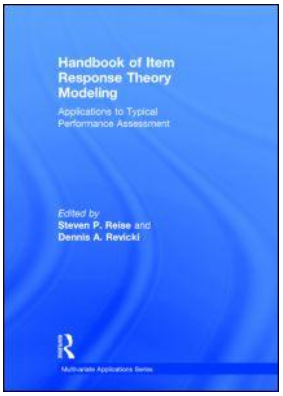

Handbook of Item Response Theory Modeling Applications to Typical Performance Assessment

Steven P. Reise, Dennis A. Revicki

Selecting Among Polytomous IRT Models

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch14
Remo Ostini, Matthew Finkelman, Michael Nering
Published online on: 16 Dec 2014

14 Selecting Among Polytomous IRT Models

Remo Ostini, Matthew Finkelman, and Michael Nering

Introduction

Polytomous item response theory (IRT) has proven fertile ground for psychological measurement model development. While unidimensional dichotomous IRT models primarily come in three basic choices, the measurement practitioner has many more potential models from which to choose when working with polytomous response data, even within the subset of unidimensional models. Ostini and Nering (2006) provide a consideration of some of the ways polytomous IRT models differ as a class from dichotomous models. Nering and Ostini (2010) provide a deeper investigation of the historical and conceptual origins of the most influential models in a collection of chapters written by the researchers most closely associated with the development of the models.

The larger range of models from which to choose results in part from attempts to address the larger number of item formats that polytomous items can take. It also reflects an increased number of possibilities for representing the relationships among category responses within an item—when there are more than two categories into which a respondent can be categorized. However, having a larger number of models from which to choose potentially complicates the question of how to select a model for a particular application, particularly if the concern is with choosing the "correct" model.

This chapter will outline some of the more commonly applied polytomous IRT models, including some of their more salient differences. This chapter will then consider a strategy for choosing among polytomous IRT models, presenting some research that has a bearing on how the strategy may play out in practice.

What Are the Choices?

Influential polytomous IRT models include a model for response data from items with categories that do not have a pre-specified order—the Nominal Response model (NRM; Bock, 1972, 1997) (see also Chapter 18). Multiple-choice items, where there is no explicit order of "correctness" to the distractor items, are often considered potential candidates for the application of the NRM. This provides an opportunity to make use of the distractor items, rather than treating the multiple-choice item as dichotomously scored items (where all distractor responses are marked as incorrect). The model is not, however, limited to multiple-choice items, as Chapter 18 shows. The mathematical form of the NRM is shown in Equation (14.1).

$$P_{jk}\left(u=k\middle|\theta;\mathbf{a},\mathbf{c}\right)=P_{jk}\left(\theta\right)=\frac{\exp\left(a_k\theta+c_k\right)}{\sum_{l=1}^{m}\exp\left(a_i\theta+c_i\right)}, \tag{14.1}$$

where $P_{jk}(\theta)$ is the probability that a response u to item j is in category k *(l = 1, 2, . . . , k, . . ., m)*, as a function of the ability or trait continuum θ, with a category slope parameter a_k and category intercept parameter c_k and with $(a_k\theta + c_k) \equiv Z_k$. The form of Equation (14.1), where the denominator is the sum of all possible numerators, is described as a divide-by-total model by Thissen and Steinberg (1986). This model is unusual in that it involves directly estimating a category response function ($P_{jk}(\theta)$) for each item response option.

One of the earliest polytomous IRT models for items with ordered response categories is the Logistic Graded Response model (L-GRM; Samejima, 1969, 1997b). This model is from a family of graded models developed by Samejima for application to polytomous items where there is a clear order to the amount of a trait required to answer each item category. Likert-type items are a clear example of such an item, but the model can be equally applied to constructed response items, which are scored into better and worse responses to a test question or item. The mathematical form of the L-GRM is shown in Equation (14.2).

$$P_{jk}(\theta) = \frac{\exp\left[a_j\left(\theta - b_{jk}\right)\right]}{1 + \exp\left[a_j\left(\theta - b_{jk}\right)\right]} - \frac{\exp\left[a_j\left(\theta - b_{j(k+1)}\right)\right]}{1 + \exp\left[a_j\left(\theta - b_{j(k+1)}\right)\right]}, \tag{14.2}$$

which can be summarized as $P_{jk} = P^*_{jk} - P^*_{j(k+1)}$, where* $P_{jk}(\theta)$ is the probability of responding in category k *(k = 0, 1, . . . , m)* of item j, P^*_{jk} represents the category boundary (threshold) function for category k of item j, a_j is the item discrimination parameter and b_{jk} is the difficulty (location) parameter for category boundary (threshold) parameter k of item j. The form of Equation (14.2), which involves subtracting an element ($*_{j(k+1)}$) from a preceding element ($*_{jk}$), is described as a difference model by Thissen and Steinberg (1986). The difference model approach to modeling response probabilities is required when polytomous category boundaries—the P* elements in the summary equation—are modeled as cumulative probabilities. This process is described in Ostini and Nering (2006) where this form of polytomous model is characterized as a cumulative boundary (CUM) model.

The Rating Scale model (RSM; Andrich, 1978a, 1978b) is another early IRT model for ordered polytomous items and like the L-GRM (Samejima, 2010) represents a specific instance of a broader approach to modeling polytomous items (Andrich, 2010). The RSM was specifically developed for measurement situations using rating scales (such as Likert scales), which were assumed or expected to operate in the same way across all items in a test, survey, or questionnaire. The model is shown in Equation (14.3).

$$P_{jk}(\theta) = \frac{\exp\sum_{v=1}^{k}\left(\theta - \left(\delta_j + \tau_v\right)\right)}{\sum_{c=1}^{m_j}\exp\sum_{v=1}^{c}\left(\theta - \left(\delta_j + \tau_v\right)\right)}, \tag{14.3}$$

where $P_{jk}(\theta)$ is the probability of responding in category k *(k = 1, 2, . . . , m)* of item j, δ_i is the item difficulty (location) parameter, and τ_k is the common category boundary (threshold) parameter for all the items using a particular rating scale. The τ_k define how far from any given item location a particular threshold for the scale is located. The form of Equation (14.3) is another example of a divide-by-total model in Thissen and Stenberg's (1986)

terminology. In the case of ordered polytomous IRT models, the divide-by-total model is constructed on the basis of category boundaries that are defined by the probability of responding in either response category adjacent to (either side of) the category boundary. Models of this type are therefore referred to as adjacent category models (ACM) by Ostini and Nering (2006) in contrast to the approach taken in CUM models such as the L-GRM. Other ACM models include the Partial Credit [14.4] and [14.5]).Generalized Partial Credit model described later in this chapter (Equations (14.4) and (14.5)).

As can be seen in Equation (14.3), the RSM approach to modeling polytomous item responses involves a distinct method for representing model item parameters with reference to a central item location. Figure 14.1 provides a graphical representation of this approach for two items b_1 and b_2, at different locations on the trait scale. The three tau at the end of the arrows for each item represent the common distance from the item location to the boundary locations for two four-category rating scale items.

This approach is quite flexible in its capacity to model item complexity while keeping the number of estimated parameters low. Examples of this approach incorporating a dispersion parameter (Dispersion Location model, DLM; Andrich, 1982), a skew parameter (Dispersion, Skew Location model, DSLM; Andrich, 1985), and a combination of rating scale and dispersion parameters that reflects Thurstone's (Edwards & Thurstone, 1952) method of successive intervals (Successive Intervals model, SIM; Rost, 1988) are described and compared in Ostini (2002). Figure 14.2 shows how these different models can be represented using the general approach shown for the RSM in Figure 14.1. Vertical lines in the figure represent category boundary locations for two different items (Item *i* and Item *j*) on equivalent θ scales. In Figure 14.1, these locations are represented by the end points of the arrows for each item category. In the mathematical representation of these models, the boundary locations are at the inflection point of the category boundary functions.

The Partial Credit model (PCM; Masters, 1982) provides a more general parameterization for ordered polytomous items than the RSM. In the PCM each category boundary is modeled separately, allowing for items within a scale to vary in the number of categories that they contain. Equation (14.4) shows the PCM.

$$P_{jk}(\theta) = \frac{\exp\sum_{v=1}^{k}\left(\theta - \delta_{jv}\right)}{\sum_{c=1}^{m_j}\exp\sum_{v=1}^{c}\left(\theta - \delta_{jv}\right)}, \tag{14.4}$$

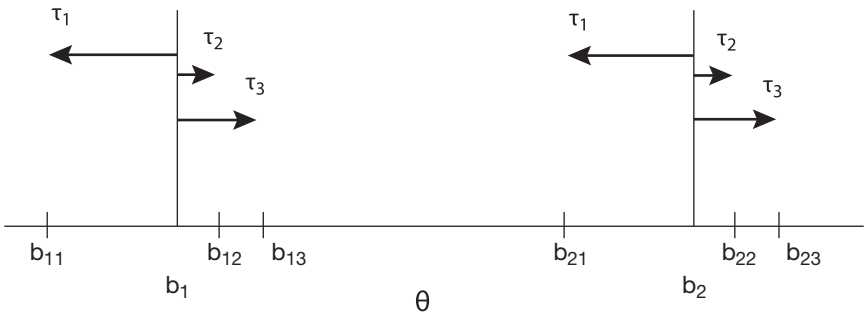

Figure 14.1 Graphical representation of the RSM approach to modeling category boundaries.

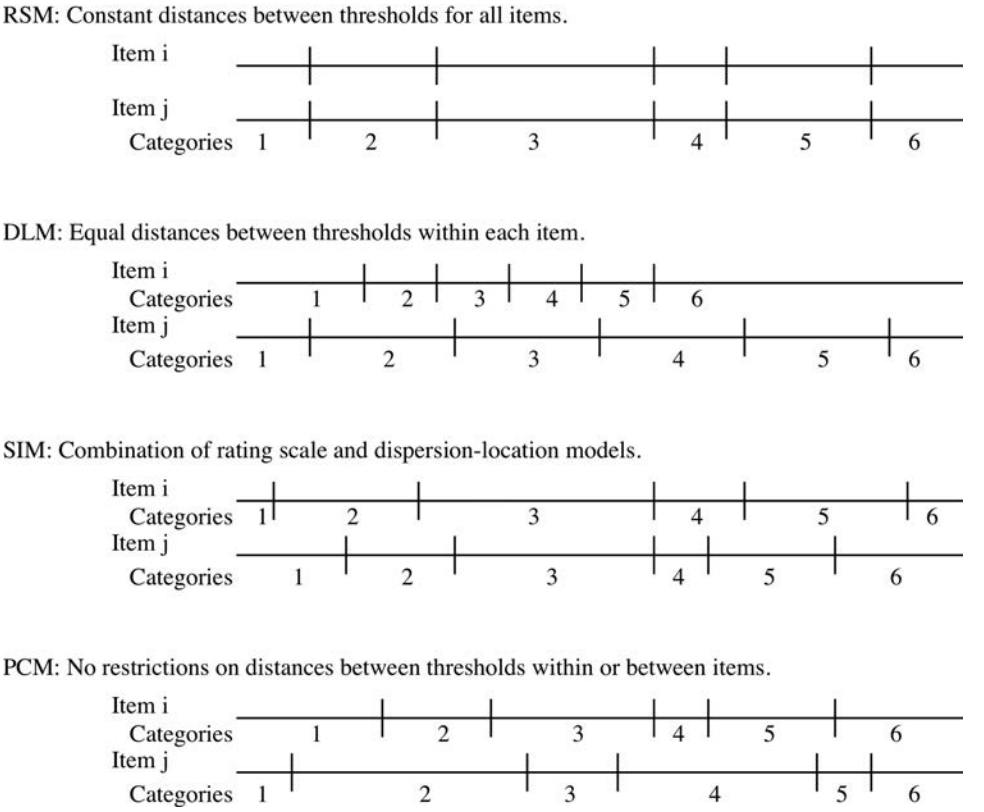

Figure 14.2 Graphical representation of different ways to simplify models by constraining the relationship between category boundaries using the RSM modeling approach.

where $P_{jk}(\theta)$ is the probability of responding in category *k (k = 1, 2, . . . , m)* of item *j*, δ_{jv} is the difficulty (location) parameter for category boundary (threshold) parameter v of item *j*. The RSM and the PCM are both Rasch-type polytomous IRT models.

The Generalized Partial Credit model (GPCM; Muraki, 1992) extends the PCM by modeling a separate discrimination parameter for each item. In this respect, the GPCM is equivalent to the L-GRM although it becomes so by extending the PCM, and is therefore an adjacent category model whereas the L-GRM is a cumulative model. The GPCM also uses the central item location parameterization approach used in the RSM and other Andrich models, as is shown in Equation (14.5).

$$P_{jk}(\theta) = \frac{\exp\sum_{v=1}^{k} 1.7a_j(\theta - b_j + d_v)}{\sum_{c=1}^{m_j}\exp\sum_{v=1}^{c} 1.7a_j(\theta - b_j + d_v)}, \tag{14.5}$$

where $P_{jk}(\theta)$ is the probability of responding in category *k (k = 1, 2, . . . , m)* of item *j*, a_j is the item discrimination parameter, b_j is the item difficulty (location) parameter, and d_v is the category boundary (threshold) parameter for an item. The d_v define how far from

an item location a threshold is located. Adding a discrimination parameter results in the GPCM no longer adhering to the tenets of Rasch measurement—for example, specific objectivity.

Other more recently developed, distinctive polytomous IRT models include the Sequential model (Tutz, 1990, 1997) and the Acceleration model (Samejima, 1995). The Sequential model is a hybrid model incorporating elements of adjacent category and cumulative boundary modeling to represent boundaries as discrete sequential steps in responding to an item (Ostini & Nering, 2006). The Acceleration model is a rare example of a model in Samejima's heterogeneous class of graded models that was developed to provide a way to model complex cognitive tasks in detail (Ostini & Nering, 2006). It is an exponential model that treats test items as cognitive tasks made up of a number of problem-solving steps.

Distinctions Among Models

Differences among polytomous IRT models might be considered important when deciding which model to use in a specific measurement context. As with dichotomous IRT, polytomous IRT models can be distinguished in terms of whether or not a guessing parameter is employed, and whether or not item discrimination is modeled as a separate parameter. Polytomous IRT models also differ with respect to whether or not they were developed within the Rasch measurement approach. The distinction between polytomous Rasch and non-Rasch models is not as clear or philosophically rigorous as it is in the dichotomous case and largely revolves around the question of whether or not sufficient statistics are available for model parameter estimation.

A major distinction that applies only to polytomous IRT models pertains to the way that category boundaries within an item are modeled. Boundary locations can either be modeled across an item, in terms of cumulative category responses (GRM-type models), or locally, with respect to adjacent category responses only (Rasch-type models). They can also be modeled in a combination of adjacent and cumulative boundaries as in sequential models. The statistical basis of this distinction is described in Agresti's (1997) work on the classification of response process and in Mellenberg's (1995) conceptual treatment of polytomous IRT models.

Table 14.1 shows how the category boundary function locations differ between a cumulative category model (L-GRM) and an adjacent category model (GPCM) for an item from a scale measuring moral conceptualization (Ostini, 2010). While these parameters were estimated using the same software program and a common set of data, the type of boundary categorization used by each model is still only one of the factors affecting these parameter estimates. Boundary location values are also influenced by differing discrimination parameter estimates and differences in parameter estimation routines across IRT

Table 14.1 Item Category Parameter Comparison for an Item Modeled by a Cumulative Model (L-GRM) and an Equivalent Adjacent Category Model (GPCM)

Model	*Category Boundary*			
	1	*2*	*3*	*4*
L-GRM	–2.502	–0.766	0.848	2.419
GPCM	–2.415	–0.688	0.992	2.111

models—even within a single software program. The differences in boundary location estimates shown in the table are therefore only indicative of the effects of the two ways of modeling category boundaries.

Polytomous models can also be constrained in pre-specified ways. The most common example of this is where polytomous items elicit responses using a rating scale. Rating scale versions of both cumulative (RS-GRM; Muraki, 1990) and adjacent category (RSM) models exist. Conversely, polytomous models can be generalized to become able to operate with data where the ordering of response categories is unknown or not specified (NRM) or where the data to be modeled are essentially continuous (Samejima, 1973).

A factor that might be considered important in selecting among the variety of polytomous IRT models available for use is the question of how they differ in terms of measurement outcomes. In a comparison of eight different polytomous IRT models with model parameters estimated using seven different software programs (for 26 different model *x* software conditions) applied to two common data sets, Ostini (2002) investigated measurement outcomes in terms of variation in item and respondent trait level parameter estimates. The results of that research inform some of our consideration of model selection issues later in this chapter.

A Selection Strategy

In the remainder of this chapter we will outline a strategy for making an informed decision about which polytomous IRT model to use in a specific context. Cella and colleagues (2007) provide a good outline addressing the more fundamental question of why IRT would be used in preference to classical test theory. These ideas are explained in more detail in the work of Lord (1980; Lord & Novick, 1968) and Hambleton and Swaminathan (1985).

Background Considerations

A number of preliminary issues must be considered before IRT model selection takes place. These include the assumptions that underpin IRT, the question of measurement philosophy, and the possible processes that give rise to polytomous item responses.

IRT Assumptions

IRT is known as a strong measurement theory because it makes strong assumptions about the condition of the data to which the model is being fit. To begin with, item data must be of known dimensionality. In practice, especially for polytomous models, this usually comes down to data being unidimensional—that is, only one latent trait is needed to account for test item responses. A second assumption is that test items produce data that are locally independent. This assumption is clearly described in a number of places (Hambleton & Swaminathan, 1985; Lord & Novick, 1968; Weiss & Yoes, 1991). For a test measuring just one trait, the assumption of unidimensionality is equivalent to the assumption of local independence, although this is not the case for multidimensional tests (Crocker & Algina, 1986)—making model assumption testing much more difficult for multidimensional tests. A third key assumption is that the latent variable is monotonically related to item response probability. That is, people with greater levels of a trait have higher probabilities of responding in the item category that indicates the presence (or greater presence for polytomous items) of the trait (Lord, 1980). In the simplest case, for a test of ability

consisting of dichotomous items, a response indicating "greater presence" of the trait typically refers to a correct response.

A further issue that arises for polytomous IRT models is the question of whether each response category for an item is, for some range of the trait continuum, the most probable response category. The contention is that an item is not working as intended or expected if it has modeled response categories that are never the most probable response category. In essence, the argument is that an item category that is never the most probable category is not required. Having such a category suggests that an item has too many categories for use in making meaningful measurement distinctions. This is seen as an important diagnostic tool (Andrich, 1995) with implications for combining data across response categories (joining assumption) and for category boundary parameterization. For polytomous models, such as Rasch models and the GPCM, which use ACM boundaries, the presence of unordered boundary locations (boundary reversals) is a clear indicator of an item category that is never most probable. For models built on CUM boundaries, category response functions need to be explicitly plotted for such categories to be identified. In the CUM case, a category response function that never rises in probability above the height of all other category functions at some point on the trait scale is a category that is never the most probable category.

Testing of IRT and model assumptions is sometimes seen as something to be done prior to model fitting; however, assumption testing usually requires fitting IRT models as part of the testing process. This was the approach explicitly used in the development of PROMIS® item banks, where, for example, items selected for the Emotional Distress domain were calibrated using the GRM and the response categories for each item were inspected to determine whether they were "most probably" categories at any point on the trait continuum (Cella et al., 2007). Our research suggests that substantive differences among polytomous IRT models are more pronounced to the extent that data do not meet IRT model assumptions (Ostini, 2002). That is, where data meet IRT model assumptions, substantive differences in the measurement outcomes of different polytomous IRT models tend to be small. Examples of how this can occur are described over the following sections of this chapter.

Measurement Philosophy

While the practical difference between Rasch and non-Rasch models is less distinct for polytomous IRT, the difference in underlying measurement philosophy remains. If measurement practitioners are convinced that specific objectivity is an important feature of their measurement enterprise, they will need to use a Rasch model. Specific objectivity can be thought of as the capacity to provide invariant comparisons among people and among items. At the model parameter level, this requires that comparisons among item parameter values be independent of person parameter values and vice versa. In practice this means expecting, or being willing to treat items as if they are equally discriminating across an item set. IRT models that include a separately modeled discrimination parameter are not therefore Rasch models.

An important element of measurement philosophy is the question of whether tests should be designed so that responses fit measurement theory or whether the goal is to find a model that reflects how people actually respond to tests. The argument from a Rasch perspective is that meaningful measurement only occurs if specific objectivity holds, which requires data that fit a Rasch model. In practice, items typically vary in their discriminating power but this cannot be modeled in a separate discrimination parameter (with the

potential decrease in model fit) for specific objectivity to hold. Considering the response processes that have been proposed for different polytomous IRT models helps clarify the distinction between selecting a model based on data and selecting data to fit a model.

Response Processes

The logistic ogive is the mathematical function at the heart of IRT. It represents the probability of a person responding to an item with a specific response in terms of the standing of that person on the trait that the item is measuring. Figure 14.3 shows the shape of a logistic ogive modeling a hypothetical response to a health outcome item. In Figure 14.3, Person H has more of the trait in question (the health outcome) and therefore has a higher probability of endorsing this item.

This function was adopted for this purpose early in the development of IRT, in part because it was found to closely resemble the item response behavior that people exhibit (see, e.g., Lord, 1980). There are a number of further assumptions implicit in the different types of polytomous IRT models about the way responses are generated when a respondent is faced with a test item. Different polytomous models make different assumptions about the psychological processes that might lead to a particular person responding in a

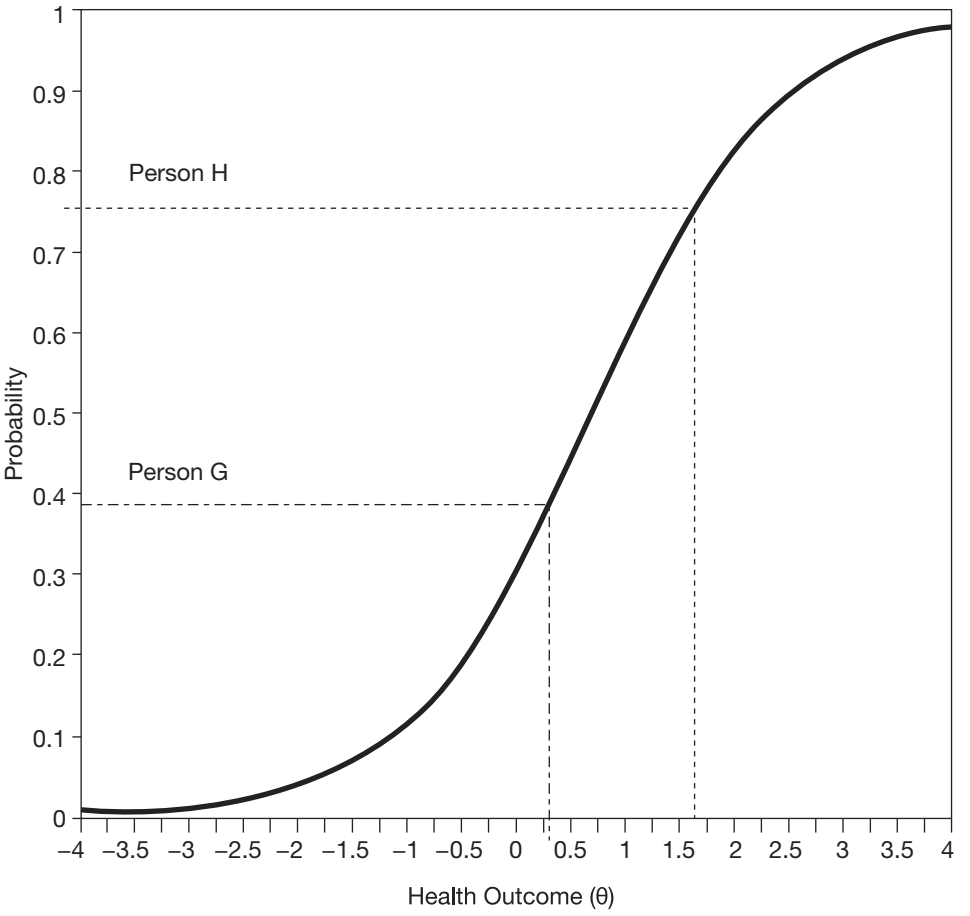

Figure 14.3 An example of a logistic ogive modeling a hypothetical health outcome item response.

particular category of a polytomous item. Curiously, and unfortunately, research into the response processes that people actually employ when faced with different types of testing situations is very rare. There is therefore little empirical data to suggest which models might be more appropriate for a given testing situation.

Cognitive interviewing can provide a method for assessing the response processes that people use when responding to polytomous items (see, e.g., Castillo-Díaz & Padilla, 2013). It could conceivably be used to test the hypothesized response processes for polytomous IRT models. For example, Samejima (1972) proposed a response process for cumulative models in her Graded Response framework whereby a respondent who has been attracted by a response category ($g - 1$) to item i is subsequently further attracted by the next category (g). This hypothesized response process is the building block of her cumulative polytomous IRT model—and could conceivably be tested, through a process such as cognitive interviewing.

RATING SCALE

The fact that the RSM models item category boundaries in terms of adjacent categories implies a response process that treats each category boundary as a separate dichotomous item (Masters, 1982; Masters & Wright, 1984; Wright & Masters, 1982). This suggests that a person giving a response in the third category of an ordered polytomous item has effectively responded positively to two successive dichotomous items (the boundary between the first and second item categories, and then again at the boundary between the second and third categories).

Rather than focusing on how people respond to polytomous items, in more recent work, Andrich (2010) begins with the theoretical measurement requirement of invariant comparisons among people and among items (i.e., specific objectivity) and shows how this is accomplished in a general polytomous Rasch model. Applying different parameter restrictions then results in specific models, such as the PCM and the RSM. The parameterization of the RSM and the PCM specifically were described earlier.

Andrich (2010) then focuses on two properties of the RSM and polytomous Rasch models that can seem counterintuitive. The first is that, with polytomous Rasch models, it is not possible to combine responses from two adjacent categories into a single response category without changing how an item is measuring the trait in question or changing the model-data fit for the item—that is, these models violate the "joining assumption" (Jansen & Roskam, 1986). The second property that he addresses is how the trait level ordering of Rasch model category boundary locations need not be the same as the intended order of the item categories (i.e., boundary reversals). Rather than treating these properties as problems with polytomous Rasch models, Andrich (2010) argues that a strength of polytomous Rasch models is that they provide a way to test whether ordered, categorical items are operating in the manner intended. For example, finding boundary reversals shows that an ordered categorical item is not functioning as intended.

Our empirical comparison of item parameters and boundary functions across different polytomous IRT models suggested that the rating scale parameterization of the RSM was more plausible than the within-item dispersion parameterization of the DLM but produced greater modeled discrepancies in comparisons with more general polytomous models such as the DSLM, PCM, and SIM (Ostini, 2002). In trait parameter comparisons, RSM trait estimates correlated perfectly with almost all other Rasch models in an IRT assumption-fitting data set. However, specific, estimated individual trait locations differed more for people modeled with the RSM than other Rasch models. This set of results suggests that the rating scale assumption may be overly restrictive—even with data collected using a rating scale format (Ostini, 2002). That is, while trait estimate rank ordering was

very similar for RSM and more general models, trait estimates across test respondents showed larger differences for RSM modeled estimates compared to more general models than among the different general models themselves. This suggests that keeping category "widths" constant in a test using an ostensibly common rating scale (such as a Likert scale) may not be empirically supported.

PARTIAL CREDIT MODEL (AND GENERALIZED PCM)

The idea that polytomous Rasch models (and adjacent category models generally) model successive dichotomous choices at each category boundary was initially proposed in the context of the development of the PCM (Masters, 1982; Masters & Wright, 1984; Wright & Masters, 1982). In early descriptions of this response process, each category boundary decision was described as an independent step. However, Tutz (1990), Molenaar (1983), and Verhelst and Verstralen (1997) showed clearly that adjacent category boundaries for an item are not modeled independently of each other—and Masters (1988) ultimately conceded that this was the case.

In his more recent exposition of the development of the PCM, Masters (2010) focuses on the importance of specific objectivity—describing it in terms of the independent estimation of different people's abilities that allows "objective" comparisons. As with Andrich (2010) as described in the previous section, the argument essentially becomes a justification of the adjacent category response process (whatever it may represent) because it allows specific objectivity.

This essentially leaves the GPCM without a clear underlying response process because the estimation of a discrimination parameter renders the model unable to meet the requirements of specific objectivity; and the critique of the step notion of a response process applies equally to the GPCM as to polytomous Rasch models.

GRADED RESPONSE MODEL FRAMEWORK

Samejima (1972) describes a response process for her class of models (including the L-GRM), which is based on a function that denotes the probability that someone who has been attracted by an item response category will be further attracted by the next category. The probability of being attracted to a specific response category is the serial product of this function for all preceding categories. The probability of responding in a particular response category is then the probability of being attracted to that category minus the probability of being attracted to the next response category. This cumulative response process applies specifically to polytomous items with ordered response categories.

More recently, Samejima (1996, 2010) has proposed a set of five criteria with which to evaluate polytomous IRT models. The first criterion explicitly states that models and their accompanying assumptions must fit with the psychological processes that gave rise to the data. The remaining four criteria are mathematical criteria that define requirements for meeting the joining assumption; that describe the general shape of category response functions; and that define the ordered nature of response categories.

Samejima appears to allow for the possibility that different psychological response processes will operate in different testing contexts. This has led her to introduce a number of new polytomous models within her graded response model framework, which operate quite differently from the L-GRM for which she is best known. These include a logistic positive exponent family of models (Samejima, 2008) and the acceleration model (Samejima, 1995). The details of the psychological processes that these models are intended to capture are described in the relevant publications, but the driving assumption behind

Samejima's work is that IRT models should be selected on the basis of their conformity to the nature of the response data. This insistence has most recently taken Samejima into the realm of nonparametric estimation methods (Samejima, 2010).

Parameter Estimation

Fitting a polytomous IRT model to a set of data effectively comes down to estimating a set of model parameters—for both item characteristics and person trait levels. Modeling multiple categories for each item adds a level of complexity to polytomous IRT modeling beyond that entailed in dichotomous IRT, particularly for models that allow for different numbers of categories for different items in a test. Simultaneously estimating invariant item and person parameters on the same measurement scale is computationally complex and intensive. It also requires data to provide information about each item parameter, in sufficient quantity to estimate the parameters with precision.

Amount of Data

Another related consideration is the amount of data being collected. As a general rule, the more item parameters being estimated, the more data that will be required to accurately estimate those parameters. Another consideration here is that more complex data collection processes can result in more missing data (planned or unplanned). Data need to be collected in a way that ensures that there are enough responses in each response category of every item to allow category boundaries to be estimated with some degree of precision.

In general terms, this requires that enough information is available in the data (item responses) to estimate a parameter with some precision. This typically requires a larger sample for more complex models, such as those that require estimation of a discrimination parameter. However, a number of factors come into play when considering sample size requirements, and no simple, generally applicable sample size guidance is possible. Data sets should be checked to identify response categories that are rarely or never selected prior to attempting to fit an IRT model. Their presence will cause estimation difficulties and indicate that more data is required—or that responses across categories must be collapsed into a combined (post hoc) category. The standard error of parameter estimates can also be checked after model fitting to identify possible estimation problems—one cause of which may be insufficient response data.

For example, the distribution of the trait in a sample, relative to the trait location of an item, is a complicating factor. The problem can be seen in an item that is not difficult to endorse fully (or answer correctly in an ability test), which may fail to elicit responses in response categories that indicate less of a trait (e.g., the "disagree" categories in a rating scale) in a sample of respondents with moderate or high levels of the trait. This can result in difficulty in accurately estimating category boundaries between the categories that are infrequently selected simply because little information is available to inform the parameter estimation.

The PROMIS® Block item collection procedure is a good example of ensuring sufficient data is collected to allow accurate parameter estimation. In the first wave of data collection, responses were obtained from a sample of 21,133 people, 7,005 of whom completed the full item bank, with 14,128 completing overlapping blocks of items (Cella et al., 2007; Hays, Morales, & Reise, 2000).

In previous research we found parameter estimation convergence to be a problem for some estimation programs, when small numbers of items were being modeled—even when there was data from a large number of respondents, with data present in every response

category of every item (Ostini, 2002). This problem can sometimes be avoided by changing the default estimation procedures in a program. An example of this that we experienced with the now superseded Multilog program (Thissen, 1991) required changing person parameter estimation from maximum likelihood to maximum *a posteriori*.

Type of Parameter

Beyond the question of amount of data, however, is the question of the type of parameter being modeled. Modeling a discrimination parameter for each item will require a broad range of responses to each item in a measurement scale. Modeling a "guessing" parameter, as, for example, in the three-parameter logistic model for dichotomous data, will require a substantial amount of data at low trait levels for the parameter to be estimated with any precision.

In a health outcome context, "guessing" types of responding could occur with any self-report measure that provides an exhaustive list of response options from which respondents can choose (such as multiple-choice, True/False, or rating scale formats). This situation allows respondents to select a trait-indicating response (e.g., health status) that is not a true indication of their health status. Such responding can occur in situations where people are disinterested in responding to a test, fatigued, or distracted, where the response becomes essentially random. Responding that is not a true reflection of a person's trait can also arise in measures of typical performance, such as self-reported health outcomes, when respondents deliberately choose a response that suggests that their health is worse than it really is. For example, a respondent may select 1 = Very Poor on a five-point scale (from Very Poor to Very Good) in order to increase their chance of obtaining a desired health care outcome. The converse is also possible with, for example, a question regarding health risk behavior, where a respondent may select a more socially desirable response indicating low-risk behavior when their true behavior is in fact more risky. While the latter two examples differ from the more random forms of responding typically associated with a guessing parameter, they contribute to potential estimation difficulties when they represent responding at low trait levels.

Aside from the "guessing" parameter, the discrimination parameter is the most difficult to estimate. This was certainly found to be the case in our model comparison research where parameter estimation problems only arose for models that required the estimation of a discrimination parameter (Ostini, 2002). Such parameter estimation problems had flow-on effects on respondent trait estimates. That is, when an estimation routine was unable to converge on a parameter estimate, the parameter (discrimination in this case) had to be artificially constrained and this process influenced the person parameter estimates that were subsequently calculated.

This issue was also more pronounced for data that demonstrated worse fit to IRT model assumptions—particularly the unidimensionality assumption. While different parameter estimation methods could overcome this problem, this in turn also differentially affected modeled trait estimate outcomes (Ostini, 2002). Indeed, it was disconcerting to find that a software program's default estimation method setting could have a greater influence on trait distribution characteristics than choice of polytomous model appeared to have (Ostini, 2002).

Another finding from our model comparison research was that discrimination parameter estimates were strongly related across models and estimation software—except for the RS-GRM. This suggests that the rating scale approach of constraining boundary parameters to be equivalent across items substantially influences the estimation of the discrimination parameter for an item (Ostini, 2002). This is perhaps unsurprising when it is recalled that discrimination in the polytomous IRT case is composed of a combination of an

item discrimination component and the widths of item categories. Constraining category boundaries across items will then necessarily require the item discrimination parameter to be entirely responsible for modeled differences in discrimination across items.

In the context of the computationally intensive parameter estimation demands of polytomous IRT models, the availability of sufficient statistics for Rasch model parameter estimation can be an advantage. The availability of sufficient statistics can improve the robustness of the parameter estimation process, irrespective of any measurement philosophy considerations. On the other hand, this robustness may also mask instances of poor model-data fit (Ostini, 2002), indicating an interaction between model type, parameter estimation, and model data fit.

In previous research we found evidence of interaction between polytomous IRT models and estimation software due to the different estimation routines used; for example, maximum likelihood versus Bayesian methods (Ostini, 2002). A general finding from our previous model comparison research was that parameter estimation effects tended to be more pronounced for the data set that more poorly met IRT assumptions (Ostini, 2002).

Model Fit

Model fit is an important tool in the test construction process but also has a potential role in model selection (see Chapter 6). Ideally, model fit testing will provide an indication of whether the response process underlying the model being fitted to a set of response data accurately represents the way the response data was produced or generated. In practice, it is not clear that either the response specification for most polytomous IRT models or the available tests of model fit, are adequate for this task.

An outline of some of the different mathematical approaches to assessing model-data fit are listed in Ostini and Nering (2006) together with a discussion of the problems associated with the fit statistics available for polytomous IRT model testing. That book also provides a worked example comparing the different results produced by two fit statistics (x^2, and Q_i) when they are applied to a five-category item modeled by four different polytomous IRT models (RSM, PCM, GPCM, and L-GRM).

The discussion in Ostini and Nering (2006) focuses on model-data fit and this is the most obvious level at which to assess fit as part of the model selection process. In IRT, however, fit can also be evaluated at the individual item or the specific person level. While this level of fit assessment is often part of the test construction process, it can also be implicated in model misfit during the model selection process. In part this is because item selection decisions made during test construction have model selection implications. More generally, the processes of test construction and model selection are rarely completely separate.

Broadening the idea of statistical model selection beyond model specific fit statistics raises the notion of a simultaneous consideration of the goodness of fit of a model and the number of parameters required to achieve that fit. This typically involves a penalty term that increases with the number of parameters in the fitted models, allowing the comparison of models with different numbers of parameters. The Akaike information criterion (AIC; Akaike, 1977) and various modifications of the AIC implement an approach to model selection along these lines.

A recent comprehensive evaluation of six model selection methods (Whittaker, Chang, & Dodd, 2012), including the likelihood ratio test and five information criteria methods, found that all of the methods were able to accurately select models from data generated by simple IRT models. However, these methods for assessing overall model-data fit were unable to correctly assess model fit for models involving discrimination or "guessing" parameters.

Glas (2010) provided an innovative framework for evaluating polytomous IRT model fit. The framework allows for the assessment of item and person fit, as well as evaluation of the form of the item response function, local stochastic independence, and subpopulation invariance, the latter of which is also known as differential item functioning (DIF). This potentially allows a more integrated approach to assessing model-data fit and testing fundamental IRT assumptions together. While this framework has not been implemented in IRT modeling software, Glas provides examples of methods that can be used to apply this framework.

Two methods for assessing fit were described by Glas (2010) in the context of two common IRT parameter estimation methods: marginal maximum likelihood (MML) and Bayesian estimation. Modification indices based on Lagrange multiplier statistics were developed for the MML estimation context, and demonstrated for three polytomous IRT models (GPCM, Sequential model, and L-GRM) and for person fit. Glas (2010) also developed the application of posterior predictive checks for the evaluation of model fit in a Bayesian parameter estimation context. It is not yet clear whether this promising approach can become a practical tool in the model selection toolkit or whether it will primarily remain a research tool.

In our earlier model comparison research, identification of misfitting items was not consistent even within the same model as fit by different software programs. Furthermore, more general models did not show better item-model fit than more restrictive models beyond the very restrictive rating scale models (Ostini, 2002). This could be a reflection of model fit; an example of the deficiencies among available model fit statistics; an example of Samejima's contention that estimation procedures unduly affect fit results (Samejima, 1997a); or some combination of these three factors.

Item-model fit analysis was the area that showed the largest potential to substantively affect measurement outcomes in our model comparison research (Ostini, 2002). This was due to the fact that applying the fit results from the different model and parameter estimation software conditions would have led to the retention of different items in the tests being modeled. However, the major substantive differences that would have resulted from this process did not appear to be based on sound fit test criteria (Ostini, 2002).

The process of ascertaining whether a chosen polytomous IRT model is the best model for a set of response data involves a number of interacting factors. The fundamental question is whether the modeled response process corresponds to the process evident in the response data—whether that process is pragmatically conceived (per Samejima) or driven by measurement theory (per Rasch). Any test of this proposition, however, is influenced by the presence of items that are not operating as intended; people who are not responding as expected; software that is imperfectly estimating model parameters; and fit statistics that are unable to provide an unbiased test of the data's deviation from model expected responses. Given this interwoven complexity and the need for substantial amounts of data for parameter estimation purposes, the *p*-value of any given fit test is unlikely to be the best way to select a polytomous IRT model for any given testing situation. For some model comparison procedures, in certain model comparison situations, however (e.g., with nested models), the comparative magnitude of the fit test statistic may be a pragmatic factor in selecting among models.

Reporting and Interpretation of Results

It is generally easier to report and explain results from simpler models. For example, modeling discrimination may result in respondents' latent trait ranking among a testing cohort being different from their summed ("number correct") score ranking. This can be

difficult to explain and justify to those who commissioned a testing program or who took a particular test, for whom "number correct" may make intuitive sense. Because Rasch trait scores are based on number correct scores given their reliance on sufficient statistics for their estimation, this can make them easier for people to understand.

In terms of specific models, Smits, Cuijpers, and van Straten (2011) selected the L-GRM over the GPCM for their modeling project because of the ease with which L-GRM item parameters can be interpreted for items generated by a Likert rating scale and because of the authors' perception that a model with cumulative category boundaries is simpler to explain to test users.

Interpreting the meaning of score categories is assisted by the capacity to equate IRT outcomes across different groups of test takers. Unlike CTT, this means score categories and IRT test results generally can be meaningfully interpreted across testing settings—temporally and across groups of test takers.

Equating is more complicated when a discrimination parameter is involved, and this may be a consideration for model selection. Equating processes are also more readily accomplished within some parameter estimation programs than others—and the polytomous IRT models for which those programs can estimate parameters may then influence model selection decision making.

Available Model Estimation Software

The availability and associated features of model estimation software can be a pragmatic factor that influences model selection decision. Early RSM software availability helped the RSM to become a model of choice (Hambleton, van der Linden, & Wells, 2010). This is not always the case, however, as in the example of WinMira software (von Davier, 1998), which had one of the earlier and more useable graphical user interfaces. The availability of this feature nevertheless did not make the Successive Integers Model or the Q fit statistic that WinMira estimated more widely adopted. Graphical user interfaces are now much more common but sometimes still require knowledge of Fortran data specification code, simply to read data in to an estimation routine.

Different parameter estimation software often provides different fit statistics, making it difficult to obtain consistent conclusions on model fit, even for a single specific polytomous IRT model when fit by different software. Furthermore, program documentation rarely gives clear guidance on how to use the fit indicators provided (Ostini, 2002). This contributes to the effect described earlier where different programs, using different methods to test fit, identify different items as misfitting—even when fitting the same polytomous model. The process of identifying and removing "misfitting" items in this way has more influence on modeled outcome than model type.

Other diagnostic features, particularly graphic displays of results such as plots of parameter estimates, are very helpful in understanding how a model is operating with a particular set of data. Most programs provide these in some form; for example, even a generic statistical program like R (R Development Core Team, 2011) has libraries such as MIRT (Chalmers, Pritikin, & Zoltak, 2014) and SIRT (Robitzsch, 2014) that provide histograms and graphs of item and person parameter results. The reliance of IRT modeling on specific forms of response functions makes graphical representations of modeled outcomes particularly useful for seeing what is actually happening in a modeling situation.

Measurement Differences

To the extent that substantive measurement differences are the result of selecting test items based on item-data fit, the fact that different model-estimation program combinations

provide highly inconsistent guidance on items to retain (Ostini, 2002) is a matter of significant concern.

In terms of the distinction between CUM and ACM models, our model comparison research suggests that differences in modeled probabilities of test-taker responses in a given item category between CUM and ACM models are smaller than the difference found among different ACM models. For example, L-GRM and GPCM results were more similar than PCM and GPCM modeled probabilities, even though the latter two are both ACM models. The presence of an estimated discrimination parameter has more effect on modeled outcomes than does the CUM—ACM distinction (Ostini, 2002). This becomes an important consideration in model choice for polytomous models, both because discrimination parameter estimation can be more difficult to achieve without imposing external constraints, and because polytomous item discrimination is modeled through a combination of a slope parameter and the width between item-category boundaries (Muraki, 1992; Muraki & Bock, 1999).

In the data set that better met IRT assumptions, the most general model (GPCM) was the only model that produced trait distribution estimates that differed markedly from those produced by other models—largely as a result of trait estimate convergence problems resulting from a small number of cases with extreme parameter estimates (Ostini, 2002). Those estimates were themselves evidently the product of discrimination parameter estimation difficulties. Aside from these GPCM results, substantive measurement differences across the 25 remaining model conditions were not reflected in large differences in trait estimate distributions (Ostini, 2002).

More generally, when a data set meets fundamental IRT assumptions, and to the extent that substantive measurement differences are thought of in terms of respondent trait estimates—especially the rank ordering of those estimates—different polytomous IRT models produce very similar results (Ostini, 2002). Therefore, while concerns remain about the utility of available item-model fit procedures, once a set of items has been selected, the many different polytomous IRT models produce remarkably similar respondent measurement. This indicates that the most important factor when using polytomous IRT models is to invest care and effort in the initial item analysis and IRT assumption testing stages. Selection of the particular polytomous IRT model that is applied to a given set of data appears to be a far less important issue (Ostini, 2002).

Results from our model comparison work (Ostini, 2002) suggest that models can be too restrictive (e.g., RSM, DLM, RS-GRM). However, the most general models investigated (L-GRM & GPCM) did not produce substantially different results from less general models such as the PCM, SIM, and DLSM. This result reaffirms a perception among polytomous IRT practitioners. For example, when considering whether to use the PCM or the L-GRM in a measurement project, Smits and colleagues (2011) noted that the choice between the two models made little difference in terms of the resulting person parameters.

Strategy Simplified

In summary, selecting from among the polytomous IRT models available for a given measurement task can be thought of as requiring an answer to three basic questions, following which, two important tests of data condition are required. These steps are outlined next.

(1) *Why are you measuring?*

- is there an implied response process?
- have you ascribed to a particular philosophy of measurement?

- how are you going to use the results (decision making; explain to respondents or test commissioning body)?

(2) *What data are (will be) available?*

- less data may require fitting a simpler model (which is likely to fit the data less well)
- fewer parameters to estimate; or—if responses are relatively evenly distributed across item response categories for all items—perhaps choose a Rasch model because then you have sufficient statistics on which to base parameter estimation.

(3) *What software is available?*

- Understanding the parameter estimation methods used by specific software and the way outlying estimates are constrained may be important for parameters that are difficult to estimate

Test IRT assumptions—including assumption of uniform discrimination if you do not want to estimate a discrimination parameter.

Test model fit—both item-model fit and person-model fit, as well as overall model-data fit. This is primarily to get to know your particular set of data better, rather than for the purpose of item or model selection.

Summary

Our research suggests that the most important factor when thinking about applying a polytomous IRT model is to test model assumptions carefully and thoroughly, with special consideration of the unidimensionality assumption. Deviation from the unidimensionality assumption has undue, differential effects on parameter estimation across different polytomous models and different software. One consequence is that, even though polytomous Rasch model estimation procedures are more robust in the face of violation of the unidimensionality condition, the resulting item and person parameters are nevertheless affected by the assumption's violation.

It is also important to consider whether the availability of sufficient statistics in the parameter estimation procedure would be advantageous—if, for example, there is a limited amount of data available. This would suggest using a Rasch model. However, this would also mean that respondent trait order would be the same irrespective of which Rasch model is used. And, as mentioned in the previous paragraph, special care will need to be taken to ensure that IRT assumptions are met before modeling takes place.

Awareness of the default settings of the parameter estimation software being used to model data is also important, as they can have undue effects on the values of the parameters that are estimated. This can occur when estimation routines fail to constrain parameter estimates; or conversely, when estimation routines unduly restrict parameter estimates, thereby artificially restricting parameter variability across items or respondents.

Finally, and to repeat the first point, it is crucial that IRT model assumptions are carefully tested and item selection results should be based on assumption testing outcomes, rather than IRT item-model fit statistics, as fit results may not be reliable and have a disproportionate effect on modeled outcomes based on the inconsistent guidance that they give across different models.

References

Agresti, A. (1997). Connections between loglinear models and generalized Rasch models for ordinal responses. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 209–218). Münster: Waxmann.

Akaike, H. (1977). On entropy maximization principle. In P.R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27–41). Amsterdam: North Holland.

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581–594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispertion parameters. *Psychometrika, 47*(1), 105–113.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N.B. Tuma (Ed.), *Social methodology* (1st ed., pp. 33–80). San Francisco, CA: Jossey-Bass.

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

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* (pp. 123–152). New York: Taylor & Francis.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51.

Bock, R.D. (1997). The nominal categories model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33–49). New York: Springer.

Castillo-Díaz, M., & Padilla, J.-L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research, 114*(3), 963–975.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–S11.

Chalmers, P., Pritikin, J., & Zoltak, M. (2014). Multidimentsional item response theory, 1.3. York: Authors.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.

Edwards, A.L., & Thurstone, L.L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17*(2), 169–180.

Glas, C.A.W. (2010). Testing fit to IRT models for polytomously scored items. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* (pp. 185–208). New York: Taylor & Francis.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Nijhoff.

Hambleton, R.K., van der Linden, W.J., & Wells, C.S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* (pp. 21–42). New York: Taylor & Francis.

Hays, R.D., Morales, L.S., & Reise, S.P. (2000). Item response theory and health outcomes measurement in the 21st century. *Med Care, 38*(9 Suppl), II28–42.

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

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Masters, G.N. (1988). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Latent traits and latent class models* (pp. 11–29). New York: Plenum Press.

Masters, G.N. (2010). The partial credit model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* (pp. 109–122). New York: Taylor & Francis.

Masters, G.N., & Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*(4), 529–544.

Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*(1), 91–100.

Molenaar, I.W. (1983). Item Steps: Heymans Bulletins Psychological Institute, University of Groningen.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*(1), 59–71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Muraki, E., & Bock, R.D. (1999). PARSCALE: IRT item analysis and test scoring for rating-scale data. (Version 3.5). Chicago: Scientific Software International.

Nering, M., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory*. New York: Taylor & Francis.

Ostini, R. (2002). *Identifying substantive measurement differences among a variety of polytomous IRT models*. PhD dissertation, University of Minnesota, Minneapolis. PsycINFO database.

Ostini, R. (2010). Measuring conceptualisations of morality: Or how to invent a construct and measure it too. In G. Walford, M. Viswanathan, & E. Tucker (Eds.), *The SAGE handbook of measurement* (pp. 337–352). Los Angeles, CA: Sage.

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

R Development Core Team. (2011). R: A language and environment for statistical computing. (Version 2.13.1). Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org/.

Robitzsch, A. (2014). Supplementary item response theory models, 0.45–23. Salzburg: Author.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12*, 397–409.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement No. 17*.

Samejima, F. (1972). A general model for free response data. *Psychometrika, Monograph Supplement No 18*.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*(2), 203–219.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60*(4), 549–572.

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika, 23*(1), 17–35.

Samejima, F. (1997a). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika, 62*(4), 471–493.

Samejima, F. (1997b). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Samejima, F. (2008). Graded response model based on the logistic positive exponent family of models for dichotomous responses. *Psychometrika, 51*, 567–577.

Samejima, F. (2010). The general graded response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory* (pp. 77–107). New York: Taylor & Francis.

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: a simulation study. *Psychiatry Research, 188*(1), 147–155.

Thissen, D. (1991). Multilog, 6.0. Chicago: Scientific Software International.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*, 39–55.
Tutz, G. (1997). Sequential models for ordered responses. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York: Springer.
Verhelst, N.D., & Verstralen, H.H.F.M. (1997). Modeling sums of binary responses by the partial credit model. Cito, Arnhem, The Netherlands.
von Davier, M. (1998). WINMIRA 32—Online users manual. (Version Windows 95). Kiel: Institute for Science Education (IPN).
Weiss, D.J., & Yoes, M.E. (1991). Item response theory. In R.K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 69–95). Boston, MA: Kluwer.
Whittaker, T.A., Chang, W., & Dodd, B.G. (2012). The performance of IRT model selection methods with mixed-format tests. *Applied Psychological Measurement, 36*(3), 159–180.
Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL: MESA.