Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2014, Vol. 33, No. 1, pp. 36–48

An NCME Instructional Module on Polytomous Item Response Theory Models

Randall David Penfield, *The University of North Carolina at Greensboro*

A polytomous item is one for which the responses are scored according to three or more categories. Given the increasing use of polytomous items in assessment practices, item response theory (IRT) models specialized for polytomous items are becoming increasingly common. The purpose of this ITEMS module is to provide an accessible overview of polytomous IRT models. The module presents commonly encountered polytomous IRT models, describes their properties, and contrasts their defining principles and assumptions. After completing this module, the reader should have a sound understating of what a polytomous IRT model is, the manner in which the equations of the models are generated from the model's underlying step functions, how widely used polytomous IRT models differ with respect to their definitional properties, and how to interpret the parameters of polytomous IRT models.

Keywords: item response theory, polytomous items, partial credit model, graded response model, nominal response model

Item response theory (IRT) has become a widely adopted framework for the development and scaling of educational assessments. The widespread use of IRT stems from the many advantages it offers in solving practical testing problems, including linking and equating, establishing the psychometric properties of items and assessments, optimizing the efficiency of test delivery through tailored assessment systems, and coupling assessment development and scoring procedures with the cognitive attributes involved in generating responses to items.

While many of the early developments in IRT focused on dichotomously scored items (Lord, 1980), the past 30 years have seen growing application of IRT to items having more than two scored outcomes. Such items are commonly referred to as *polytomous* items. Polytomous items are used in a variety of settings, including the scoring of rated tasks, the scoring of testlets or groups of dependent dichotomous items, innovative item types, multiple-choice items for which the distinction between all distractors are retained for scoring purposes, and rating scales used to measure a host of psychological and behavioral traits. Now, more than 40 years since the seminal work of Bock (1972) and Samejima (1969, 1972) in developing the first widely adopted polytomous IRT models, numerous polytomous IRT models have been proposed and applied in practice. These models are more complex than their dichotomous counterparts, having more parameters and a more sophisticated mathematical form. As a result, individuals in the fields of assessment and measurement often have a weaker understanding of polytomous IRT models than of widely adopted dichotomous models.

Randall David Penfield, Professor and Chair of the Department of Educational Research Methodology, The University of North Carolina at Greensboro, 1300 Spring Garden, St. Greensboro, NC 27412; rdpenfie@uncg.edu.

This instructional module provides an accessible overview of the most widely used polytomous IRT models. By design, this account is intended to be simple, nontechnical, and focused on the most frequently encountered polytomous models in practice. Individuals seeking a more comprehensive treatment of polytomous IRT models and associated estimation techniques are referred to the works of Baker and Kim (2004), de Ayala (2009), Nering and Ostini (2010), Ostini and Nering (2006), van der Linden and Hambleton (1997), and Wright and Masters (1982). In addition, polytomous IRT model summaries that are more technical in nature are provided by Mellenbergh (1995), Samejima (1996), Thissen and Steinberg (1986) and van der Ark (2001).

Overview of Dichotomous IRT Models

Before embarking on a discussion of polytomous IRT models, it will prove useful to review some general IRT principles and terminology in the context of dichotomous items. To this end, consider an assessment consisting of a series of items, whereby each item is scored into a specified number of categories appropriate for estimating the respondent's level of the trait measured by the assessment, what I will refer to hereafter as the *target trait*. Most often in educational assessment the target trait is a particular knowledge, skill, or ability. The scored outcomes to the ith item of the assessment are denoted here by Y_i. In the case of dichotomous items, the outcomes of Y_i are typically represented by $Y_i = 0$ (incorrect) and $Y_i = 1$ (correct). At its core, IRT is based on establishing a unique model for each outcome of Y_i that specifies the probability of observing the outcome as a function of the target trait. For a dichotomous item there are two possible outcomes of Y_i (0 and 1), and thus IRT generates two models; one specifying the probability of $Y_i = 0$ and one specifying the probability of $Y_i = 1$. Each of these two models is referred to

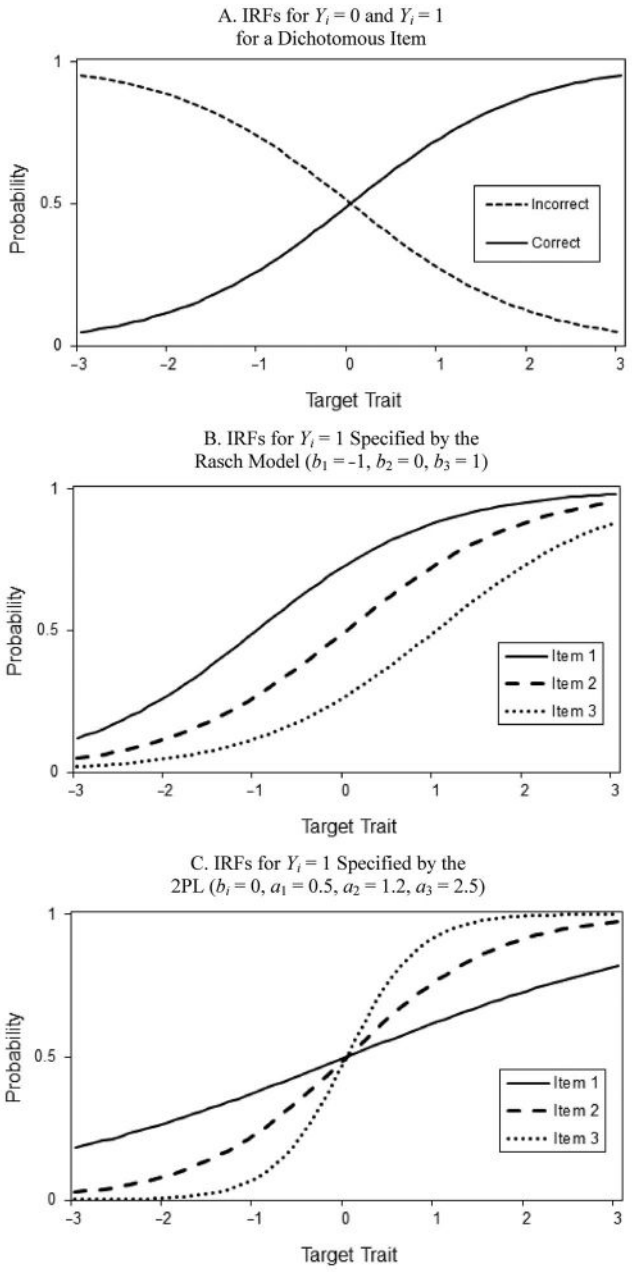

FIGURE 1. IRFs for dichotomous items.

here as an item response function (IRF) because they specify the probability of Y_i (0 or 1) as a function of the target trait. A visual representation of the IRFs for a dichotomous item is presented in Figure 1A. The horizontal axis of Figure 1A reflects the target trait continuum, and the vertical axis represents the probability of observing each outcome of Y_i. The IRF for $Y_i = 0$ is near unity at low levels of target trait, and diminishes toward zero at high levels of target trait. Conversely, the IRF for $Y_i = 1$ is near zero at low levels of target trait, and increases toward unity at high levels of target trait. It should be noted that the IRFs shown in Figure 1A correspond to the Rasch model for the IRFs of dichotomous items, which is just one of several models in widespread use for the IRFs of dichotomous items. More information about the Rasch model is provided in the paragraphs that follow.

Parametric IRT approaches for dichotomous items employ an equation to specify the IRFs for $Y_i = 1$ and $Y_i = 0$. The most common equation used for dichotomous items is a logistic model, which provides the "S-shaped" curves seen in Figure 1A. Several different logistic models are in common use, and I introduce two such models here as they play a fundamental role in the development of polytomous IRT models. The first dichotomous IRT model is the Rasch model (Rasch, 1960), which specifies the IRF for $Y_i = 1$ by

$$P_{i1}(\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}. \tag{1}$$

In Equation 1, the symbol θ represents the target trait, $P_{i1}(\theta)$ denotes that this equation specifies the probability of $Y_i = 1$ for a particular value of θ for item i, and b_i represents a difficulty parameter of item i. The value of b_i determines the horizontal location of the IRF for $Y_i = 1$; as b_i increases, the IRF shifts to the right, and the item difficulty increases. More specifically, b_i corresponds to the value of θ at which $P_{i1}(\theta) = .5$ (i.e., probability of correct response equals .5). The IRF for $Y_i = 1$ shown in Figure 1A follows the Rasch model with $b_i = 0$, and thus $P_{i1}(\theta) = .5$ occurs at the target trait value of $\theta = b_i = 0$.

Just as there exists an equation for the IRF for $Y_i = 1$ (Equation 1), there exists an equation for the IRF for $Y_i = 0$. Because dichotomously scored items maintain the property that $P_{i0}(\theta) + P_{i1}(\theta) = 1$ for any particular value of θ, it must be the case that $P_{i0}(\theta) = 1 - P_{i1}(\theta)$. As a result, the Rasch model IRF for $Y_i = 0$ can be obtained directly from the IRF of $Y_i = 1$ using

$$\begin{aligned} P_{i0}(\theta) &= 1 - \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \\ &= \frac{1}{1 + \exp(\theta - b_i)}. \end{aligned} \tag{2}$$

The item difficulty parameter, b_i, is the same parameter as that used in the model for the IRF of $Y_i = 1$ shown in Equation 1. Thus, the Rasch model employs a single item parameter to specify the IRFs for both $Y_i = 0$ and $Y_i = 1$. An example of an IRF for $Y_i = 0$ following the Rasch model with $b_i = 0$ is shown in Figure 1A. Examining the IRFs in Figure 1A for $Y_i = 0$ and $Y_i = 1$ reveals another interpretation of b_i as the value of target trait at which the IRFs for $Y_i = 0$ and $Y_i = 1$ intersect. For the IRFs shown in Figure 1A the intersection occurs at $\theta = 0$, which is the value of b_i.

Admittedly, for dichotomous items the IRF for $Y_i = 0$ is simply the mirror image of that for $Y_i = 1$ (see Figure 1A), and thus when we visually examine the IRFs for a dichotomous item it is strict convention to display only the IRF for $Y_i = 1$ so as to avoid the unnecessary redundancy inherent in the display of the IRF for $Y_i = 0$. An example of this convention is demonstrated in Figure 1B, which presents the IRF for $Y_i = 1$ specified by the Rasch model for three hypothetical items having $b_1 = -1$, $b_2 = 0$, and $b_3 = 1$. Notice that the IRF for $Y_i = 1$ shifts to the right as b_i increases. Despite the convention to portray only the IRF for $Y_i = 1$, I make explicit up front that a dichotomous item actually has two IRFs (one for $Y_i = 0$ and one for $Y_i = 1$) to reinforce the defining property of IRT as specifying an IRF for each scored outcome of Y_i. This defining property of IRT will be more salient, and more important, when we extend the discussion to polytomous IRT models.

Equations 1 and 2 present the models of the IRFs for $Y_i = 1$ and $Y_i = 0$ under the Rasch model, which contains only a single location parameter (b_i) for each item. A more flexible model is the two-parameter logistic (2PL) model (Lord, 1980) form that specifies the IRF for $Y_i = 1$ by

$$P_{i1}(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \quad (3)$$

and the IRF for $Y_i = 0$ by

$$P_{i0}(\theta) = \frac{1}{1 + \exp[a_i(\theta - b_i)]}. \quad (4)$$

The 2PL model is similar to the Rasch model, but includes an additional parameter, a_i, that determines the steepness of the IRFs for $Y_i = 0$ and $Y_i = 1$. As the value of a_i increases, the steepness of the IRFs increases. As the steepness of the IRFs increase, the item is better able to differentiate, or discriminate, between individuals residing in different ranges of the target trait continuum. For this reason, a_i is commonly referred to as a discrimination parameter. Figure 1C presents the IRF for $Y = 1$ for three dichotomous items for which $a_1 = .5$, $a_2 = 1.2$, and $a_3 = 2.5$.

The flexibility of the 2PL model to specify IRFs with differential steepness across different items (as depicted in Figure 1C) is not available when IRFs are specified using the Rasch model. As can be seen in Figure 1B, when IRFs are specified according to the Rasch model there is a common IRF steepness (and thus common discrimination) shared across all items. Note that fixing $a_i = 1$ in Equations 3 and 4 for the 2PL model leads to the Rasch model shown in Equations 1 and 2. As a result, IRFs following the Rasch model can be viewed as a constrained form of the 2PL whereby all a_i are fixed to unity.

Visualizing Polytomous IRT Models

Having described the central IRT concepts in the context of dichotomous items, let us now extend these concepts to polytomous items. Similar to dichotomous items, IRT applied to polytomous items specifies a unique IRF for each outcome of Y_i; the only difference being that for polytomous items Y_i assumes more than two outcomes and thus there will be more than two IRFs for each item. As an example, let us consider a polytomously scored item—say, a rated task—for which there are three score categories that are scored as $Y_i = 0$, 1, and 2. For the time being let us assume that the score categories are in increasing order with respect to the target trait, such that $Y_i = 2$ reflects a higher degree of success than does $Y_i = 1$, which in turn reflects a higher degree of success than does $Y_i = 0$. A pictorial representation of the IRFs for $Y_i = 0$, 1, and 2 for this item is displayed in Figure 2A. Note that the IRF for $Y_i = 0$ (the lowest score category) is high at low levels of target trait and decreases to near zero as target trait increases, indicating that there is a high chance of observing a score of 0 for individuals low in target trait and this chance decreases as level of target trait increases. Conversely, the IRF for $Y_i = 2$ (the highest score category) is near zero at low levels of target trait, and then increases to near unity at high levels of target trait. The IRF for $Y_i = 1$ is near zero at low and high levels of the target trait, and increases for moderate levels of the target trait. At any particular value of target trait the sum of the height of the three IRFs equals unity, which can represented as $P_{i0}(\theta) + P_{i1}(\theta) + P_{i2}(\theta) = 1$.

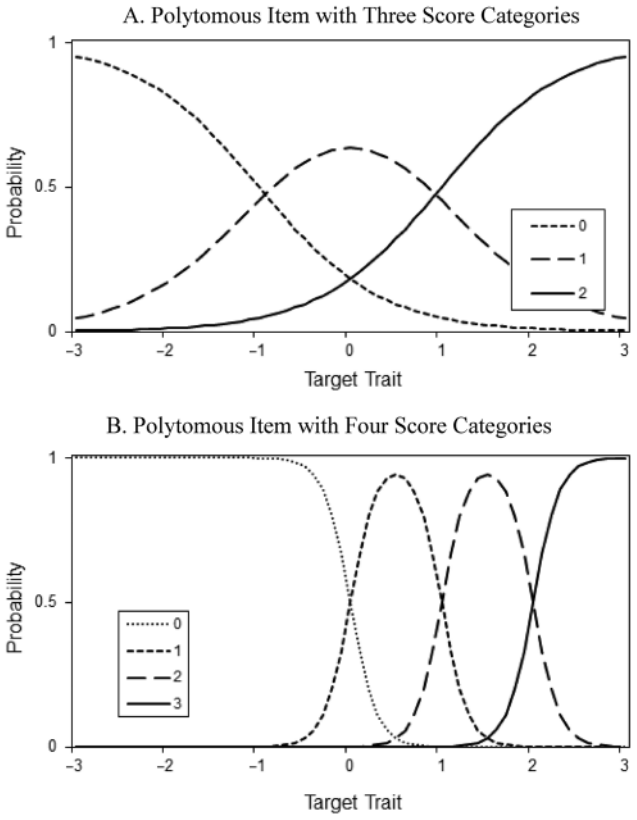

FIGURE 2. IRFs for polytomous items.

The IRFs for the polytomous item shown in Figure 2A reflect just one of many forms that the IRFs can assume; different polytomous items will have IRFs having different locations and shapes. As was the case with dichotomous items, the particular form of a polytomous item's IRFs determines the item's difficulty and discrimination. However, because polytomous items have more than two IRFs, defining item difficulty and discrimination in polytomous items is less transparent than in dichotomous items. For example, recall that item difficulty in dichotomous items can be quantified by the value of b_i, which is the target trait value at which the IRFs for $Y_i = 0$ and $Y_i = 1$ intersect. For polytomous items, however, there are more than two IRFs and thus a single point of intersection between the IRFs does not exist. As a result, item difficulty cannot be completely and unambiguously represented by a single parameter value. Rather, a comprehensive quantification of item difficulty for polytomous items must consider the location of all IRFs, which involves the consideration of more than one b_i parameter.

Let's now elaborate on our discussion of item difficulty in the context of polytomous IRT models. An item's difficulty corresponds to the target trait value(s) where the item's scored outcomes provide the greatest information concerning the respondent's target trait. In general, the greatest amount of information is generated at target trait values where adjacent IRFs intersect because it is at these target trait values where individuals are expected to transition from one outcome (e.g., $Y_i = 0$) to a higher outcome (e.g., $Y_i = 1$), and thus where the item response most clearly differentiates between individuals in different ranges of the trait continuum. For example, the dichotomous item shown in Figure 1A has a point of intersection at $\theta = 0$, and it is here that the item provides

its greatest information because the item's outcomes most clearly differentiate between individuals having low and high levels of target trait. This item provides little information at high levels of target trait (e.g, $\theta = 2$) because nearly all individuals at high levels of target trait will have a correct response, and thus the item is ineffective at differentiating between individuals with target trait values this high. Similarly, the item provides little information at low levels of target trait because nearly all individuals at low target trait levels will answer incorrectly. In contrast to dichotomous items, polytomous items have multiple points of intersection, and thus item difficulty concerns the location of intersection for each pair of adjacent outcomes of Y_i. For the polytomous item depicted in Figure 2A, this amounts to consideration of the value of target trait at which the IRFs for $Y_i = 0$ and $Y_i = 1$ intersect ($\theta = -1$), as well as the value of target trait at which the IRFs for $Y_i = 1$ and $Y_i = 2$ intersect ($\theta = 1$). Thus, when conceptualizing item difficulty in polytomous items, one must consider the range of target trait continuum about which the multiple points of intersection occur. For the item shown in Figure 2A, the range is approximately from $\theta = -1$ to $\theta = 1$, and thus is centered at $\theta = 0$. As a result, this item can be viewed as being of moderate difficulty because the outcomes of this item are most effective at differentiating between individuals with moderate values of target trait.

Figure 2B displays the IRFs for a different polytomous item for which there are four score categories. This item differs from that shown in Figure 2A with respect to difficulty. The item shown in Figure 2B has a relatively high level of difficulty, reflected by the fact that the intersections of adjacent IRFs occur at moderate-to-high target trait values. Specifically, these intersections occur at target trait values of $\theta = 0$, $\theta = 1$, and $\theta = 2$. Because these points of intersection span the range of 0 to 2, the outcomes of this item are most effective at differentiating between individuals with moderate and high values of target trait. As a result, this item reflects a higher level of difficulty than that of the polytomous item shown in Figure 2A.

As was the case for dichotomous items, the steepness of the IRFs for polytomous items informs the item's discrimination. As an example, the item shown in Figure 2B has IRFs that are steeper than those of the item shown in Figure 2A, representing a higher degree of discrimination for the item shown in Figure 2B. As will be described shortly, the steepness of the IRFs for polytomous items is described by one or more a_i parameters in a manner similar to that of the Rasch and 2PL models for dichotomous items.

The Building Blocks of Polytomous IRT Models: The Step Function

In the previous section, we examined visual representations of IRFs for polytomous items. Because IRT is based on specifying an IRF for each outcome of Y_i, the application of IRT to polytomous items requires the development of a mathematical model that can accomplish this in a manner similar to how the Rasch and 2PL models are used to specify the probability of $Y_i = 0$ and $Y_i = 1$ for dichotomous items (Equations 1–4). Yet, examining the IRFs of the example polytomous items shown in Figures 2A and 2B reveals that the forms of the IRFs for polytomous items are more complex than those of dichotomous items; not only are there more than two IRFs for polytomous items, but the IRFs can assume a wide range of shapes. The task of polytomous IRT models, then, is to specify the IRFs of polytomous items with sufficient flexibility to appropriately fit the wide range of potential IRF forms while using a relatively small number of parameters so that the models may be applied under the practical sample size constraints of real testing situations. To address this task (offering IRF flexibility using few parameters), polytomous IRT models have been developed using the concept of the *step function*, a term I borrow from Masters (1982), Muraki (1992), and Tutz (1990). Step functions are not the IRFs themselves, but are intermediary models that allow us to specify a wide range of IRF forms using relatively few item parameters.

To describe the concept of the step function, let us consider a sample polytomous item with score categories denoted by $Y_i = 0, 1, 2, 3$, reflecting ordered levels of performance with respect to a rated task. Let us assume that $Y_i = 0$ corresponds to no portion correct, $Y_i = 1$ corresponds to partially correct, $Y_i = 2$ corresponds to mostly correct, and, $Y_i = 3$ corresponds to completely correct. One can conceptualize the score an examinee receives as being determined by the success that she has had in transitioning, or stepping, to successively higher score categories. In the case of our example item, we have three possible steps: (a) step 1, reflecting the transition from "no portion correct" to "partially correct"; (b) step 2, reflecting the transition from "partially correct" to "mostly correct"; and (c) step 3, reflecting the transition from "mostly correct" to "completely correct."

For each of the three steps in our sample item we can consider the probability of being successful at the step as a function of target trait. For example, we can consider the probability of being successful at the first step (i.e., successful transition from "no portion correct" to "partially correct") as a function of target trait. We would expect that the probability of being successful at the first step increases with target trait; individuals with higher target trait will have a higher probability of successfully transitioning to "partially correct" and potentially beyond. A mathematical equation expressing the probability of success at the first step is the first step function. Using this same rationale, corresponding step functions can be considered for the probability of success at the second step and the probability of success at the third step. A pictorial representation of the three step functions underlying our example polytomous item is shown in Figure 3A. Notice that for each step function the probability of success increases as target trait increases. In this example, the step functions are equally spaced along the target trait continuum. In addition, each successive step function is shifted to the right, representing an increase in difficulty of transitioning across higher score categories.

In comparison to the step functions shown in Figure 3A, Figure 3B depicts the step functions associated with a different item having four score levels. The step functions of this item are steeper than those shown in Figure 3A, reflecting a higher degree of discrimination at each step. Also, the step functions shown in Figure 3B are not equally spaced; the first two step functions require a relatively low level of target trait to have a high chance of success, while the third step function requires a relatively high level of target trait to have a high chance of success.

An item with $m + 1$ score levels (i.e., $Y_i = 0, 1, 2, \ldots, m$) will have m step functions. For example, an item with three score levels ($Y_i = 0, 1, 2$) has $m = 2$ step functions;

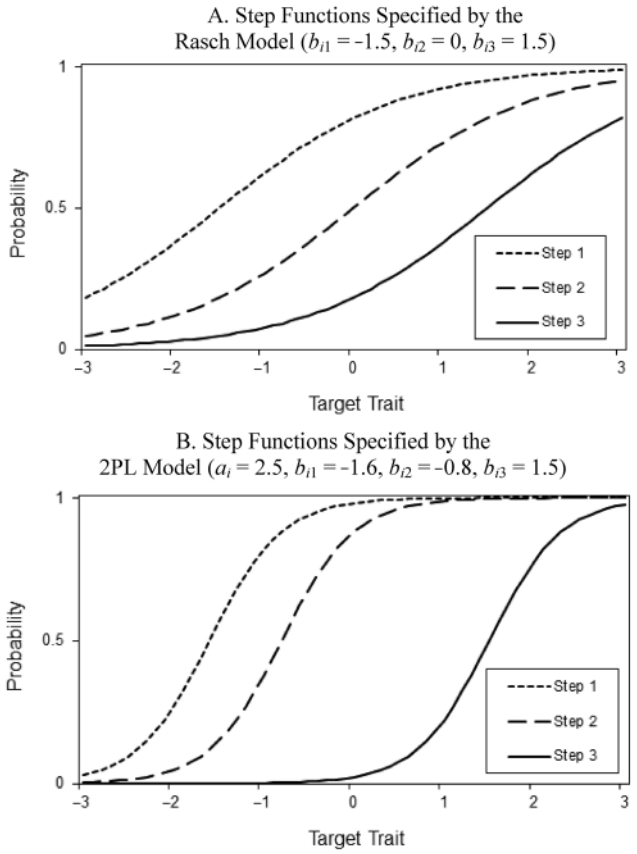

FIGURE 3. Step functions associated with a four-category polytomous item.

step function 1 reflects the probability of transitioning from $Y_i = 0$ to $Y_i = 1$, and step function 2 reflects the probability of transitioning from $Y_i = 1$ to $Y_i = 2$. Similarly, an item with five score levels ($Y_i = 0, 1, 2, 3, 4$) has $m = 4$, and thus will have $m = 4$ step functions. It is germane to note that a dichotomous item can be conceptualized as a particular case of a polytomous item with two score levels ($Y_i = 0, 1$), such that $m = 1$. In this instance, there is only a single step function reflecting the probability of successfully transitioning from $Y_i = 0$ to $Y_i = 1$, and this step function is given by the IRF for $Y_i = 1$.

From Figures 3A and 3B, we see that each step function represents a dichotomous event that looks very much like the IRF for $Y_i = 1$ for a dichotomous item. For example, the IRFs for $Y_i = 1$ for three dichotomous items depicted in Figure 1B have a similar appearance to the three step functions of a polytomus item shown in Figure 3A. Indeed, we can use our familiar models for the IRF for $Y_i = 1$ of dichotomous items to specify the step functions. For example, just as the Rasch model can be used to specify the probability of $Y_i = 1$ for a dichotomous item, it can also be used to specify the probability of success at the kth step. Let us denote the probability of success at the kth step for a particular value of θ by $\Psi_{ik}(\theta)$. If the Rasch model is used to specify the probability of success at the kth step, then the kth step function is expressed by

$$\Psi_{ik}(\theta) = \frac{\exp(\theta - b_{ik})}{1 + \exp(\theta - b_{ik})}. \quad (5)$$

I intentionally use the symbol $\Psi_{ik}(\theta)$ to denote the probability of success at the kth step, so that it is not confused with the probability of observing a particular value of Y_i (e.g, P_{i0}, P_{i1}, P_{i2}, etc.) as specified by the IRFs for Y_i. The b_{ik} parameter in Equation 5 reflects the difficulty of the kth step, specifying the value of target trait at which there is a .5 probability of success at that step, or simply where $\Psi_{ik}(\theta) = .5$. Figure 3A shows an example of step functions following the Rasch model having b_{ik} values of $b_{i1} = -1.5$, $b_{i2} = 0$, and $b_{i3} = 1.5$.

Using the Rasch model to specify the step functions, as was done in Figure 3A, has the advantage of parsimony, but it does not allow the step functions to vary with respect to their discrimination across different items. To allow the discrimination of the step functions to vary, we can specify step functions using the 2PL model as

$$\Psi_{ik}(\theta) = \frac{\exp[a_i(\theta - b_{ik})]}{1 + \exp[a_i(\theta - b_{ik})]}. \quad (6)$$

Notice that in Equation 3A there is a separate value of b_{ik} for each step function, as well as an item-level discrimination parameter a_i that is constant across all m steps. An example of step functions following the 2PL model is depicted in Figure 3B, whereby the three step functions have parameters $b_{i1} = -1.6$, $b_{i2} = -.8$, $b_{i3} = 1.5$, and $a_i = 2.5$.

Up to this point we have only considered the equations for step functions, $\Psi_{ik}(\theta)$, and not the equations for the actual IRFs of polytomous IRT models. As will be described in subsequent sections of this module, the equations for the step functions are used to generate the polytomous IRT models that specify the IRFs. The specific way in which the step functions are used to generate different polytomous IRT models is discussed throughout the remainder of the module.

Although all of the widely used polytomous IRT models are based on the concept of the step function, most polytomous IRT models define step functions using one of four different approaches: (a) the adjacent categories approach, (b) the continuation ratio approach, (c) the cumulative approach, and (d) the nominal approach. Figure 4 presents a diagram of the four approaches to defining step functions for an item having four response categories. Each of the four approaches shown in Figure 4 involves $m = 3$ step functions, but across the four approaches the step functions involve different outcomes of Y_i and define success (S) and failure (F) at each step differently. For example, the adjacent category approach defines the first step function using only score categories $Y_i = 0$ and $Y_i = 1$, such that $\Psi_{i1}(\theta)$ reflects the probability of successful transition from a score of $Y_i = 0$ (F) to a score of $Y_i = 1$ (S). This is consistent with the language used thus far in describing step functions. In contrast, the continuation ratio approach defines the first step function using score categories $Y_i = 0$, 1, 2, and 3 such that $\Psi_{i1}(\theta)$ reflects the probability of successful transition from a score of $Y_i = 0$ (F) to any other higher score category ($Y_i = 1, 2, 3$ are classified as S). The cumulative and nominal approaches provide two other approaches for defining step functions. Differences in the definition of the step functions across these four approaches lead to important differences in the applicability and interpretations of the respective polytomous IRT models. As a result, the approach used to define the step function serves as a fundamental characteristic of each polytomous IRT model. For this reason, I categorize polytomous IRT models into four major classes depending on which of the four approaches for defining the step functions is used. The remainder of this

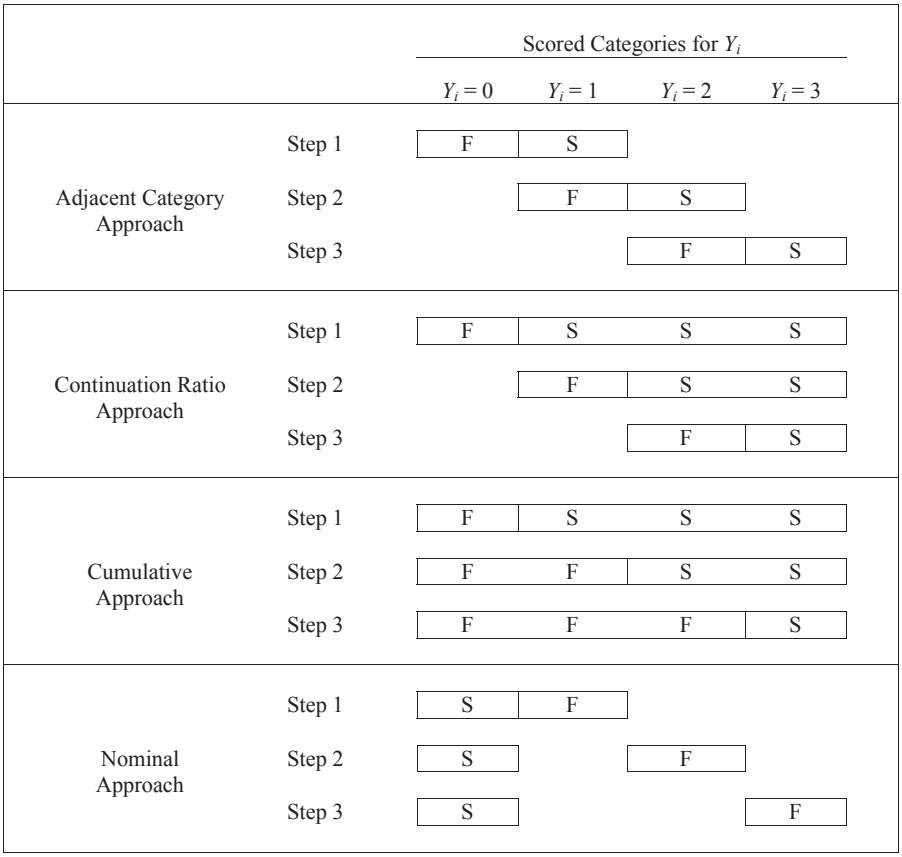

FIGURE 4. Description of the step functions for a four-category polytomous item. Within each step, S represents success and F represents failure. Under the nominal approach for defining step functions, the outcome $Y_i = 0$ represents the correct option of a multiple-choice item and $Y_i = 1, 2, 3$ represent distractor options.

module presents polytomous IRT models according to each of these four classes of step functions.

Adjacent Category Models

The first series of polytomous IRT models we discuss define step functions using the adjacent category approach, which defines the kth step function using only the adjacent pair of score categories $Y_i = k - 1$ and $Y_i = k$. For example, the first step function involves only the adjacent score categories $Y_i = 0$ and $Y_i = 1$, and specifies the probability of success on the first step as the probability that $Y_i = 1$ given that $Y_i = 0$ or $Y_i = 1$. Similarly, the second step function involves only the adjacent score categories $Y_i = 1$ and $Y_i = 2$ where the probability of success is defined as the probability that $Y_i = 2$ given that $Y_i = 1$ or $Y_i = 2$. The remaining step functions are defined in an analogous manner. The top portion of Figure 4 describes the steps associated with the adjacent categories approach. Note that under the adjacent category approach, each step involves only two score categories of Y_i, whereby success (S) at the step is assigned to the higher score category and failure on the step (F) is assigned to the lower score category and all other score categories are ignored.

A useful property of all adjacent category models is a relatively simple interpretation of b_{ik} (the location of the kth step). Recall that b_{ik} specifies the value of target trait at which $\Psi_{ik}(\theta) = .5$ (i.e., the value of the target trait at which the probability of success on the kth step equals .5). Because the adjacent category definition of step function defines the kth step function as the probability of $Y_i = k$ given that $Y_i = k$ or $k - 1$, the target trait value at which $\Psi_{ik}(\theta) = .5$ is necessarily the same value at which $P_{ik}(\theta) = P_{ik-1}(\theta)$, which reflects the value at which the IRFs for $Y_i = k$ and $Y_i = k - 1$ intersect. Thus, for adjacent category models b_{ik} reflects the value of target trait at which the IRFs of adjacent score categories intersect.

Of the widely used polytomous IRT models, three are based on the adjacent category definition of the step function: the partial credit model (PCM), the generalized PCM, and the rating scale model (RSM). While these three models all use the adjacent category definition of the step function, they differ with respect to the model used to specify the step functions and the constraints placed on the relative locations of the m step functions. The following sections describe these three models in more detail.

Partial Credit Model

The PCM (Masters, 1982) employs step functions defined according to the adjacent category approach and specifies the probability of success at the kth step, $\Psi_{ik}(\theta)$, using the Rasch model as shown in Equation 5. An example of step functions of the PCM is shown in Figure 3A. To describe how we arrive at the PCM equations specifying the IRFs for Y_i, let us take a closer look at the specific meaning of the adjacent category step functions. Recalling that the kth adjacent category step function specifies the probability that $Y_i = k$ given that $Y_i = k$ or $Y_i = k - 1$, we can express this

probabilistically as $\Psi_{ik}(\theta) = P_{ik}(\theta)/[P_{ik-1}(\theta) + P_{ik}(\theta)]$. For example, we can express the first step function as $\Psi_{i1}(\theta) = P_{i1}(\theta)/[P_{i0}(\theta) + P_{i1}(\theta)]$, the second step function as $\Psi_{i2}(\theta) = P_{i2}(\theta)/[P_{i1}(\theta) + P_{i2}(\theta)]$, the third step function as $\Psi_{i3}(\theta) = P_{i3}(\theta)/[P_{i2}(\theta) + P_{i3}(\theta)]$, and so on. Yet, we also have specified the form of each step function according to the Rasch model with parameter b_{ik}. Because each $\Psi_{ik}(\theta)$ can be expressed both as a function of IRFs ($P_{ik-1}(\theta)$ and $P_{ik}(\theta)$) and Rasch model parameters (Equation 5), there exists an algebraic link between the individual IRFs and the step function parameters. Using this algebraic link, we can conduct an algebraic manipulation of the terms associated with each $\Psi_{ik}(\theta)$ to express $P_{i0}(\theta), P_{i1}(\theta), \ldots, P_{im}(\theta)$ as a function of $b_{i1}, b_{i2}, \ldots, b_{im}$. I do not present the entire derivation here, but a complete description can be found in Masters (1982) and Muraki (1992). The resulting algebraic manipulation leads the IRF for $Y_i = 0$ to have the form

$$P_{i0}(\theta) = \frac{1}{1 + \sum_{r=1}^{m} \left[\exp \sum_{k=1}^{r} (\theta - b_{ik})\right]}, \tag{7}$$

and the IRF for $Y_i = j$, where $j > 0$, to have the form

$$P_{ij}(\theta) = \frac{\exp\left[\sum_{k=1}^{j} (\theta - b_{ik})\right]}{1 + \sum_{r=1}^{m} \left[\exp \sum_{k=1}^{r} (\theta - b_{ik})\right]}. \tag{8}$$

Equations 7 and 8 represent the PCM. Immediately apparent from Equations 7 and 8 is that the mathematical forms underlying the IRFs for the PCM are more complex than those of dichotomous IRT models. Specifically, unlike dichotomous IRT models that involve a single exponent term in the denominator, the PCM involves m exponent terms in its denominator, one for each step (where step is denoted by $r = 1, 2, \ldots, m$ in the denominator summation). The summation across the m exponent terms in the denominator is accomplished by the outer summation. The inner summation aggregates terms $(\theta - b_{i1}) + (\theta - b_{i2}) + \ldots + (\theta - b_{ir})$. For example, if we expanded the denominator of Equations 7 and 8 for a polytomous item having four score levels, the resulting expanded denominator term would take the form of $1 + [\exp(\theta - b_{i1})] + [\exp(\theta - b_{i1} + \theta - b_{i2})] + [\exp(\theta - b_{i1} + \theta - b_{i2} + \theta - b_{i3})]$. Also apparent from Equations 7 and 8 is that the PCM does not involve a discrimination parameter, a_i, which is a consequence of using the Rasch model to define the m step functions underlying the PCM.

Three characteristics of Equations 7 and 8 are important to note. First, the only item parameters involved are the m values of b_{ik}. Thus, for a set of n items there will be $n \times m$ item parameters. Second, the denominators of Equations 7 and 8 are identical, and thus the only component of the PCM equations that vary across the outcomes of Y_i is the numerator. This is similar to the IRFs for $Y_i = 0$ and $Y_i = 1$ of dichotomous IRT models (Equations 1–4), for which the denominators are identical for $Y_i = 0$ and $Y_i = 1$. Third, when there are only two outcomes of Y_i (i.e., $Y_i = 0, 1$), $m = 1$, and the PCM form shown in Equations 8 and 9 reduces down to the dichotomous Rasch model shown in Equation 1 (for $Y_i = 1$) and Equation 2 (for $Y_i = 0$).

Figure 5A displays the IRFs of a four-category item following the PCM with parameters $b_{i1} = -1.5$, $b_{i2} = 0$, and $b_{i3} = 1.5$. The values of b_{ik} correspond to the value of target trait at which the IRFs for adjacent categories intersect; the IRFs for $Y_i = 0$ and $Y_i = 1$ intersect at a target trait value of $\theta = -1.5$, the IRFs for $Y_i = 1$ and $Y_i = 2$ intersect at $\theta = 0$, and the IRFs for $Y_i = 2$ and $Y_i = 3$ intersect at $\theta = 1.5$. It may be helpful to note that the step functions for this very item are those depicted in Figure 3A (three step functions following the Rasch model with parameters $b_{i1} = -1.5$, $b_{i2} = 0$, and $b_{i3} = 1.5$). Thus, the IRFs shown in Figure 5A are derived directly from the step functions shown in Figure 3A through an algebraic manipulation of the step function terms.

The PCM receives widespread use, partly due to its parametric parsimony that allows it to be appropriately applied with sample sizes that might otherwise be insufficient to support higher parameterized models. While the appropriate sample size for a particular PCM calibration will depend on the number of items, the distribution of θ in the sample, and the intended uses of the polytomous items, researchers have demonstrated stable estimation of PCM parameters using sample sizes on the order of 300 (see the discussion in de Ayala, 2009, pp. 198–199, for a review of relevant literature). Ostini and Nering (2006) and Wright and Masters (1982) provide extensive descriptions of the PCM, its advantageous properties, and its application to applied assessment contexts. Despite the appeal of the PCM, the use of only m b_{ik} parameters limits its flexibility in generating IRFs of varying steepness. This limitation has led to the development of the more flexible generalized partial credit model (GPCM), as described in the following section.

Generalized Partial Credit Model

Like the PCM, the GPCM (Muraki, 1992) defines the step functions using the adjacent category definition. Unlike the PCM, however, the GPCM specifies the probability of success on the kth step, $\Psi_{ik}(\theta)$, using the 2PL model form in Equation 6 that includes an item-level discrimination parameter, a_i. The inclusion of a_i allows the GPCM more flexibility in the steepness of the step functions than that afforded by the PCM, and thus the potential for improved fit to the data compared to the PCM. The GPCM expresses the IRF for $Y_i = 0$ using the form

$$P_{i0}(\theta) = \frac{1}{1 + \sum_{r=1}^{m} \left(\exp \sum_{k=1}^{r} [a_i\,(\theta - b_{ik})]\right)}, \tag{9}$$

and the IRF for $Y_i = j$, where $j > 0$, using the form

$$P_{ij}(\theta) = \frac{\exp\left(\sum_{k=1}^{j} [a_i\,(\theta - b_{ik})]\right)}{1 + \sum_{r=1}^{m} \left(\exp \sum_{k=1}^{r} [a_i\,(\theta - b_{ik})]\right)}. \tag{10}$$

Note the similarity to the PCM equations for the IRFs shown in Equations 7 and 8; the only difference being the inclusion of a_i. Because a_i is constant across all m steps, any given item following the GPCM will have a total of $m + 1$ parameters; m values of b_{ik} and one value of a_i.

An example of an item following the GPCM is shown in Figure 5B for which $a_i = 2.5$, $b_{i1} = -1.6$, $b_{i2} = -.8$, and $b_{i3} = 1.5$. As was the case with the PCM, the IRFs for adjacent categories intersect at the target trait values equal to the corresponding value of b_{ik}; the IRFs for $Y_i = 0$ and $Y_i = 1$ intersect at the target trait value of $\theta = -1.6$, the IRFs for $Y_i = 1$ and $Y_i = 2$ intersect at $\theta = -.8$, and the IRFs for $Y_i = 2$ and $Y_i = 3$ intersect at $\theta = 1.5$. The step functions for this item are those shown in Figure 3B (three step functions following the 2PL model with parameters $a_i = 2.5$, $b_{i1} = -1.6$, $b_{i2} = -.8$, and $b_{i3} = 1.5$). Thus, the IRFs shown in Figure 5B are

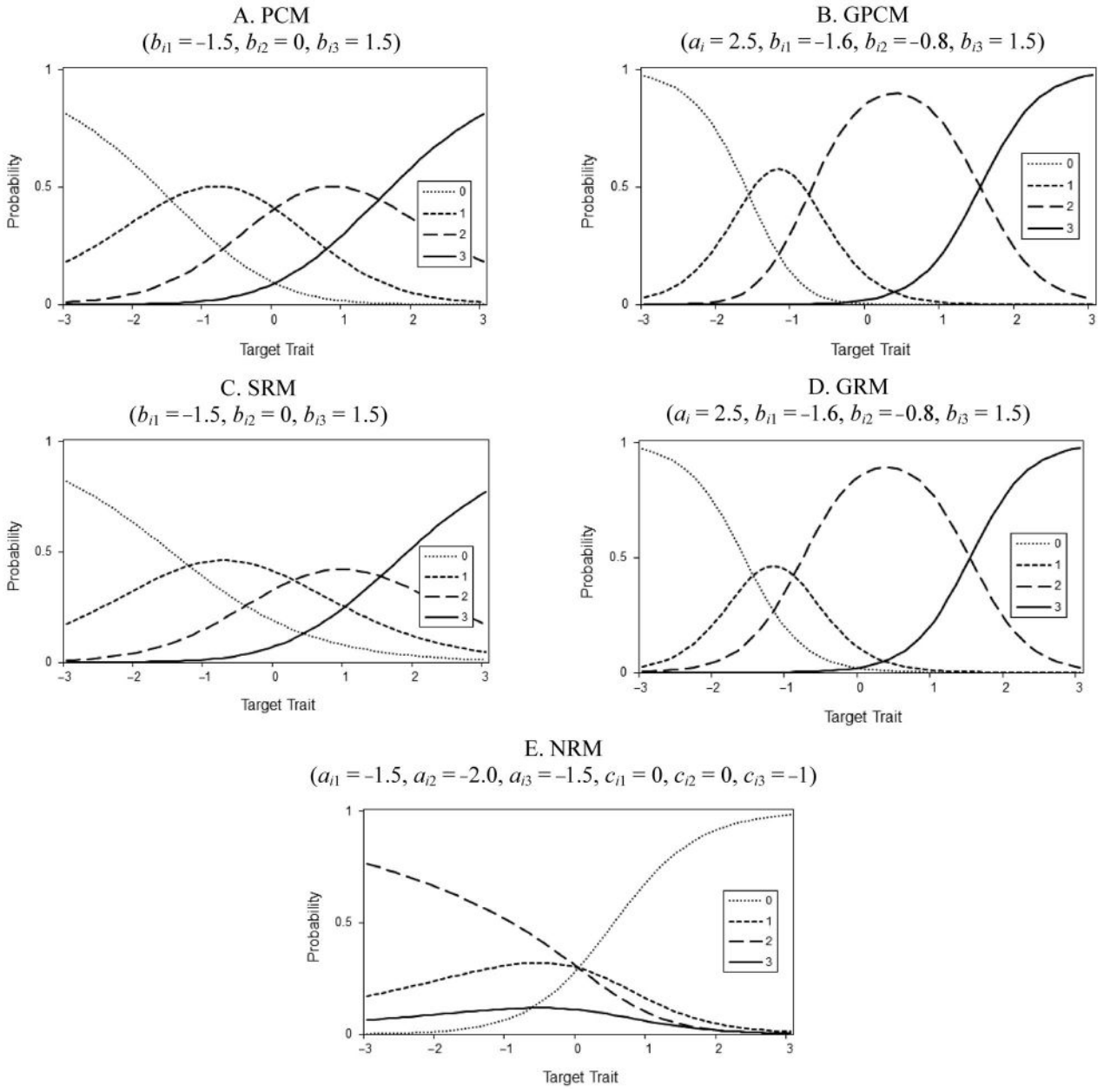

FIGURE 5. IRFs for a four-category polytomous item specified by five different models.

derived directly from the step functions shown in Figure 3B. Unlike the evenly spaced IRFs shown in Figure 5A, the item shown in Figure 5B has IRFs that are not evenly spaced, an outcome that results from the unevenly spaced step functions underlying the item (see Figure 3B). In particular, the IRFs for $Y_i = 0$, 1, and 2 are tightly bunched together, while the IRF for $Y_i = 3$ is centered a good distance above the other three score categories.

Rating Scale Model

There exist instances for which it is reasonable to assume that the distance between each pair of successive b_{ik} values of adjacent category step functions is identical for all items used in a particular instrument or scale (e.g., the distance between b_{i1} and b_{i2} is the same for all items, the distance between b_{i2} and b_{i3} is the same for all items, etc.). This might be the case, for example, in a rating scale whereby all items share a common set of anchor descriptors for the response categories, such as *strongly disagree, disagree, agree*, and *strongly agree*. If we can assume that each anchor has a common affective intensity across all items of the rating scale (e.g., the meaning of *strongly agree* is the same across all items), then it may be a reasonable assumption that the relative distance between the step functions associated with each anchor is constant across all items of the rating scale. This constancy in relative spacing of the b_{ik} across all items allows IRFs to be modeled using fewer parameters than what is used by the PCM or GPCM, which then allows IRFs to be modeled using smaller sample sizes than that required by the PCM and GPCM.

To describe how we can make use of this assumption, let us conceptualize each item as having an overall center, denoted by d_i, where the i subscript in d_i is used to make explicit that d_i is specific to the ith item. The value of d_i is defined as the

value of target trait about which the item's steps are centered, so that the average distance between the m values of b_{ik} for the ith item and d_i is zero. The value of d_i that satisfies this condition is given by $d_i = \sum b_{ik}/m$, or the average of the m values of b_{ik}. For example, the item in Figure 5A has $b_{i1} = -1.5$, $b_{i2} = 0$, $b_{i3} = 1.5$, and thus $d_i = (-1.5 + 0 + 1.5)/3 = 0$. Similarly, the item in Figure 5B has $b_{i1} = -1.6$, $b_{i2} = -0.8$, $b_{i3} = 1.5$, and thus $d_i = (-1.6 - .8 + 1.5)/3 = -.3$.

We can consider the distance of each b_{ik} from d_i, and denote this distance by $t_k = b_{ik} - d_i$. The value of t_k reflects how far the kth step lies from the item's center, d_i. The subscript i has been omitted from t_k because the relative spacing of the b_{ik} values is assumed constant across all items of a rating scale, thus constraining the values of t_k to be constant across all items. For example, the item shown in Figure 5B for which $b_{i1} = -1.6$, $b_{i2} = -.8$, $b_{i3} = 1.5$, and $d_i = -.3$, has values of t_k given by: $t_1 = b_{i1} - d_i = -1.6 - (-.3) = -1.3$, $t_2 = b_{i2} - d_i = -.8 - (-.3) = -.5$, and $t_3 = b_{i3} - d_i = 1.5 - (-.3) = 1.8$. Thus, the value of b_{i1} is -1.3 units from the item's center, the value of b_{i2} is $-.5$ units from the item's center, and the value of b_{i3} is 1.8 units from the item's center. The sign of t_k indicates whether the kth step function lays above or below the item's center. The situation of $t_k < 0$ indicates that the kth step lies below the item's center, and thus the kth step is relatively easy given the set of steps contained in the item. The situation of $t_k > 0$ indicates that the kth step lies above item's center, and thus is relatively difficult given the set of steps contained in the item. The situation of $t_k = 0$ indicates that the kth step resides at the item's center. The values of t_k are often referred to as *threshold* parameters (Andrich, 1978).

Using the d_i and t_k parameters, we can conceptualize b_{ik} (the location of the kth step) as being an aggregate of two components: (a) the item's center, d_i, which is unique to the ith item; and (b) the affective value of the kth transition across the score categories, t_k, which is constant across all items of the rating scale. We can thus express b_{ik} as the sum of these two components, given by $b_{ik} = d_i + t_k$. Substituting the term $[d_i + t_k]$ for b_{ik} in the PCM (Equations 7 and 8) leads to the RSM (Andrich, 1978), for which the IRF for $Y_i = 0$ is given by

$$P_{i0}(\theta) = \frac{1}{1 + \sum_{r=1}^{m} \left[\exp \sum_{k=1}^{r} (\theta - d_i - t_k)\right]}, \qquad (11)$$

and the IRF for $Y_i = j$, where $j > 0$, is given by

$$P_{ij}(\theta) = \frac{\exp\left[\sum_{k=1}^{j} (\theta - d_i - t_k)\right]}{1 + \sum_{r=1}^{m} \left[\exp \sum_{k=1}^{r} (\theta - d_i - t_k)\right]}. \qquad (12)$$

Because the RSM is a constrained form of the PCM, such that t_k are constrained to be constant across all items, the RSM requires fewer parameters than the PCM. For a set of n items the RSM contains only $n + m$ parameters; n values of d_i (one for each item) and only one set of m values of t_k that are constant across all items. This is notably fewer parameters than the $n \times m$ parameters required for the PCM. As a result, it is often possible to appropriately fit the RSM using smaller sample sizes than what is typically required for the PCM. Indeed, researchers have found stable item parameter estimation for the RSM with sample sizes on the order of 200 or fewer (see de Ayala, 2009, p. 199, for a broader discussion).

Continuation Ratio Models

Recall that the adjacent category models (PCM, GPCM, RSM) define the kth step using the score categories $Y_i = k-1$ and $Y_i = k$, whereby success at the kth step is defined as the outcome $Y_i = k$. The continuation ratio approach to defining the kth step takes a different approach, whereby the kth step is defined as advancing to score category k or higher ($Y_i \geq k$) given that the respondent has achieved at least score category $k-1$ ($Y_i \geq k-1$). Thus, the kth step defines a failure as $Y_i = k-1$ and a success as $Y_i \geq k$. This is shown schematically in Figure 4. As was the case with the adjacent categories approach, there are m steps under the continuation ratio approach. Furthermore, the mth step (the final step) of the continuation ratio approach is identical to the mth step of the adjacent categories approach; in both instances the mth step involves outcomes $Y_i = m-1$ and $Y_i = m$, for which $Y_i = m$ is designated as success.

The continuation ratio approach to defining the kth step is often referred to as a sequential step approach because it describes sequential advancement across successively higher score categories when the score categories reflect the successful application of a hierarchically ordered set of skills or processes. As an example, consider a math task involving the application of three components (components 1, 2, and 3) such that the demonstration of the three components is assumed to be hierarchical; success on component 3 presumes success on component 2, which in turn presumes success on component 1. The first step addresses whether or not the respondent is successful on component 1. Due to the assumed sequential process underlying the score categories, respondents who are not successful on component 1 are assigned $Y_i = 0$, and thus cannot advance to a higher score category. However, respondents who are successful at component 1 (and thus have $Y_i \geq 1$) advance to the second step, which addresses whether or not the respondent is successful on component 2. Respondents who are not successful on component 2 are assigned $Y_i = 1$, and thus cannot advance to a higher score category. Respondents who are successful on component 2 (and thus have $Y_i \geq 2$) advance to the third step, which addresses whether or not the respondent is successful on component 3. Respondents who are not successful at component 3 are assigned a score of $Y_i = 2$ and those who are successful are assigned a score of $Y_i = 3$.

As was the case with the adjacent category approach, the probability of success at the kth step, $\Psi_{ik}(\theta)$, can be specified using the Rasch model as described in Equation 5, or the 2PL model as described in Equation 6. Using the Rasch model to specify the step functions yields the continuation ratio model commonly referred to as the sequential Rasch model (SRM; Tutz, 1990). The sequential nature of the step functions underlying the SRM allows a relatively simple formulation of the IRFs for Y_i based on the m step functions. In particular, the outcome $Y_i = k$ reflects a situation of success at steps 1, 2, ..., k, and failure at step $k + 1$. For example, $Y_i = 2$ corresponds to a success at step 1, success at step 2, and then failure at step 3. It follows that the probability of $Y_i = k$ can be specified by the joint probability of success at steps 1, 2, ..., k and failure at step $k + 1$, where the probability success and failure at the kth step are given by $\Psi_{ik}(\theta)$ and $[1 - \Psi_{ik}(\theta)]$, respectively.

As an example of how to obtain the IRFs for the SRM, consider a rated task scored in relation to a hierarchically

ordered set of three steps, thus yielding four score categories. The IRF for $Y_i = 0$ is equal to the probability of failure at the first step, which is obtained by $P_{i0}(\theta) = 1 - \Psi_{i1}(\theta)$. The IRF for $Y_i = 1$ is equal to the probability of success at the first step and failure at the second step, which is obtained by $P_{i1}(\theta) = \Psi_{i1}(\theta) \times [1 - \Psi_{i2}(\theta)]$. The IRF for $Y_i = 2$ is equal to the probability of success at the first and second steps and failure at the third step, which is obtained by $P_{i2}(\theta) = \Psi_{i1}(\theta) \times \Psi_{i2}(\theta) \times [1 - \Psi_{i3}(\theta)]$. Finally, the IRF for $Y_i = 3$ is equal to the probability of success at all three steps, which is obtained by $P_{i3}(\theta) = \Psi_{i1}(\theta) \times \Psi_{i2}(\theta) \times \Psi_{i3}(\theta)$. In all of these instances, the values of $\Psi_{ik}(\theta)$ are obtained using the Rasch model as specified in Equation 5.

The general equations used in obtaining the IRFs for the SRM for any particular number of score categories can be obtained by substituting the Rasch model for $\Psi_{ik}(\theta)$ across the m steps. Upon making this substitution, we find that the SRM specifies the IRF for $Y_i = 0$ by

$$P_{i0}(\theta) = \frac{1}{1 + \exp(\theta - b_{i1})} \tag{13}$$

and the IRF for $Y_i = j$, where $j > 0$, by

$$P_{ij}(\theta) = \frac{\prod_{k=1}^{j} \exp(\theta - b_{ik})}{\prod_{k=1}^{j+1} \{1 + \exp(\theta - b_{ik})\}}, \tag{14}$$

where $\exp(\theta - b_{ik}) = 0$ for $k = m + 1$.

An example of the IRFs for the SRM is presented in Figure 5C for a four-category item having parameters $b_{i1} = -1.5$, $b_{i2} = 0$, and $b_{i3} = 1.5$. The step functions underlying the IRFs in Figure 5C are those shown in Figure 3A. You will notice that the IRFs look similar to those of the PCM having the same item parameters in Figure 5A, but not identical. In particular, the symmetric form of the PCM in Figure 5A does not hold for the SRM in Figure 5C. In addition, the interpretation of the b_{ik} for the SRM is different than that of the PCM; under the PCM the value of b_{ik} corresponds to the value of θ at which the IRFs for $Y_i = k - 1$ and $Y_i = k$ intersect, while under the SRM b_{ik} corresponds to the value of θ at which the height of the IRF for $Y_i = k-1$ equals the sum of the height of the IRFs for $Y_i \geq k$.

Generalized versions of the SRM have also been proposed (Hemker, van der Ark, & Sijtsma, 2001; Kim, 2002; Mellenbergh, 1995) for which the kth step function is specified using the 2PL model. One such generalization of the SRM employs a unique value of a_{ik} and b_{ik} for each of the m step functions, yielding a total of $2m$ parameters for each item. Another generalization of the SRM (Hemker et al., 2001) includes a unique value of b_{ik} for each step function, but only one item-level value of a_i that is constant across all m step functions. To date, these generalized versions of the SRM have been given limited attention in the literature.

Cumulative Models

A third approach to defining step functions is the cumulative approach, in which the kth step function defines a failure as $Y_i < k$ and a success as $Y_i \geq k$. Figure 4 depicts the cumulative approach to defining the step functions for an item with four score categories. The first step function contrasts $Y_i = 0$ (F) to $Y_i = 1, 2, 3$ (S) such that the probability of success on the first step, $\Psi_{i1}(\theta)$, represents the probability that $Y_i = 1, 2$, or 3. The second step function contrasts $Y_i = 0, 1$ (F) to $Y_i = 2, 3$ (S) such that the probability of success on the second step, $\Psi_{i2}(\theta)$, represents the probability that $Y_i = 2$ or 3. The third step function contrasts $Y_i = 0, 1, 2$ (F) to $Y_i = 3$ (S) such that the probability of success on the third step, $\Psi_{i3}(\theta)$, represents the probability that $Y_i = 3$. A salient property of the cumulative definition of step functions is that each step function involves all $m + 1$ score categories; the only difference between the m step functions is the point of dichotomization that differentiates failure from success. Note that the first-step function of the cumulative approach is identical to the first step function of the continuation ratio approach; both contrast the lowest score category with all other higher score categories.

The predominant model employing the cumulative step function is the graded response model (GRM; Samejima, 1969, 1972). The GRM specifies each step function using the 2PL form shown in Equation 6, whereby there is a common value of a_i for all m steps of the ith item, and a separate b_{ik} parameter for each step of the item. A salient characteristic of the GRM is that there is a separate value of a_i for each item, which affords the GRM considerable flexibility in fitting different IRF forms across different items. An appealing property of the GRM is that the IRFs can be obtained directly from the difference between adjacent step functions. Specifically, the IRF for $Y_i = k$ is obtained directly from the difference between $\Psi_{ik}(\theta)$ and $\Psi_{ik+1}(\theta)$. As an example, consider an item with four score categories. The IRF for $Y_i = 0$ is given by $P_{i0}(\theta) = 1 - [P_{i1}(\theta) + P_{i2}(\theta) + P_{i3}(\theta)]$, which is equivalent to $1 - \Psi_{i1}(\theta)$. The IRF for $Y_i = 1$ is equivalent to $P_{i1}(\theta) = [P_{i1}(\theta) + P_{i2}(\theta) + P_{i3}(\theta)] - [P_{i2}(\theta) + P_{i3}(\theta)]$, which is equivalent to $\Psi_{i1}(\theta) - \Psi_{i2}(\theta)$. The IRF for $Y_i = 2$ represents $P_{i2}(\theta) = [P_{i2}(\theta) + P_{i3}(\theta)] - P_{i3}(\theta)$, which is equivalent to $\Psi_{i2}(\theta) - \Psi_{i3}(\theta)$. Lastly, the IRF for $Y_i = 3$ represents $P_{i3}(\theta)$, which is equivalent to $\Psi_{i3}(\theta)$.

In general, for an item with $m + 1$ ordered score categories, the GRM specifies the IRF of $Y_i = 0$ using

$$P_{i0}(\theta) = 1 - \Psi_{i1}(\theta), \tag{15}$$

and the IRF of $Y_i = j$, where $j > 0$, using

$$P_{ij}(\theta) = \Psi_{ij}(\theta) - \Psi_{ij+1}(\theta), \tag{16}$$

where $\Psi_{im+1}(\theta) = 0$. Because the IRFs for any outcome of Y_i cannot be negative, it must be the case that $\Psi_{ij}(\theta) \geq \Psi_{ij+1}(\theta)$. In order to satisfy this condition, the values of b_{ik} must be successively increasing across the m steps, such that $b_{ij} \leq b_{ij+1}$.

Figure 5D shows the IRFs for a four-category item following the GRM with parameters $a_i = 2.5$, $b_{i1} = -1.6$, $b_{i2} = -0.8$, and $b_{i3} = 1.5$. The step functions for this item are shown in Figure 3B. Note that the IRFs shown in Figure 5D are obtained directly from the difference between the adjacent step functions of Figure 3B. For example, the IRF for $Y_i = 1$ in Figure 5D is equal to the difference between the step function 1 and step function 2 in Figure 3B. Similarly, the IRF for $Y_i = 2$ is equal to the difference between step function 2 and step function 3.

The cumulative definition of the step function leads to a different interpretation of the b_{ik} than what is observed for the adjacent category and continuation ratio definition. Under the cumulative definition, b_{i1} corresponds to the target trait value at which $P_{i0}(\theta) = .5$, b_{im} corresponds to the target trait value at which $P_{im}(\theta) = .5$, and for $1 > k < m$ the value of $(b_{ik} + b_{ik+1})/2$ corresponds to the target trait value at which

the IRF for $Y_i = k$ peaks (i.e., the modal point of the IRF for $Y_i = k$). For example, the IRFs displayed in Figure 5D are based on the GRM having $b_{i1} = -1.6$, $b_{i2} = -.8$, and $b_{i3} = 1.5$. In this case we know that the IRF for $Y_i = 0$ is .5 at $\theta = -1.6$, the IRF for $Y_i = 1$ is highest at $\theta = [-1.6 + (-.8)]/2 = -1.2$, the IRF for $Y_i = 2$ is highest at $\theta = (-.8 + 1.5)/2 = .35$, and the IRF for $Y_i = 3$ is .5 at $\theta = 1.5$.

Nominal Models

Up to this point in the discussion, we have considered polytomous items for which there is a known ordering of the outcomes of Y_i prior to fitting the IRT model. We now turn our attention to polytomous items where the outcomes have no implicit a priori ordering across all $m + 1$ outcomes. The most common example of this situation in educational testing is the multiple-choice item when all options—correct option and distractors alike—are retained in the scoring of the item, rather than collapsing all distractors into a single "incorrect" category. In this context, the test developer typically would not be aware of the ordering of the m distractors of the item prior to data analysis, and thus ordinal polytomous items described thus far in the module would not be appropriate. The nominal set of models provides the flexibility of not requiring the a priori specification of score category ordering.

Nominal models are based on the nominal step function. To describe the nominal step function, let us consider a multiple-choice item for which we arbitrarily assign the correct option to $Y_i = 0$, and the m distractors to $Y_i = 1, 2, \ldots, m$. The nominal approach to defining step functions defines the jth step function as the probability that $Y_i = 0$ given that $Y_i = 0$ or $Y_i = k$. The nominal step function is shown in the bottom portion of Figure 4 for the situation of a four-category multiple-choice item. The first step function contrasts $Y_i = 0$ (S) to $Y_i = 1$ (F) such that the probability of success on the first step, $\Psi_{i1}(\theta)$, represents the probability that $Y_i = 0$ given that $Y_i = 0$ or $Y_i = 1$. The second step function contrasts $Y_i = 0$ (S) to $Y_i = 2$ (F) such that the probability of success on the second step, $\Psi_{i2}(\theta)$, represents the probability that $Y_i = 0$ given that $Y_i = 0$ or $Y_i = 2$. The third step function contrasts $Y_i = 0$ (S) to $Y_i = 3$ (F) such that the probability of success on the third step, $\Psi_{i3}(\theta)$, represents the probability that $Y_i = 0$ given that $Y_i = 0$ or $Y_i = 3$. It is important to note that for didactic purposes I have used $Y_i = 0$ (the correct option) as a reference category that is employed in each of the m step functions. Doing this allows useful interpretation of the obtained item parameters. In practice, however, any option could be used as the reference category, but the interpretation of the item parameters will depend on which option is selected as the reference category.

The nominal definition of the step function is similar to the adjacent category definition of step function in that they both contrast two outcomes of Y_i. In the adjacent category approach, the kth step function contrasts $Y_i = k$ to $Y_i = k - 1$, and in the nominal approach, the kth step function contrasts $Y_i = 0$ to $Y_i = k$. In this manner, the adjacent category approach can be viewed as a constrained form of the nominal approach that involves only adjacent pairs of score categories.

The most widely used model adopting the nominal approach for defining the step functions is the nominal response model (NRM; Bock, 1972). Under the NRM, the probability of success at the kth step, $\Psi_{ik}(\theta)$, is specified using a modified form of the 2PL shown in Equation 6, given by

$$\Psi_{ik}(\theta) = \frac{\exp(-c_{ik} - a_{ik}\theta)}{1 + \exp(-c_{ik} - a_{ik}\theta)}. \quad (17)$$

In this parameterization, each step has a unique value of c_{ik} and a_{ik}, where c_{ik} serves as a location parameter that is similar, although not identical, to the b_{ik} parameter used for the adjacent categories, cumulative, and continuation ratio models. While the b_{ik} parameter used for the adjacent categories, cumulative, and continuation ratio models can be interpreted as the location of the kth step function (i.e., the value of θ at which the probability of success on the kth step is .5), the location of the kth step function underlying the NRM occurs at $\theta = -c_{ik}/a_{ik}$. Thus, based on the parameterization shown in Equation 17 for the kth step function, the intersection of the IRFs for $Y_i = 0$ and $Y_i = k$ occurs at $\theta = -c_{ik}/a_{ik}$. In this manner, the value of c_{ik} represents a rescaling of the point of intersection of the IRFs for $Y_i = 0$ and $Y_i = k$; that is, the value of c_{ik} corresponds to the factor of $-a_{ik}$ multiplied by the value of θ at which the IRFs for $Y_i = 0$ and $Y_i = k$ intersect.

The IRFs of the NRM are obtained by rearranging the terms of the nominal step functions in a manner similar to that described for the PCM. The resulting IRF for $Y_i = 0$ (correct response in this instance) has the form

$$P_{i0}(\theta) = \frac{1}{1 + \sum_{k=1}^{m} \exp(c_{ik} + a_{ik}\theta)}, \quad (18)$$

and the IRF for $Y_i = j$, where $j > 0$ (each of the m distractors), has the form

$$P_{ij}(\theta) = \frac{\exp(c_{ik} + a_{ik}\theta)}{1 + \sum_{k=1}^{m} \exp(c_{ik} + a_{ik}\theta)}. \quad (19)$$

Any item following the NRM will contain a total of $2m$ parameters corresponding to the m values of c_{ik} and the m values of a_{ik}.

An example of the IRFs for an item following the NRM is shown in Figure 5E. For this item, $a_{i1} = -1.5$, $a_{i2} = -2.0$, $a_{i3} = -1.5$, $c_{i1} = 0$, $c_{i2} = 0$, $c_{i3} = -1$. As would be expected, the IRF for $Y_i = 0$ (correct option) increases as the target trait increases. However, the IRFs for $Y_i = 1$, 2, and 3 (the three distractors) assume a range of forms. The outcome of $Y_i = 2$ is most attractive to individuals with very low values of target trait, the outcome of $Y_i = 1$ is most attractive to individuals with moderate levels of target trait, and the outcome of $Y_i = 3$ is relatively unlikely to be selected across all levels of target trait. Notice that the target trait value at which the IRFs for $Y_i = 1$, 2, and 3 intersect the IRF for $Y_i = 0$ is equal to $-c_{ik}/a_{ik}$. For example, the IRF for $Y_i = 3$ intersects the IRF for $Y_i = 0$ at $1.0/-1.5 = -.67$.

A drawback of the NRM applied to multiple-choice items is that the IRF for the correct option has a lower asymptote of zero (i.e., approaches zero as the target trait level becomes very low), and thus the NRM lacks the flexibility to account for the chance of guessing the correct option by individuals with low levels of target trait. For this reason, several variations on the NRM have been proposed that account for guessing (Revuelta, 2005; Samejima, 1979; Thissen & Steinberg, 1984; Thissen, Steinberg, & Fitzpatrick, 1989), and these models are often termed multiple-choice models. The multiple-choice

models contain additional parameters over those contained in the NRM that allow the IRFs of all options to have nonzero lower asymptotes to reflect the chance of guessing. These models hold potential promise for increasing the information generated by multiple-choice items and for improving our understanding the properties of multiple-choice items.

Concluding Remarks

As assessment practices continue to advance and technology allows for increasing use of innovative item formats, polytomous models will no doubt become increasingly common. At the same time, advances in parameter estimation techniques, particularly those related to Markov chain Monte Carlo estimation (Kim & Bolt, 2007; Patz & Junker, 1999), will likely facilitate the application of complex polytomous IRT models to practical assessment contexts. This will bring with it continued need for understanding and applying polytomous IRT models, and this instructional module can serve as an accessible introduction to the topic. It is important to acknowledge that this module describes only the most commonly encountered models. Numerous other polytomous models have been proposed, and a complete description or listing of these is beyond the scope of this module. However, most of these other models are variations on the models described here. For this reason, the contents of this module should serve as a useful primer for readers exploring polytomous IRT models not described here.

A final issue that warrants mention is how one decides which polytomous IRT model is most appropriate for a given context. This decision should be based on consideration of several factors. One factor is the theoretical assumptions of the response process. For example, does one step function form (e.g., adjacent category, cumulative, or continuation ratio) or a particular constraint on location parameters (as are associated with the RSM) have theoretical appeal for a given context? If so, then there may be a theoretical justification using a particular polytomous IRT model. A second factor is sample size. In the presence of a relatively small sample size (say, less than 400), those models without a discrimination parameter (a_i) are likely the most viable (i.e., PCM, RSM, or SRM). A third factor involves the empirical testing of model–data fit; the model that best fits the data may be the preferred model to employ. A wide range of methods for evaluating model–data fit are available and the reader is referred to Ostini and Nering (2006) and Kang, Cohen, and Sung (2009) for more information on these methods in relation to polytomous IRT models. While the relative importance of each of these three factors (theoretical assumptions, sample size, model-data fit) will depend on the particular testing situation, each of these factors should be considered when developing a defensible argument for model selection.

Self-Test

1. How many step functions are associated with a polytomously scored item containing four score categories?
2. Which polytomous models employ an adjacent category definition of step function?
3. Which step function definition is adopted by the GRM?
4. Which polytomous models described in this module employ the dichotomous Rasch model in specifying each step function?
5. Describe what the first step function of the PCM represents.
6. An item having three score categories and following the GRM has values of its two step functions at $\theta = -1$ equal to $\Psi_{i1}(-1) = .5$ and $\Psi_{i2}(-1) = .2$. Based on this information, what is the probability of observing $Y_i = 0$, $Y_i = 1$, and $Y_i = 2$ on this item for an individual having $\theta = -1$?
7. Consider an item having three score categories that follow the sequential Rasch model. The values of the two step functions at $\theta = 0$ are given by $\Psi_{i1}(0) = .7$ and $\Psi_{i2}(0) = .4$. Based on this information, what is the probability of observing $Y_i = 0, Y_i = 1$, and $Y_i = 2$ on this item for an individual having $\theta = 0$?

Answers to Self-Test

1. Three step functions
2. PCM, GPCM, RSM
3. Cumulative
4. PCM, RSM, SRM
5. The probability of observing $Y_i = 1$ given that $Y_i = 0$ or $Y_i = 1$
6. $P_{i0} = .5, P_{i1} = .3, P_{i2} = .2$
7. $P_{i0} = .30, P_{i1} = .42, P_{i2} = .28$

References

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

Baker, F. B., & Kim, S. -H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

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

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Hemker, B. T., van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*, 487–506.

Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*, 499–518.

Kim, S.-H. (2002, June). *A continuation ratio model for ordered category items*. Paper presented at the annual meeting of the Psychometric Society, Chapel Hill, NC.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38–51.

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

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

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

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

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

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

Patz, R., & Junker, B. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institute.

Revuelta, J. (2005). An item response model for nominal data based on the rising selection ratios criterion. *Psychometrika*, *70*, 305–324.
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. *Pyschometrika*, Monograph Supplement No. 18.
Samejima, F. (1979). *A new family of models for the multiple choice item*. Research Report No. 79–4. Knoxville: Department of Psychology, University of Tennessee.
Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika*, *23*, 17–35.
Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501–519.
Thissen, D. J., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.
Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice items: The distractors are also part of the item. *Journal of Educational Measurement*, *26*, 161–176.
Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55.
van der Ark, L. A. (2001). Relationship and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*, 273–282.
van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

