytical problems. In fact the PCM is suitable for use with any test format which provides a finite set of ordered response options (Masters & Wright, 1997). It should be noted however, that the further one moves from the archetypal, multi-part mathematics problem, the more effort is required to conceptualize the category boundaries as successive steps that must be completed in turn for a respondent to locate themselves in a particular response category.

To illustrate this, consider an example that Masters (1982) provides, of a multiple-choice geography item. The response options have been explicitly ordered in terms of geographical proximity to the correct answer. The question is: "What is the capital of Australia?" The response options are:

a. Wellington … (1pt)
b. Canberra … (3pts)
c. Montreal … (0pts)
d. Sydney … (2pts)

It is difficult to imagine that a respondent who selected the correct answer successively chose Wellington over Montreal, then Sydney over Wellington, and then Canberra over Sydney. Further, it is difficult to accept that respondents with the lowest level of geographical ability would simply select the most wrong answer, Montreal, and fail to progress further. Such a response process is, however, implied by the step conceptualization of the model.

Category "Steps"

Masters (1982) argues that an advantage of the Rasch dichotomization approach is that each CBRF (step) is independent of all others in an item. It transpires however, that even with such items as the archetypal mathematics item, the PCM does not, in fact, validly model a response process characterized by sequential, independent steps between categories within an item. This fact also renders invalid the associated implication that the CBRF location parameters (b_{i_g}) indicate the difficulty, or trait level, of the respective boundaries (Molenaar, 1983; Verhelst & Verstralen, 1997).

Tutz (1990) provides a way to think about why the PCM cannot legitimately model a sequential steps process. In his explanation he focuses on the fact that the PCM defines a step entirely locally, fixed at a point on the response scale that is delineated by the categories *g* and *g*-1. The first step (category boundary) in the PCM is therefore determined by a dichotomous Rasch model which assumes that the person's score is 1 or 2, but no other value. Thus the next "step" from category 2 to category 3 cannot be thought of as taking place after the first step takes place because the first step only allows for the possibility that the outcome is 1 or 2. It does not accommodate the possibility of subsequent responses. This means that to model sequential steps the PCM would, in a sense, need to be able to see into the future and know the outcome of subsequent steps (Tutz, 1990). This explanation neatly encapsulates Verhelst and Verstralen's (1997) demonstration that the PCM models a series of dependent binary variables.

Ultimately Masters (1988a) agrees that the steps notion had been misinterpreted in early PCM research and development, warning that the location parameter of the category boundaries (steps) *cannot* be interpreted as the "difficulties" or location of independent parts of an item. Instead he highlights the important fact that even though the boundary dichotomizations (steps) are locally defined, they occur in the context of an entire item and are therefore not completely independent.

The notion of item steps has nevertheless proved to be stubbornly enduring in the PCM literature, perhaps because of its intuitive appeal. It probably also persists because it was such a central part of the early expositions of the model (e.g., in Masters, 1982; 1988a; Masters & Wright, 1984; Wright & Masters, 1982).

The principal advantage of constructing an ordered polytomous IRT model on the basis of the local comparisons that define the adjacent category dichotomizations (b_{i_g}) is that this approach allows parameter separability (Molenaar, 1983). That is, the person and item parameters can be conditioned out of the estimation of the respective alternate parameter. This occurs because this type of dichotomization, with the associated method for subsequently combining the category boundary information, provides easily obtainable sufficient statistics for each type of parameter (Masters, 1982). The result is that specific objectivity is maintained in this model (Masters, 1982; 1988a; Molenaar, 1983) which, in turn, is why this type of model qualifies as a Rasch model. Glas and Verhelst (1989) suggest that the advantages of having a model with parameter separability, in most cases, outweigh the difficulties arising from the complicated parameter interpretation.

The Mathematical Model

The PCM is described mathematically by the general form for all polytomous Rasch models shown in Equation 1.3. By incorporating a location parameter (b) for each category boundary (g) and each item (i) we obtain a flexible model where categories can vary in number and structure across items within a test (Masters, 1982). That is, even for items with the same number of categories, the boundary locations are free to vary from item to item. As described in the prelude to Table 3.1, the Rasch approach to combining category boundary information results in a general equation for the probability of responding in a specific item category (P_{i_g}) for the PCM, contingent on θ, of the form

$$P_{i_g} = \frac{e^{\sum_{g=0}^{l}\left(\theta - b_{i_g}\right)}}{\sum_{h=0}^{m} e^{\sum_{g=0}^{h}\left(\theta - b_{i_g}\right)}} \quad , \tag{3.1}$$

Each adjacent pair of response categories is modeled by a simple logistic ogive with the same slope (Masters & Wright, 1984; Wright & Masters, 1982) to produce category boundaries that differ only in location. In other words, the CBRFs are obtained by applying the dichotomous Rasch model to each successive, adjacent pair of response categories for an item. The dichotomous model used to obtain CBRFs takes the form

$$P_{i_g|g-1,g} = \frac{e^{(\theta - b_{i_g})}}{1 + e^{(\theta - b_{i_g})}} \quad , \tag{3.2}$$

where, in this case, since CBRFs are being modeled, the response probability is designated $P_{i_g|g-1,g}$ representing the probability of responding in category g rather than g-1, given that the response must be in g or g-1, with g=1,2, …, m. Equation 3.2 is identical to the usual equation for the dichotomous Rasch model except for the subscript on b which differentiates the item category boundaries. Equation 3.2 plus the requirement that a respondent must respond in one of the m+1 categories leads to the general expression for PCM ICRFs shown in Equation 3.1. Furthermore, for notational convenience,

$$\sum_{g=0}^{0} \left(\theta - b_{i_g}\right) \equiv 0 \quad , \tag{3.3}$$

so that,

$$\sum_{g=0}^{h} \left(\theta - b_{i_g}\right) \equiv \sum_{g=1}^{h} \left(\theta - b_{i_g}\right) \quad . \tag{3.4}$$

A complete derivation of Equation 3.1 is given in Masters and Wright (1984). In Equation 3.1, l is the count of the boundary locations up to the category under consideration. The numerator only contains the locations of the boundaries prior to the specific category, g, being modeled. The denominator is the sum of all m+1 possible numerators (Wright & Masters, 1982). An example of what this would look like if one were modeling the third category of a five-category item is shown in Equation 1.4.

The use of simple counts as the basic data for the model is a distinguishing characteristic of Rasch models (Wright & Masters, 1982). This is contrasted with other models that weight category scores, typically by employing a discrimination parameter when modeling category boundaries. The Rasch approach of using simple counts makes it possible to obtain the probability of a given response vector, conditional on θ, without reference to the θ level of the person responsible for the responses (Wright & Masters, 1982). This demonstrates the separability of the person and item parameters and this approach can be extended to an entire response matrix to show that all item and person parameters can be estimated independently of each other (Wright & Masters, 1982). This is part of the notion of specific objectivity.

An interesting consequence of this type of parameterization is that, for any given item calibration dataset, there must be responses in every available response category of an item . Otherwise it is not possible to estimate values for all item parameters (Masters & Wright, 1997). This is a feature of all polytomous Rasch models. It suggests that polytomous Rasch models may suffer in practical testing situations from a similar problem to that described by Drasgow et al. (1995) for the NRM, namely, that it is difficult to implement in sparse data settings. The unusual situation also follows that the different boundary location parameter estimates for an item are typically based on different numbers of respondents. This has interesting implications for differing precision in the estimates of the model's item parameters. Standard errors will always be worse for parameters that model categories with fewer respondents in a manner directly comparable to the situation Tutz (1997) describes for his sequential model.

Information

The notion of statistical item or test information is rarely employed in the Rasch measurement literature where the issue of measurement precision is subordinated to issues surrounding measurement validity. This distinction is blurred by Lord's (1980) description of information as a form of validity. However, common usage in the IRT literature concentrates on information as an indicator of measurement precision while model fit and issues concerning the plausibility of a model for a specific dataset are key elements of the measurement validity discussions more common in the Rasch literature. The de-emphasis of information in the Rasch literature is at least partly due to the fact that, in the dichotomous case, the fixed slope parameter of the Rasch model results in all items having identical information functions. Of course, test information still varies across tests with different items. In any case, Smith (1987) notes that Rasch measurement tends to employ an index called reliability of person separation (R_p) instead of CTT reliability and in preference to test information for assessing measurement precision.

Nevertheless, it is entirely possible to calculate item information for the PCM in exactly the same manner as for other IRT models, that is, as the inverse of the variance of the maximum likelihood estimate of θ, evaluated across θ (Mellenbergh, 1995). Wright and Masters (1982, Chapter 8) provide plots of measurement error at the test level in terms of variance which, as Mellenbergh's (1995) comment indicates, is the inverse of test information.

The earliest use of the concept of statistical information in a PCM context was the work of Dodd (1984). She derived the formulae for test and item information for the PCM based on Birnbaum's (1968) definitions of item and test information functions, and on Samejima's (1969) application of Birnbaum's formulation to the GRM. This amounted to a close approximation of Samejima's ICRF based equations to the PCM ICRFs.

Details of Samejima's (1969) approach to calculating information will be discussed in more detail in the upcoming chapter on her models. There is also another approach to calculating information based on PCM IRFs rather than ICRFs. Explaining the practical steps involved in calculating item information in this alternative approach requires reparameterization of the PCM in a form that differs from Masters' (1982) parameterization of the model. This particular parameterization forms the basis of Andrich's (1978a) development of the RSM. A detailed description of this approach to calculating item information in the polytomous Rasch model context will therefore be provided as part of the later description of the RSM. Dodd and Koch (1994) found that empirically, both approaches produce almost identical results.

Relationship to other IRT models

As one of the most general unidimensional, polytomous Rasch models, the PCM can be considered – conceptually and structurally – as the basis for almost all other unidimensional polytomous Rasch models. This includes the Rating Scale, Binomial Trials, and Poisson Counts models (Wright & Masters, 1982), as well as other models for rating data developed by Andrich (1982) and by Rost (1988a). Furthermore, in the special case of an item with two ordered categories, the PCM becomes the dichotomous Rasch model. These are all divide-by-total models in Thissen and Steinberg's (1986) taxonomy. Among early polytomous Rasch models the only examples that are not in the PCM family of models are the multidimensional polytomous model developed by Rasch (1961) himself and a unidimensional version of this model developed by Andersen (1973). A multidimensional model is simply one that simultaneously models more than one latent trait for each respondent. Recent work however has involved extending polytomous Rasch models in various ways including approaches that generalize Rasch's (1961) early model itself into a much more comprehensive framework. These Rasch model extensions will be introduced in upcoming sections following the introduction of the RSM.

Samejima (1972; 1988; 1996) argues that the PCM itself represents a special case of the NRM. She demonstrates how this is the case in the context of her general framework for items with multiple responses. It is worth noting that Samejima's argument regarding the PCM and the NRM reaches the same conclusion as Mellenbergh (1995), even if the form of their arguments differ.

It is important to be aware that the PCM is not a simplified version of Samejima's (1969) usual graded response model (GRM). This is in marked contrast to the dichotomous case where the Rasch model can be considered a simplified variation on the 2PL and 3PL IRT models.

Variations

Since the PCM is among the most general of the class of Rasch models described in this text it could be argued that all other unidimensional polytomous Rasch models are variations of the PCM. While this may be technically true, the rating scale model (RSM) elaborated by Andrich (1978a) predates the development of the PCM. It also contains features that warrant a separate discussion.

The binomial trials and Poisson counts models however, are true variants of the PCM. The binomial trials model is designed to accommodate data from a response format that allows respondents a fixed number (*m*) of attempts at an item (Wright & Masters, 1982; Masters &

Wright, 1984). The number of successes a respondent has on a given item represents the "category" that they score for the item, such as hitting a target 7 times in ten attempts. The Poisson counts model extends the idea of the binomial trials model to the situation where there is no finite limit to the number of successes (or failures) a person might have at a given task (Wright & Masters, 1982; Masters & Wright, 1984). Rasch (1960) developed this model himself to analyze reading errors, where there was no fixed number of errors that a student could make. These two models modify the PCM by adjusting the way the location parameter is represented in ways analogous to the examples listed in Table 3.1.

Another way to modify the PCM is to insert into the model a positive discrimination parameter (a) which is free to vary from the implied a=1 of the usual PCM. This would allow the discrimination for the boundary location functions (CBRFs) to vary from boundary to boundary. In effect this means modeling the CBRFs with the 2PL dichotomous model instead of the dichotomous Rasch model. Note again that this does not however, result in a Thurstone/Samejima type of polytomous model since it is still only adjacent categories that are being dichotomized.

There are three ways that discrimination parameter values can vary across a set of polytomous items. Most generally, the parameters could vary across boundary locations within an item as well as across equivalent locations in different items – where items have the same number of categories. Alternatively, the discrimination parameter could be fixed to be the same for each equivalent boundary location (g) for all items in a test (Hemker, 1996, Chapter 5). This would again require that all items in a test had the same number of categories. The third alternative is for discriminations to be held constant across CBRFs within an item but for them to be free to vary across items. Items would then again be free to vary in the number of categories per item. This variation of the PCM describes the rationale behind Muraki's (1992) generalized partial credit model (GPCM). This particular variation has also been called Thissen and Steinberg's ordinal model (TSOM) by Maydeu-Olivares, Drasgow and Mead (1994).

The features of the GPCM that pertain to its scoring function formulation will be further explored in the context of the upcoming representation of the RSM. In contrast to the GPCM, the

TSOM (Maydeu-Olivares, et al., 1994; Thissen & Steinberg, 1986) is explicitly formulated in the form of Equation 3.1. It simply substitutes a 2PL model for the dichotomous Rasch model when modeling CBRFs, with the additional constraint that the discrimination parameters be equal across CBRFs within an item.

Wilson (1992; Wilson & Adams, 1993) developed the ordered partition model (OPM) as an extension of the PCM. The OPM is designed to model data that is neither entirely nominal nor completely ordered. In particular, it is envisioned as appropriate for items that have ordered levels, but where there may be a number of nominal response categories within any given level. Wilson suggests that the OPM may be particularly relevant for data resulting from the performance assessment paradigm in education.

Wilson and Masters (1993) developed a procedure for modeling data with the PCM when one or more available item categories are not used. This could occur either when a category is observed to be without responses in an empirical setting, or when a category is predefined to be empty on theoretical grounds in a particular application of an item. Categories that are left unused are called null categories. Such a model may be of considerable importance since, as Ludlow and Haley (1992) note, PCM estimation typically obtains item parameter estimates from one non-zero frequency (non-null) category to the next. Masters and Wright (1984) address this issue by pointing out that all the item parameters for the model can only be estimated if there is data in each of the available response categories. This, in turn, means that the boundary parameters that are estimated may not be those presented to test respondents since null categories are essentially automatically collapsed into subsequent categories in this process.

Wilson and Masters' (1993) modeling of null categories is accomplished by reparameterizing the PCM of Masters (1982) in a manner similar to the way the NRM is parameterized. Specifically, rather than being constructed on the basis of category boundaries, item parameters are provided to model ICRFs directly, using weight and location parameters. The weight parameters are considered to be the slope parameters in the NRM and other similar formulations. The null categories are then modeled by setting the weight and location parameters for those categories to both equal zero. By incorporating null categories into the response model one is able to avoid the necessity for collapsing categories which may be inappropriate for polytomous Rasch models (Jansen & Roskam, 1986; Rasch, 1966; Roskam, 1995, Roskam & Jansen, 1989).

The approach outlined by Wilson and Masters (1993) for handling null categories in the PCM can be extended to the entire family of polytomous Rasch models. A particularly valuable potential use for this approach is in handling sparse data conditions to which polytomous Rasch models may be particularly susceptible.

PCM Summary

Masters and Wright (1997) suggest that the strength of the PCM is its simplicity, which derives from applying a dichotomous Rasch model to each pair of adjacent item categories. The result is a polytomous IRT model with only two parameters, where both are location parameters; one for items b_{i_g} and one for people θ_j. As we have seen, the dichotomizing approach employed in this model makes it sensitive to sparse data conditions which may be common in practice. Furthermore, the resulting item location parameters have ambiguous meaning. A practical consequence of employing this particular type of boundary location parameters can also be seen in the example that Masters and Wright (1997) provide. In their example they show how regions of a trait continuum are defined by different essay graders. However, to do this they must use Thurstone/Samejima CBRF location parameters since the PCM b_{i_g} cannot be used for this purpose.

Masters and Wright (1997) also highlight the fact that the algebraic formulation of the PCM gives it the statistical properties of a Rasch model. This means that it has separable person and item parameters. It is therefore specifically objective within and between person and item parameter comparisons and has the sufficient statistics for all model parameters required for conjoint additivity and fundamental measurement. It could be argued that the possibility of fundamental measurement might outweigh the weaknesses of the PCM, if the model can be shown to fit a given dataset.

Rating Scale Model

Items used in psychological testing contexts such as attitude, interest, personality surveys, inventories, or questionnaires often have a consistent structural form across all of the items for a given measurement instrument. Often this is a Likert or similar type of rating scale format where people are asked to respond to an item using a pre-defined set of responses and where the same set of response alternatives is applied to all the items in the test. A typical example of such a response format is where people are asked to respond to each item on a test using a five-point scale ranging from Strongly Disagree, through Neutral, to Strongly Agree.

The rating scale model (RSM) is uniquely suited to this type of response format. The fundamental assumption of the model, which distinguishes it from other polytomous IRT models, is that the use of a single rating response format across items in a test implies that the response format is intended to function in the same way (i.e. be consistently applied) across all items, since the categories have a constant definition. The implied intent is reflected in the model by the exposition of a common parametric form for all the items using a given rating scale. The result is that, in the RSM, any given category on the response scale is modeled to be the same "size" across all items, although within an item on the scale the different response categories need not be the same size as each other.

Thus, in contrast to the PCM, all items being modeled by the RSM in a given setting must have the same number of categories, and the categories must, in effect, have the same set of labels across the items. Of course, if items in a test use two rating scales with different numbers of categories or the categories have different labels they are, by definition, different rating scales and the basic assumption of the RSM would apply to each scale separately.

Masters (1988a; Masters & Wright, 1984) sometimes describes the RSM in terms of a constrained version of the PCM. While this is formally correct, the RSM also contains several distinctive features. Firstly, the RSM, as an explicated mathematical model predates the PCM by several years. Furthermore, while the PCM was not developed as an elaboration of Rasch's (1961) polytomous model, the RSM is explicitly linked to Rasch's model. Indeed, the model's originators describe the RSM as a special case of Rasch's (1961) multidimensional, polytomous model (Andersen, 1997; Andrich, 1978b).

Another feature of the RSM is that it is presented in a mathematical form that facilitates a fuller understanding of some especially interesting elements of polytomous items and IRT models, specifically with respect to the notion of discrimination. Finally, the RSM is described without reference to the notion of item steps. Instead, category boundaries are described as thresholds, which is the more common term in attitudinal scaling. This change of terminology also helps reduce the temptation to think of the boundaries as embodying independent, sequential dichotomizations, which is so closely attached to the steps concept.

The thresholds still represent locally defined dichotomizations in keeping with the polytomous Rasch approach to defining category boundaries. However, rather than describing these boundaries as independent sequential steps, Andrich (1978a) conceives the response process in terms of respondents making simultaneous, independent, dichotomous choices at each threshold. This leads to some interesting possibilities. For example, when there are four possible responses to a three-category item, a person could respond positively at both thresholds (choose category 3); respond negatively at both thresholds (choose category 1); respond positively at the first and negatively at the second threshold (choose category 2); or respond negatively at the first and positively at the second threshold. This last response set is logically inconsistent and Andrich (1978a) suggests that respondents recognize this and respond accordingly. Naturally the possibilities for logically inconsistent responses grows dramatically as items with more categories are considered.

The Mathematical Model

The RSM is a special case of Rasch's (1961) polytomous model in that it begins with the postulate that a single dimension is modeled in contrast to Rasch's general model which is multidimensional in terms of both the person parameter and with respect to item response categories (Andrich, 1978b; Rasch, 1961). Andersen (1973) built the foundations for the RSM by providing an estimation method for the item parameters of Rasch's (1961) model when it is applied to questionnaire items. This included a discussion of how to obtain the sufficient statistics needed for the estimation procedure.

Andrich (1978a) suggests that his own work complements that of Andersen by providing a substantive interpretation of various parameters in Rasch's (1961) model to go with the statistical properties that Andersen provides. This is important because Rasch (1961) describes his parameters in only the most general theoretical terms. Specifically, Andrich (1978a) describes how the category coefficients of Rasch's model can be interpreted in terms of thresholds on a latent continuum, where thresholds are a form of category boundaries, while Rasch's scoring function is interpreted in terms of discriminations at the thresholds.

Model parameters

Since the PCM was not developed with reference to Rasch's (1961) model we have not yet been introduced to the category coefficient and scoring function parameters. In this section we will introduce these parameters through a rearrangement of model parameters that we have already seen. Recall that from Table 3.1 the general equation for the RSM is obtained as

$$P_{i_g} = \frac{e^{\sum_{g=0}^{l}\left[\theta-\left(b_i+\tau_g\right)\right]}}{\sum_{h=0}^{m} e^{\sum_{g=0}^{h}\left[\theta-\left(b_i+\tau_g\right)\right]}} \quad , \tag{3.5}$$

where h=0, 1, …, g, …, m with g representing the specific category being modeled from among m+1 categories. This shows that in the RSM the usual polytomous Rasch category boundary parameter (b_{i_g}) is decomposed into two elements, b_i and τ_g. The b_i parameter becomes an item location parameter which is estimated for each individual item in a test while the τ_g are the threshold parameters which define the boundary between the categories of the rating scale, relative to each item's trait location. That is, the τ_g tell you how far each category boundary is from the location parameter. These parameters are estimated once for an entire set of items employing a given rating scale, in keeping with the rating scale assumption described above. Given the response format in question and the fact that the τ_g are estimated once for an entire set of items, Andrich (1978b) requires that the threshold parameters be ordered, in contrast to PCM category boundaries.

The form of the RSM given in Equation 3.5 reflects Master's (1988a) conceptualization of the RSM as a constrained version of the PCM. Andrich (1978a) however, shows that it is useful to expand the numerator in Equation 3.5 and that this becomes particularly meaningful if a discrimination parameter (a) is explicitly included. Equation 3.5 then becomes

$$P_{i_g} = \frac{e^{\sum_{g=0}^{l}\left[a_g\left(\theta-\left(b_i+\tau_g\right)\right)\right]}}{\sum_{h=0}^{m} e^{\sum_{g=0}^{h}\left[a_g\left(\theta-\left(b_i+\tau_g\right)\right)\right]}} \quad . \tag{3.6}$$

Note that the discrimination parameter is indexed only with respect to the relevant category boundary (g). Andrich (1978a) points out that it is quite possible, within this mathematical framework, to also index the discrimination by individual items (a_{i_g}). However, since the Rasch model requires items to have the same discrimination this idea is not pursued (Andrich, 1978a).

Expanding the numerator of Equation 3.6 then, produces

$$\sum_{g=0}^{l}\left[a_g\left(\theta-\left(b_i-\tau_g\right)\right)\right] = \sum_{g=0}^{l} -a_g\tau_g + \sum_{g=0}^{l} a_g\left(\theta-b_i\right) \quad . \tag{3.7}$$

If we then define parameters ϕ_g and κ_g as a set of sums of relevant values up to the specific category being modeled at any time, such that

$$\phi_g = \sum_{g=0}^{l} a_g \quad , \tag{3.8}$$

and

$$\kappa_g = -\sum_{g=0}^{l} a_g\tau_g \quad , \tag{3.9}$$

then the general for of the RSM (Equation 3.5) becomes

$$P_{i_g} = \frac{e^{\left[\kappa_g+\phi_g\left(\theta-b_i\right)\right]}}{\sum_{h=0}^{m} e^{\left[\kappa_h+\phi_h\left(\theta-b_i\right)\right]}} \quad , \tag{3.10}$$

where h=0,1,…,g,…,m for m+1 categories;

and $\phi_g = \kappa_g = 0$ for g=0.

Also note that there is now no longer a summation sign in the numerator of Equation 3.10 and only one in the denominator, in contrast to Equation 3.5.

Equation 3.10 is the unidimensional form of Rasch's (1961) polytomous model where the κ_g are the category coefficient parameters and the ϕ_g are the scoring function parameters (Andrich, 1978a; Masters, 1982). The interpretation provided by Andrich (1978a) for these parameters is that Rasch's category coefficients are a negative sum of the products of the threshold locations (τ_g) and their respective discriminations (a_g) up to the category (g) under consideration. The scoring functions, in turn, are simply a sum of the threshold discriminations up to g. More generally, Andersen (1973) notes Rasch's description of the ϕ_g as the scoring values of the response categories, that is, the category scores. Andrich (1978a) here shows what comprises these category scores.

In this form however, Equation 3.10 still describes a very general model since, as Rasch (1961, p. 330; Fischer & Ponocny, 1995) intended, the scoring functions must be estimated from the data. Note that this form of Equation 3.10 also implies that within a legitimate Rasch model, discrimination can vary from threshold to threshold, or in the more general terminology used so far, it is possible for category boundary discriminations to vary within an item.

Equation 3.10 can be simplified by requiring that all threshold discriminations be the same, such that $a_1 = a_2 = \ldots = a_m = a$. Then

$$\phi_g = a \times g \quad , \tag{3.11}$$

and

$$\kappa_g = -\sum_{g=0}^{l} a\tau_g \quad , \tag{3.12}$$

and threshold discrimination becomes a constant so that the scoring function need no longer be estimated. Indeed, the magnitude of a is no longer an issue since it is absorbed into the person, item, and threshold parameters (Andrich, 1978a). The key RSM requirement is simply that the category score values in ϕ increase by a constant, a condition that Equation 3.11 fulfils. Absorbing the value of a into the remaining model parameters essentially fixes the threshold discrimination value equal to 1.0. The category coefficient then simply becomes the negative sum of the threshold location values up to the category in question and the scoring function becomes the number of thresholds crossed, or in our more general terminology, the number of category boundaries traversed. When the m+1 categories are numbered in the usual manner from 0, 1, …, m, then the scoring function values for any given category is simply the category score or category label value. Under this simplification

$$\phi_g = g \quad , \tag{3.13}$$

and

$$\kappa_g = -\sum_{g=0}^{l} \tau_g \quad . \tag{3.14}$$

Equation 3.10 with the scoring functions (ϕ_g) and category coefficients (κ_g) defined as in Equations 3.13 and 3.14 respectively is the usual form of the RSM. Equation 3.13 is described by Andrich (1978a) as an integral scoring function though it has become known in the literature as a linear integer scoring function.

This form of the scoring function is intuitively appealing because it corresponds to the usual Likert scoring approach of weighting successive category responses by the respective category scores, which are usually successive integers. It also has important statistical consequences. Andersen (1973; 1977) showed that the scoring function must take the form of Equation 3.11 for there to be a real valued sufficient statistic for θ that is independent of the item parameters. Since Equation 3.13 also has this form, it has the same property. Andersen (1977, p. 69) calls this "equidistant scoring". Andrich (1978a, 1978b) stresses that the scoring function approach defined in Equation 3.14, where successive categories are scored with successive integers, is a result of having equal discriminations at the thresholds and not equal distances between thresholds.

Sufficient statistics and other considerations

The actual sufficient statistic for θ, is simply a person's total score (Andrich, 1978b), or equivalently in this formulation, the number of thresholds, across all items, that the person has crossed (Andrich, 1978a). Furthermore, the total number of responses, across all people and all items, in the category *g* is a sufficient statistic for κ_g. Also, the sum of all the responses to an item weighted by category score, that is, the total score for an item *i*, is a sufficient statistic for b_i, the item location parameter.

Andersen (1983) presents a more general RSM model which does not require category scores to increase by a constant value. Andersen (1983; 1995) is careful to point out however, that in more general Rasch models, with unequal threshold discriminations resulting in category scores that do not increase uniformly, one must have known category scores. To be more specific, it is not possible to estimate the category score values (scoring function) within the Rasch framework. This is because it is not possible to obtain the requisite distribution free sufficient statistics for item parameter estimation unless the values of a_g are known, or prespecified. Of course, the usual RSM linear integer scoring function involves prespecified discrimination values. They are essentially prespecified to equal 1.0. It is interesting to note that when the scoring function is not the linear integer function, that ICRFs are no longer symmetrical. Furthermore, category boundaries are now also no longer the location on the trait continuum where there is an equal probability of responding in either of the adjacent categories (Muraki, 1993)

Another feature of the scoring function formulation of the RSM is that it sheds light on the notion of collapsing categories. Andersen (1977) shows that having the same category score for two categories collapses the categories. Andrich (1978a) points out that this situation, where $\phi_{g+1} = \phi_g$, results when $a_{g+1} = 0$, meaning that the threshold does not discriminate and the categories on either side can therefore be pooled. This is a consequence of the fact that the ϕ effectively function to order the answer categories (Andersen, 1973). If successive ϕ values are not ordered, the answer categories will not be ordered, but will rather be nominal categories in the sense of Bock's (1972) NRM.

The process of elaborating the scoring function formulation of the RSM, that is, going from Equation 3.5 to Equation 3.10, therefore helps clarify the role of category scores, thresholds, and of item and threshold discrimination in polytomous Rasch models. It also shows that it is possible to reparameterize formally identical models in strikingly different ways.

In summary, the RSM is composed of a location parameter for each item (b_i) and a set of parameters for the response alternatives across all items in a test or questionnaire (τ_g) which require estimation. Thus the RSM has fewer parameters to estimate than the PCM [n item location parameters + m threshold (category boundary) parameters versus $n \times m$ item (category boundary) parameters for the PCM]. Masters (1988a) warns that this parsimony may come at the cost of lost information about interactions between items and response alternatives that may be subtle but psychologically interesting. It also means that the only modeled difference between items with the RSM is their location, b_i, on the θ scale (Masters, 1988a). This is shown graphically for two hypothetical, four-category items in Figure 3.1. The ends of the arrows indicate the category boundary locations relative to each item's location. Note that each boundary (τ_g) is the same distance from the location parameter of each item. The specific trait scale location for each boundary is indicated on the ordinate by b_{i_g} which also represents the point on the θ scale at which the ICRFs for the relevant categories would intersect, were they plotted.

INSERT FIGURE 3.1 ABOUT HERE

Masters (1988b) suggests that, in addition to modeling rating scale items, the RSM might also be useful when there is insufficient data to provide reliable estimates for all the parameters required to model the PCM. Wilson (1988) notes that applying the RSM to items that have no data in some of their categories (typically extreme categories) essentially results in the model imputing parameters for those categories from data from the other items in the analysis. This, of course, is the case whether one is attempting to resolve PCM parameter estimation problems as described by Masters (1988b) or when one is legitimately applying the RSM to sparse rating scale type data.

Information

Understanding the concept of information with respect to polytomous Rasch models is facilitated by the scoring function formulation of these models. While this formulation has, thus far, only been shown with respect to the RSM, its application to the PCM and the GPCM, follow in a straightforward manner, as will be shown later in the RSM variations section. A useful starting point with respect to information is the realization that the scoring function formulation of these models allows us to easily calculate an IRF for a polytomous item.

Expected values and response functions

The ease of calculating a polytomous item IRF follows from the fact that an IRF can be thought of as describing the rate of change of the expected value of an item response, as a function of the change in θ relative to an item's location b_i (Andrich, 1988b). More succinctly, this can be thought of as a regression of the item score onto the trait scale (Chang & Mazzeo, 1994; Lord, 1980).

In the case of a dichotomous item, the expected value of a score for an item is the regression of the positive response on θ, which is the usual IRF for that item (e.g., Figure 1.1). From this perspective, a negative response equates with failure to score. In more formal terms, the expected value of response *x* for item *i* is

$$E\left[X_i\right] = \left(P_i = 1 | \theta\right) \quad . \tag{3.15}$$

As noted earlier, in dichotomous IRT models the IRF provides the probability of responding in a given category, which function is performed by the ICRFs in polytomous IRT models. However, if we maintain the regression, or the expected value, description of an IRF then it is possible to obtain an IRF for polytomous items also. This is a single function, generally monotonically increasing in θ, which describes the "conditional mean of item scores at a given θ" (Muraki, 1997, p. 156). The scoring function for a polytomous item makes it a simple matter to obtain the function for this conditional mean/IRF/expected value. When the usual linear integer scoring function is employed, the expected value of response *x* (where $x \in g = 0,\ 1,\ \ldots,\ m$) for item *i* is

$$E\left[X_i\right] = \sum_g g P_{i_g}(\theta) \quad , \tag{3.16}$$

that is, it is a weighted sum of ICRF values (Chang & Mazzeo, 1994) where the ICRF values are weighted by the scoring function.

As we will see later, one may have reason to use a scoring function other than the usual linear integer function. In that case a more general form of the expected value is

$$E\left[T_i\right] = \sum_{g=1}^{m} T_g P_{i_g}(\theta) = \overline{T}_i(\theta) \quad , \tag{3.17}$$

where T_g is the, not necessarily linear integer, scoring function for categories g=0, 1, …, m, and $\overline{T}_i(\theta)$ is the notation for item i's conditional mean given θ, that is, its IRF.

Response functions and information

Andrich (1988b) notes that the discrimination of a polytomous item, at any value of θ, is the rate of change of the IRF given by $E\left[X_i\right]$, with respect to θ. Item information in the IRT context is directly related to item discrimination, such that, $I_i(\theta)$= Squared IRF Slope/Conditional Variance, where $I_i(\theta)$ represents item information and IRF slope is synonymous with discrimination. Now that we have a ready method for obtaining an IRF for a polytomous item the equation for item information follows directly.

Masters and Evans (1986) give this form of the equation for item information as

$$\frac{\partial E\left[X_i\right]}{\partial \theta} = V\left[X_i\right] = \sum_{g=1}^{m} T_g^{\,2} P_{i_g} - \left(\sum_{g=1}^{m} T_g P_{i_g}\right)^2 \quad , \tag{3.18}$$

where the squared component on the right side of the difference symbol can be recognized from Equation 3.17 as the IRF. Category information can then be obtained as a partition of the item information due to a given category (Muraki, 1993). That is

$$I_{i_g}(\theta) = P_{i_g}(\theta) I_i(\theta) \quad . \tag{3.19}$$

Muraki (1993) provides an expanded version of Equation 3.18 incorporating the notion that item information is a sum of category information for the categories of an item. Thus Equation 3.18

is multiplied by P_{i_g}, giving category information by Equation 3.19, and then summed across categories to give

$$I_i(\theta) = \sum_{g=0}^{m} P_{i_g}(\theta)\left[\sum_{g=0}^{m} T_g^2 P_{i_g} - \left(\sum_{g=0}^{m} T_g P_{i_g}\right)^2\right] \quad , \tag{3.20}$$

$$= \sum_{g=0}^{m}\left[T_g - \overline{T}_i(\theta)\right]^2 P_{i_g} \quad , \tag{3.21}$$

where Equation 3.21 is a simplification of Equation 3.20 (Muraki, 1993).

These two equations describe item information for a polytomous model which does not have a separate item discrimination parameter, such as the PCM. In the case of models such as the GPCM which have an additional item discrimination parameter, item information is obtained by multiplying either equation by item discrimination squared, that is a_i^2 (Muraki, 1993). Similarly, these equations all have information functions on the logistic metric. If equivalence with the normal ogive metric is desired then these equations must also be multiplied by the square of the scaling factor, D (where D=1.702). So, for example, item information in the form of Equation 3.21 modified for items with an item discrimination parameter and also modified to place information on the normal ogive metric would take the form

$$= D^2 a_i^2 \sum_{g=0}^{m}\left[T_g - \overline{T}_i(\theta)\right]^2 P_{i_g} \quad . \tag{3.22}$$

Although the notion of statistical information is not a feature of usual Rasch model discussion, Andrich (1978b; 1979) does refer to the large sample standard errors of maximum likelihood parameter estimates for the RSM. In the case of the person parameter estimate, this is simply the negative inverse of the square root of the usual item information function. Andrich's discussion of the standard errors serves to highlight the possibility of obtaining information functions for any model parameter (not just θ). This is an issue rarely discussed in any IRT literature, though conceptually, this approach would provide an indication of the "measurement" precision of item, as well as the usual person, parameters. Interestingly, more recent discussions of generalized Rasch models (e.g., Adams & Wilson, 1996) explicitly refer to the information matrix as the source of the asymptotic standard errors for their parameter estimates.

Relationship to Other IRT Models

Two additional points pertain to the RSM's relationships to other IRT models which do not apply directly to the PCM in its usual form as expounded by Masters (1982; 1988b; Masters & Wright, 1984; 1997). Both points follow from the fact that the RSM was developed using a scoring function formulation. The first point is that the RSM is directly and explicitly related to Rasch's (1961) polytomous model while the PCM is not. The earlier discussion of how the PCM (Equation 3.5) can be developed to become the RSM (Equation 3.10) suggests however, that the PCM's separation from Rasch's model is somewhat artificial.

Secondly the scoring function formulation of the RSM makes more obvious the connection between the polytomous Rasch models and the NRM. The connection is simply that if the scoring function is considered to be a set of estimable parameters which need not be ordered, this is the NRM. That is, estimating a separate discrimination parameter for each item category and allowing those parameters to be unordered describes the process for obtaining parameters for nominal item categories. This is not as clear when discussing the RSM scoring function since the RSM requires ordered thresholds as a part of its modeling of ordered categories. Recall that the PCM does not require ordered category boundaries. Thus the notion of nominalizing response categories might be more intuitively obvious in the context of a scoring function formulation of the PCM.

PCM scoring function formulation and the NRM

Andrich (1988b) shows how the PCM can be expressed in terms of the scoring function formulation that he uses for the RSM, in a manner that is formally equivalent to Muraki's GPCM approach to be described in more detail in the next section. He shows that the PCM is simply Equation 3.10 with the τ_{i_g} allowed to vary across items, and therefore, estimated separately for each item rather than once, as a set, for all items in a test. This leads Andrich (1988b) to relabel the PCM as the extended logistic model (ELM)[2]. It also shows that just as the RSM can be formulated in terms of the usual PCM formulation (Table 3.1), so the PCM can be formulated in terms of a scoring function and category coefficients, where the category boundaries are defined with respect to a single item location for each item. Andrich also makes the important point that simply reformulating the PCM CBRFs in terms of thresholds does not change the fact that these locations cannot be considered separately but must be considered together, within an item.

Aside from requiring the τ_{i_g} to be estimated separately for each item, Andrich's (1988b) ELM differs from the RSM in that it does not require the τ_g to be ordered. Andrich (1988b) does however, make a convincing argument for the need to avoid reversed category boundaries in ordered polytomous models, irrespective of whether they are in b_{i_g} or τ_{i_g} form.

The scoring function equivalent of the PCM shown in Equation 3.1 therefore is simply

$$P_{i_g} = \frac{e^{\left[\kappa_{i_g} + \phi_g(\theta - b_i)\right]}}{\sum_{h=0}^{m} e^{\left[\kappa_{i_h} + \phi_h(\theta - b_i)\right]}} \quad , \tag{3.23}$$

where

$$\kappa_{i_g} = -\sum_{g=0}^{l} \tau_{i_g} \quad , \tag{3.24}$$

with τ_{i_g} now estimated separately for each item and which need not be ordered, and where ϕ_g is the usual linear integer scoring function.

If the threshold discrimination parameters a_g are estimated rather than being stipulated equal to 1.0, and are also allowed to be unordered, Equation 3.23 becomes

$$P_{i_g} = \frac{e^{\left[\kappa_{i_g} + \phi_{i_g}(\theta - b_i)\right]}}{\sum_{h=0}^{m} e^{\left[\kappa_{i_h} + \phi_{i_h}(\theta - b_i)\right]}} \quad , \tag{3.25}$$

now with

$$\kappa_{i_g} = -\sum_{g=0}^{l} a_{i_g} \tau_{i_g} \quad , \tag{3.26}$$

and

$$\phi_{i_g} = \sum_{g=0}^{l} a_{i_g} \quad . \tag{3.27}$$

Equation 3.25 is the scoring function formulation of the NRM.

Reformulating the PCM in terms of a scoring function, as in Equation 3.23, also allows Andrich (1988b) to explore a key difference between dichotomous and polytomous Rasch models. He notes that although the dichotomous Rasch model IRFs are constrained to have the same discrimination, this need not be the case for PCM (or RSM) IRFs. Appreciating this difference requires understanding the earlier discussion on how an IRF is obtained for a polytomous item by means of the expected score. The reason the difference exists is that the "distances" between the thresholds, for items modeled by both the PCM and RSM, affect the items' discrimination. The further apart the τ_{i_g} values, when the thresholds are in the same order as the item categories, the weaker the discrimination. The converse is also true, as is the fact that when thresholds are reversed the item discriminates even more sharply than for naturally ordered categories (Andrich, 1988b). This description of discrimination echoes Muraki's (1992) comment that the discrimination of a polytomous item is composed of the item's slope parameter and the set of threshold parameters for the item.

The key difference between dichotomous and polytomous Rasch models, described by Andrich (1988b) therefore, centers on the respective item location parameters. A dichotomous item response only allows the estimation of a single location parameter. In contrast, when a subset of dichotomous items, such as a testlet, or when a polytomous item is being modeled by the PCM, it becomes possible to estimate a location and the distance between two items' or two categories' thresholds. An implication of this distinction is that dichotomous IRT models that allow discrimination to vary from item to item, e.g., 2PL, may be masking item dependencies in a test (see e.g., Masters, 1988c), with the associated risks that follow from a breakdown of the local independence assumption.

Variations

Andrich elaborations

A number of variations of the RSM are possible by virtue of the fact that the relationships among thresholds can be constrained in different ways, both within and across items. Andrich (1978a), in his original explication of the RSM, showed how to modify the RSM so that all thresholds are the same distance apart. The important point he made there was that equidistant thresholds were a function of the form of the category coefficients and not due to the integer scoring function. Andrich (1982) generalized this early demonstration model by showing that it resulted from adding a dispersion parameter to the formulation of the category coefficient. This dispersion parameter can be easily estimated and while it results in equal distances, need not result in unit distances between thresholds. The important role of this dispersion parameter together with the fact that thresholds are defined, in the usual RSM manner, relative to a central item location parameter lead Andrich to refer to this as the dispersion, location model (DLM).

Andrich (1982) also showed that the DLM can be applied in two distinct ways. If one retains the rating scale assumption that all rating items in a test should function in the same manner, then only one dispersion parameter value need be estimated for an entire test. The result is that the distance between all the thresholds within an item would be modeled to be the same, and this distance would also apply across all items in a test. In other words, the size of every category within and across items using a specific rating scale would be the same, though they would not necessarily be one scale unit in size. This could be called a rating scale version of the DLM (RS-DLM).

The second way to apply the DLM is to estimate a separate dispersion parameter for every item in a test. In this case all categories within an item are modeled to be the same size but the category size can vary across items. While this variation can still apply to items that have the same number of categories, the rating scale assumption clearly no longer holds since categories are no longer functioning in the same way across items. Andrich (1982) notes that the DLM variations that allow category size to vary across items require items with at least three ordered categories. The dispersion parameter cannot be constructed for a dichotomous response.

The algebraic form of the DLM clearly shows the role of the dispersion parameter. This is

$$P_{i_g} = \frac{e^{\left[\phi_g(\theta - b_i) + g(m-g)\lambda_i\right]}}{\sum_{g=0}^{m} e^{\left[\phi_g(\theta - b_i) + g(m-g)\lambda_i\right]}} \quad , \tag{3.28}$$

where the quadratic function *g(m-g)* multiplied by the dispersion parameter λ_i replaces the usual RSM form of the category coefficients κ_g. Note that when a single dispersion parameter is estimated for an entire dataset then the subscript *i* is not required on λ.

Obtaining threshold (τ_{i_g}) values from the DLM is different to the case of the RSM where threshold values are obtained simply by subtracting successive category coefficient values such that

$$\tau_g = \kappa_{g-1} - \kappa_g \tag{3.29}$$

with

$$\tau_1 = 0 - \kappa_1 \quad , \tag{3.30}$$

The process differs for the DLM since the category coefficient component is no longer simply a sum of thresholds. Instead, Müller (1987) shows that DLM threshold values can be obtained from the simple formula

$$\tau_{i_g} = 2\lambda_i\left(g - \frac{m+1}{2}\right) \quad . \tag{3.31}$$

Andrich (1982) shows that the dispersion parameter (λ_i) of the DLM can indicate the plausibility of the response distribution for an item. This essentially provides an assessment of model plausibility and follows directly from the value of λ_i. If it is greater than zero the response distribution is unimodal and suggests a plausible response process. If it equals zero then responses are distributed uniformly and response dependencies are implied (Andrich, 1985a). However, if the dispersion parameter's value is below zero, the response distribution is U-shaped, suggesting that the model does not fit the response data. A negative dispersion parameter is directly related to the issue of reversed category boundaries, which Andrich (1988b) argues is problematic and which contravenes a basic requirement of the rating models elaborated by Andrich.

Andrich (1985b) introduces the idea that additional parameters can be incorporated into the DLM to describe more aspects of category structure. Specifically, he introduces a third item parameter, η_i, that indexes skew. This parameter can be interpreted as characterizing response set when applied to questionnaire data, and is called an asymmetry parameter (Andrich, 1985b). The skew parameter is, logically, scored by a cubic function such that the category coefficients, κ_g, of the RSM now take the form

$$\kappa_g = g(m - g)\lambda_i + g(m - g)(2g - m)\eta_i \quad . \tag{3.32}$$

Rost elaborations

Rost (1988a) develops a model that attempts to address the original concerns of Thurstone's method of successive intervals (Edwards & Thurstone, 1952; Rimoldi & Hormaeche, 1955), which he identifies as a need to scale category widths while allowing category boundaries to be dispersed differently from item to item. Rost notes that the RSM scales category boundaries while the DLM addresses variable boundary dispersions but neither accommodates both factors.

Rost's (1988a) solution is to develop the successive intervals model (SIM) by combining the RSM and DLM approaches. That is, a set of thresholds is estimated for the entire set of data from a given rating scale (as per RSM). Additionally, a dispersion parameter is estimated for each item (as per DLM). Then the final, unique, threshold locations for each item are a combination of these two item parameters. Thus thresholds are not constant across items, since a unique item dispersion parameter modifies the baseline rating scale condition. Neither are they constant within an item since the dispersion modification is to RSM-type thresholds which themselves have variable interval sizes within items. This produces a general model that accommodates varying threshold locations across and within items while still being considerably more parsimonious with regard to the total number of parameters to be estimated than the PCM.

For completeness it should be noted that Rost's (1988a) SIM is part of a much broader body of work combining latent class analysis with IRT. Through a series of expositions Rost (1988b; 1988c; 1990; 1991) incorporates increasingly less restricted versions of dichotomous and polytomous Rasch models into the latent class modeling framework. Rasch models were chosen for the IRT component because the approach relies on obtaining the item parameters without reference to the trait distribution (Rost, 1990), and this can only be done if sufficient statistics are available for the parameters. The approach is intriguing since it provides a method for quantitatively measuring people, using Rasch models, within qualitatively different groups, defined as latent classes (Rost, 1991).

Other elaborations

A Rasch model for continuous ratings, such as those produced by a graphic rating scale, was developed by Müller (1987). Müller conceived of continuous rating scales as a limiting case of ordered categorical rating scales where the number of categories becomes infinitely large. The continuous model developed by Müller is an elaboration of the DLM outlined by Andrich (1982) and is obtained by using a more general scoring formula than the usual integer scoring function.

Another set of polytomous Rasch models has also been developed in the context of the scoring function formulation used by Rasch (1961) and elaborated by Andrich (1978a). These models were developed by Fischer because there were no appropriate methods for assessing, or testing hypotheses about, treatment effects with ordered polytomous models (Fischer & Ponocny, 1994; 1995).

The approach taken in this set of models is to decompose the item parameters into linear components called basic parameters. The basic parameters are defined generally enough that they can represent any experimental factor or structural property of an item that is expected to influence the item's trait scale location (Fischer & Parzer, 1991; Fischer & Ponocny, 1994). Thus they can represent, for example, cognitive operations involved in responding to an item, the effect of some treatment or of learning on a trait, or the effect of advertising or some other communication on an attitude (Fischer & Ponocny, 1994).

Generalized Partial Credit Model

In the previous section some RSM variations were described that resulted from constraining threshold relationships in different ways. An alternative is to take advantage of Andrich's introduction of a discrimination parameter in Equation 3.6 and develop variations of the RSM by allowing this parameter to vary across items. This has important implications for the status of the resulting models as Rasch models.

The major variant of the RSM that takes advantage of a discrimination parameter that varies across items is the GPCM, which was introduced briefly in the earlier section on PCM variants. The place of Muraki's (1992) GPCM as a variant not only of the PCM (Masters, 1982) but also of the RSM (Andrich, 1978a) is demonstrated by Muraki's (1997) representation of the GPCM in terms of the scoring function formulation of polytomous Rasch models that is used by Andrich.

The equation for this formulation of the GPCM is

$$P_{i_g} = \frac{e^{a_i\left[\phi_g(\theta - b_i) + \sum_{g=1}^{i} \tau_g\right]}}{\sum_{h=1}^{m} e^{a_i\left[\phi_h(\theta - b_i) + \sum_{g=1}^{h} \tau_g\right]}} \quad , \tag{3.33}$$

where:

a_i is the slope parameter representing item discrimination (one per item);

b_i is the item location parameter (one per item);

ϕ_g is the usual scoring function (defined here as a scaling factor) that equals the category count for the specific category (g) being modeled; and

τ_{i_g} is the threshold parameter representing the category boundary locations relative to the item location parameter b_i. Note that Muraki does not represent GPCM summed thresholds as a category coefficient (κ_{i_g}) the way Andrich does for the RSM.

Muraki (1997) argues that allowing non-uniform discrimination across all items in a test means that the GPCM allows more insight into the characteristics of the test items than does the PCM. An important attraction of the GPCM is that the twin tactics of adding an item discrimination parameter and of employing the scoring function formulation to describe the model, make it a flexible framework for describing a wide range of polytomous IRT models. All of the potential models are, of course, divide-by-total models in Thissen and Steinberg's (1986) terminology. Nevertheless, by constraining the discrimination (a_i), scoring function (ϕ_g), and threshold (τ_{i_g}) parameters of test items in different ways, a collection of IRT models can be described. This strategy can also help illuminate some of the relationships among different models as well as further demonstrate some of the enormous flexibility of the polytomous item format. Some examples of this strategy are described below.

In Equation 3.33, when the slope parameter (a_i) is fixed equal to 1.0 and a set of threshold parameters (τ_{i_g}) is estimated separately for each item, with the additional condition that these parameters need not be ordered on θ, this becomes the scoring function formulation of the equation for the PCM (Masters, 1982). This is also identical to the ELM described by Andrich (1988b).

When the a_i are free to vary across items, but the τ_g are constrained to be ordered according to the item category ordering and are estimated once for all items in a set of items, then Equation 3.33 becomes the RSM (Andrich, 1978a) with a varying slope parameter; in a sense a generalized rating scale model.

A systematic listing of these model variations is provided in Table 3.2 which replicates the intent of Table 3.1 from the beginning of the section on Polytomous Rasch models. That is, it shows how different models can be obtained from a single general model. In this case however, it does so using the scoring function formulation of the various models based on Muraki's (1992) general equation (Equation 3.33) above. This formulation allows non-Rasch models to also be included in this pseudo-taxonomy.

INSERT TABLE 3.2 ABOUT HERE

It is important to remember that the RSM, the PCM, and by extension the GPCM, only become models for ordered categorical responses when the scoring function is increasing (Muraki, 1997). This means that it must be the case that $\phi_h > \phi_{h-1}$ for any category *h* within an item. Furthermore, the discrimination parameter, a_{i_g}, must also be greater than 0. If a_{i_g} is allowed to be negative the categories become nominal and the PCM becomes the NRM (Hemker, 1996).

Discrimination and Polytomous Rasch Models

By introducing a discrimination parameter Andrich (1978a) provides a substantive interpretation for the category coefficient κ_g and scoring function ϕ_g of Rasch's (1961) polytomous model. The interpretation however, rests on two important points, as Andrich (1978a) makes clear. Firstly, in Rasch models, discrimination is only indexed by threshold, a_g , and not by item and threshold a_{i_g} . Secondly, these threshold discriminations are not estimated but rather, must be a constant so that the scoring function increases by the same amount from category to category. This allows independent, minimally sufficient statistics (Andersen, 1977) and when this constant is set equal to 1.0 the usual linear integer scoring function ensues.

Muraki (1992) reintroduces item discrimination but now as a separate entity (a_i) rather than in combination with threshold discrimination (a_{i_g}). This allows Muraki to retain the linear integer scoring function and its interpretation in terms of threshold discrimination and category scores while providing a means for modeling varying discrimination on an item by item basis.

In addition to the separation of item and threshold discrimination, a key issue in the foregoing discussion is whether or not the parameter is estimated. If either the item or threshold parameters are to be estimated, where the former is the case with the GPCM, the resulting model is not a Rasch model. This is because requiring a discrimination parameter to be estimated precludes the possibility of obtaining the sufficient statistics required for parameter separation, in exactly the same manner as with the 2PL model in the dichotomous case. Thus the GPCM, and similarly any generalized RSM, are not Rasch models, any more than the NRM, even though they all have the divide-by-total form to their model equations.

Summary of Polytomous Rasch Models

The Polytomous Rasch model strategy of local dichotomizations combined in a divide-by-total framework results in models that are specifically objective. That is, they are objective because two items can be compared independently of which people responded to the items and independently of any other items in a test, while similarly, two people can be compared independently of the items they responded to and of which other people could be compared (Rasch, 1961). Further, they are specifically objective because the comparisons take place within a specified frame of reference, usually with respect to a specific trait. Specific objectivity is, in practical terms, inextricably linked to the availability of sufficient statistics for the model parameters (Andrich, 1996).

The partial credit model (PCM) and the rating scale model (RSM) are the most prominent manifestations of polytomous Rasch models. Both can be presented in terms of a scoring function formulation, although the PCM was not developed using that formulation. When presented in terms of a scoring function the PCM is often referred to as the extended logistic model (ELM) reflecting its status as a general model where no restrictions are made on the number of, or relationships among, item categories across items in a test.

Aside from differences in the form of the general equations within which the PCM and RSM were initially presented, and their different levels of generality, the two models also differ in terms of the response process posited to underlie them. It has been suggested the PCM models a series of sequential dichotomous steps, represented by the category boundaries, whereas the RSM has been suggested to model the entire set of dichotomous choices represented by the category boundaries simultaneously. While the RSM process avoids the implication associated with the sequential steps approach that requires each step to incorporate information about what occurs in subsequent steps, it retains the problem of ambiguously defined category boundary location parameters endemic to polytomous Rasch models. This ambiguity arises from the fact that even though category boundaries are locally defined, they occur in the context of a polytomous item and the location of any given boundary is therefore influenced by the locations of adjacent boundaries – which in turn have their locations influenced by the locations of boundaries adjacent to them. It is an interesting irony for polytomous Rasch models that, while they arguably have no plausible response process underlying them (van Engelenburg, 1997), they are the only polytomous IRT models that always result in stochastically ordered trait estimates (Hemker, 1996).

The local dichotomization approach of polytomous Rasch models also results in the situation where category boundary locations need not be in the same order on the trait continuum as the categories they separate. This has sometimes been described as a strength of the polytomous Rasch models, particularly in contrast to the Thurstone/Samejima approach (Masters, 1982), or at least as perfectly understandable (Adams, 1988). Boundary reversals may not however, actually be desirable.

For example, Andrich (1988b) notes that whenever category boundaries are reversed there will be one or more categories that are never the most probable. So, while reversed boundaries might be used as a diagnostic tool to help provide understanding of the cognitive processes at work in responding to an item, they probably indicate that something is wrong with the item (Andrich, 1988b). Verhelst, Glas and de Vries (1997) however, suggest that, given the ambiguous nature of the boundary location parameters in polytomous Rasch models, they are, in fact, unhelpful for diagnostic purposes.

This chapter on polytomous Rasch models began by introducing the PCM as a potential model for a large range of polytomous item types in the social sciences, including those used in personality, cognitive, and attitude measurement. It can also apply to non-independent dichotomous item sets such as those used in item bundles and testlets. Furthermore, the model accommodates items with different numbers of categories. Discussion then turned to the possibility of modifying the PCM by allowing discrimination to vary, which in one case leads to the GPCM. Variations of the PCM were also introduced which, in one case, allowed the modeling of null categories and, in another, the inclusion of multiple, nominal categories within an ordered polytomous item (i.e., the OPM).

Introduction of the RSM was notable because it provides an alternative response rationale to the contentious steps notion associated with the PCM, and because it elaborates an alternative parameterization of Rasch models, namely, the scoring function formulation. This formulation helps clarify the sources of polytomous Rasch model flexibility, specifically in terms of the scoring function and category coefficient parameters. It also highlights the context in which simple sufficient statistics for model parameters arise.

By reformulating the PCM in its scoring function formulation as the ELM, it was shown that different models could be obtained by constraining the thresholds in different ways through alternate definitions of the category coefficient. This led to such models as the “classic” RSM, as well as the DLM, the DSLM, and the SIM.

Perhaps more importantly, the scoring function formulation of Rasch models makes more obvious the fact that discrimination can occur at an item level or at a category boundary level in polytomous items. Elaborating the role of the scoring function in discrimination helped clarify such issues as the link between the NRM and the polytomous Rasch models, and the prerequisites needed for sufficient statistics to be available, particularly through the work of Andersen (1977; 1983).

The flexibility in polytomous Rasch models is enhanced by the technique of decomposing item location parameters into, once again predetermined, linear composites, as in the linear logistic models of Fischer (Fischer & Parzer, 1991; Fisher & Ponocny, 1994; 1995). Rost’s (1988b; 1988c; 1990; 1991) work integrating latent class analysis with individual differences measurement by means of polytomous Rasch models is another example of the flexibility of the basic Rasch approach.

Three Practical Examples

The remaining practical examples (PCM, RSM, GPCM; and in the next chapter, the GRM) will all use the same data in order to highlight similarities and differences in the four models in the context of a common set of data. The data come from a survey of people’s notions of morality which obtained responses using a 5-point Likert-type rating scale. Screening of the initial survey data resulted in a final set of 999 responses to 93 items. Responses were primarily obtained from students at a large Midwestern university. A dimensionality analysis of this data produced a 12 factor Principal Axis factor analysis solution which was rotated by Varimax rotation. Twelve items were identified as loading on the first factor of this 12 factor solution. This set of 12 items was selected for IRT analysis as it represents the strongest single construct dimension of the full dataset.

The items in this dataset contained responses in every response category. Category response frequencies ranged from a minimum of 24 to a maximum of 397. IRT analysis was conducted using Parscale (Muraki & Bock, 1999) for all of the practical examples using this data. This reduces the prospect of model variation due to variation in software parameter estimation routines. Item 4 was selected from the 12 items in this first dimension for use in the practical examples, primarily because it was located in the middle of the trait range, thus enhancing the presentation of the response functions.

PCM

Only one item parameter (b_{i_g}) is estimated in the PCM. However, this boundary location parameter must be estimated separately for each boundary of every item in a test or questionnaire. Note that in the adjacent category models the category boundary parameter values need not be ordered in the same sequence as are the categories, since adjacent category boundaries involve local dichotomizations. However, when boundaries are out of order this indicates that at least one item category will never be the most probable response for respondents at any point on the trait scale. It is also considered an indication of an item that is not functioning as intended (Andrich, 1988b). As we discuss specific parameter values it is also important to remember from the earlier "Item Steps" discussion that adjacent category model b_{i_g} do not actually model independent boundary locations, trait levels or "difficulties" since the local boundary dichotomizations occur in the context of an entire item. Discussion of category boundary locations below therefore refers to non-independent item parameters.

Table 3.3 contains the four b_{i_g} values that describe the PCM (non-independent) CBRF locations for our example item. Note that for this item the boundaries are sequentially ordered, indicating a well-functioning item. The ICRFs for this item are shown in Figure 3.2. This figure shows that all five categories have some segment of the trait scale along which a response in the respective categories is more probable than for any other category. The category with the shortest such segment is Category 3. In practice, the middle category in Likert-type items with odd numbers of categories is the category that most commonly is found to be "never most probable". This occurrence may be an indication that the middle category is being used by respondents as a "Don't

Know" or "Can't Decide" category instead of operating as a legitimate measure of the middle part of the trait continuum for an item.

INSERT TABLE 3.3 ABOUT HERE

INSERT FIGURE 3.2 ABOUT HERE

The middle category of the example item may be demonstrating some effects of responses that are not legitimately trait-indicating, since its ICRF's peak is lower than those of adjacent categories. However, it is also legitimately measuring the middle of the trait continuum for some respondents. This is suggested by the fact that it is the most probably response category for respondents with $-0.250 < \theta < 0.375$. For respondents in the region of the trait continuum immediately below $\theta = -0.250$ a response in Category 2 becomes most probable, while immediately above $\theta = 0.375$ a Category 4 response becomes most probable.

To facilitate comparisons across models specific response probabilities are provided in Table 3.4, for each item category, at two arbitrary points on the trait continuum ($\theta = -1.5$ and $\theta = 1.5$). Table 3.4 shows that, as modeled by the PCM, respondents at $\theta = -1.5$ have the highest probability of responding in Category 2 , while Category 5 is the most probable response category for those at $\theta = 1.5$.

INSERT TABLE 3.4 ABOUT HERE

RSM

Two types of item parameters are estimated for the RSM. Table 3.3 shows the relevant item location (b_i) and threshold location (τ_g) parameters for our example item. A total of five parameters have been estimated for this item, in contrast to only four b_{i_g} parameters for the PCM. Recall however, that the set of four τ_g shown here is only estimated once for all 12 of the items in the questionnaire dimension from which this example item was taken, and only the b_i parameter is estimated separately for each item. In general, the total number of RSM parameters for an n-length test with m categories is therefore $n+(m-1)$. Except for very short tests, this will always be fewer than the $n(m-1)$ PCM parameters that must be estimated for a test.

To obtain specific category boundary locations for an item it is necessary to combine an item's unique location parameter with the test-wide common threshold parameters (see Figure 3.1). The result of this process for our example item is shown in Table 3.3. Note that while the resulting boundary locations are similar to the PCM locations, all of the RSM boundary locations do differ from the PCM locations for the same item. In particular, the range of the trait scale covered by the boundaries (from b_{i_1} to b_{i_4}) is smaller for the RSM than it is for the PCM.

This outcome is reflected in the RSM ICRFs, shown as dashed lines in Figure 3.2. The PCM and RSM ICRFs are shown in the same figure to facilitate comparisons. For example, the smaller trait continuum range covered by the RSM boundaries is clearly evident in a comparison of RSM and PCM ICRFs for Categories 1 and 5. This narrowing of the trait range for this item also has consequences for Category 2 and Category 4 ICRFs, which for the RSM have lower maxima than the PCM ICRFs for these two categories. The ICRFs for Category 3 are however, very similar across the two models.

The effect of the different ICRFs for the two models is shown in Table 3.4. It shows that at $\theta = -1.5$, Category 1 is the most probable response as modeled by the RSM rather than Category 2 at this trait level for the PCM. At $\theta = 1.5$ Category 4 is the most probable response as modeled by both the RSM and the PCM. However, the probability of responding in this category (for respondents at this trait level) is lower for the RSM than was modeled by the PCM.

GPCM

Five parameters unique to each item must be estimated for the GPCM. The single item discrimination and four category boundary location parameters unique to our example item are shown in Table 3.3. Note that the estimated GPCM item discrimination for this item ($a_i = 0.395$) is considerably lower than the implied, but not estimated, PCM and RSM item discrimination, which is $a_i = 1.0$.

Previously, with the RSM, the application of a test-wide set of category thresholds to our example item resulted in a contraction of the range of the trait continuum measured by this item, compared to the range measured using the PCM modeled parameters. In contrast, estimating a discrimination parameter for the example item has contributed to a more than doubling of the trait continuum range between the first and last category boundaries for the GPCM, compared to the same boundary parameters for the PCM. This is demonstrated in the ICRFs shown in Figure 3.3 where the GPCM functions are noticeably flatter, and the range of most probable responding is wider, for each category when compared to the PCM ICRFs for this item.

INSERT FIGURE 3.3 ABOUT HERE

It is interesting to note that the flattening of the category response functions has not affected which are the most probable response categories for this item at $\theta = -1.5$ and $\theta = 1.5$ respectively when modeled by the GPCM compared to the PCM (see Table 3.4). Response probabilities are however lower for the GPCM modeled most probable categories at both of the θ levels shown relative to the PCM response probabilities. The flattening of the GPCM ICRFs however, also means that the probability of responding in Category 3 is higher at both $\theta = -1.5$ and $\theta = 1.5$ as modeled by the GPCM compared to the PCM. This is made quite clear in Figure 3.3.