CHAPTER 3

The Nominal Categories Item Response Model

David Thissen
The University of North Carolina at Chapel Hill

Li Cai
University of California, Los Angeles

with a contribution by
R. Darrell Bock
University of Illinois at Chicago

Introduction

Editor Introduction: This chapter elaborates the development of the most general polytomous IRT model covered in this book. It is the only model in this book that does not assume ordered polytomous response data and can therefore be used to measure traits and abilities with items that have unordered response categories. It can be used to identify the empirical ordering of response categories where that ordering is unknown *a priori* but of interest, or it can be used to check whether the expected ordering of response categories is supported in data. The authors present a new parameterization of this model that may serve to expand the model and to facilitate a more widespread use of the model. Also discussed are various derivations of the model and its relationship to other models. The chapter concludes with a special section by Bock, where he elaborates on the background of the nominal model.

The Original Context

The *nominal* categories model (Bock, 1972, 1997) was originally proposed shortly after Samejima (1969, 1997) described the first general item response theory (IRT) model for polytomous responses. Samejima's graded models (in normal ogive and logistic form) were designed for item responses that have some a priori order as they relate to the latent variable being

Y102002_Book.indb 43 2/17/10 4:26:49 PM

measured (θ); the nominal model was designed for responses with no predetermined order.

Samejima (1969) illustrated the use of the graded model with the analysis of data from multiple-choice items measuring academic proficiency. The weakness of the use of a graded model for that purpose arises from the fact that the scoring order, or relative degree of correctness, of multiple-choice response alternatives can only rarely be known a priori. That was part of the motivation for the development of the nominal model. Bock's (1972) presentation of the nominal model also used multiple-choice items measuring vocabulary to illustrate its application. Ultimately, neither Samejima's (1969, 1997) graded model nor Bock's (1972, 1997) nominal model has seen widespread use as a model for the responses to multiple-choice items, because, in addition to the aforementioned difficulty prespecifying order for multiple-choice alternatives, neither the graded nor the nominal model makes any provision for guessing. Elaborating a suggestion by Samejima (1979), Thissen and Steinberg (1984) described a generalization of the nominal model that *does* take guessing into account, and that multiple-choice model is preferable if IRT analysis of all of the response alternatives for multiple-choice items is required.

Current Uses

Nevertheless, the nominal model is in widespread use in item analysis and test scoring. The nominal model is used for three purposes: (1) as an item analysis and scoring method for items that elicit purely nominal responses, (2) to provide an empirical check that items expected to yield ordered responses have actually done so (Samejima, 1988, 1996), and (3) to provide a model for the responses to testlets. Testlets are sets of items that are scored as a unit (Wainer & Kiely, 1987); often testlet response categories are the patterns of response to the constituent items, and those patterns are rarely ordered a priori.

The Original Nominal Categories Model

Bock's (1972) original formulation of the nominal model was

$$T(u = k|\theta; \mathbf{a}, \mathbf{c}) = T(k) = \frac{\exp(z_k)}{\sum_i \exp(z_i)} \tag{3.1}$$

in which T, the curve tracing the probability that the item response u is in category k is a function of the latent variable θ with vector parameters a and c. In what follows we will often shorten the notation for the trace line to $T_{(k)}$, and in this presentation we number the response alternatives $k = 0,1,\ldots,m-1$ for an item with m response categories. The model itself is the so-called multivariate logistic function, with arguments

$$z_k = a_k\theta + c_k \tag{3.2}$$

in which z_k is a response process (value) for category k, which is a (linear) function of θ with slope parameter a_k and intercept c_k. Equations 3.1 and 3.2 can be combined and made more compact as

$$T(k) = \frac{\exp(a_k\theta + c_k)}{\sum_i \exp(a_i\theta + c_i)} \tag{3.3}$$

As stated in Equation 3.3, the model is twice not identified: The addition of any constant to either all of the a_ks or all of the c_ks yields different parameter sets but the same values of $T(k)$. As identification constraints, Bock (1972) suggested

$$\sum_{k=0}^{m-1} a_k = \sum_{k=0}^{m-1} c_k = 0 \tag{3.4}$$

implemented by reparameterizing, and estimating the parameter vectors α and γ using

$$\mathbf{a} = \mathbf{T}\boldsymbol{\alpha} \quad \text{and} \quad \mathbf{c} = \mathbf{T}\boldsymbol{\gamma} \tag{3.5}$$

in which "deviation" contrasts from the analysis of variance were used:

$$\underset{m\times(m-1)}{\mathbf{T}_{\mathbf{DEV}}} = \begin{bmatrix} \frac{1}{m} & \frac{1}{m} & \cdots & \frac{1}{m} \\ \frac{1}{m}-1 & \frac{1}{m} & \cdots & \frac{1}{m} \\ \frac{1}{m} & \frac{1}{m}-1 & \cdots & \frac{1}{m} \\ \vdots & \vdots & & \vdots \\ \frac{1}{m} & \frac{1}{m} & \cdots & \frac{1}{m}-1 \end{bmatrix} \tag{3.6}$$

With the $\mathbf{T}$ matrices defined as in Equation 3.6, the vectors (of length $m-1$) $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ may take any value and yield vectors $\mathbf{a}$ and $\mathbf{c}$ with elements that sum to zero. As is the case in the analysis of variance, other contrast ($\mathbf{T}$) matrices may be used as well (see Thissen and Steinberg (1986) for examples); for reasons that will become clear, in this presentation we will use systems that identify the model with the constraints $a_1 = c_1 = 0$ instead of the original identification constraints.

Figure 3.1 shows four sets of trace lines that illustrate some of the range of variability of item response functions that can be obtained with the nominal

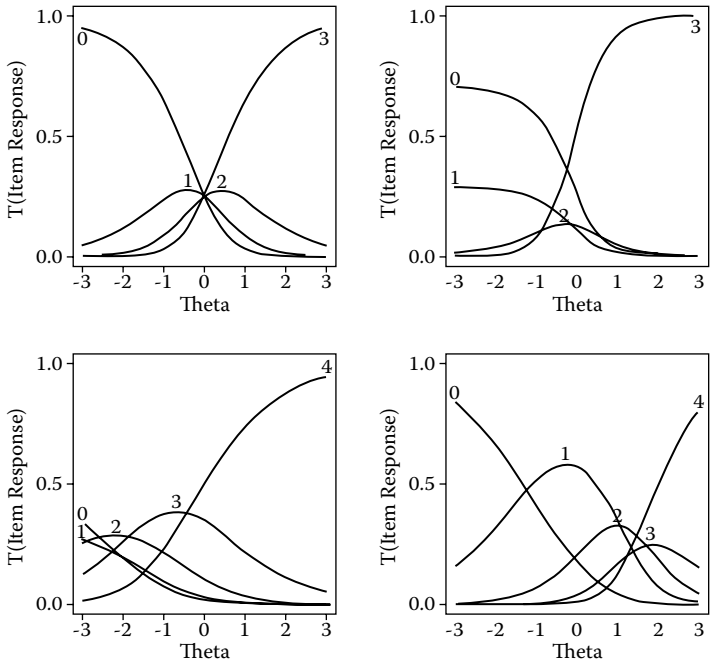

FIGURE 3.1 Upper left: Trace lines for an artificially constructed four-alternative item. Upper right: Trace lines for the "Identify" testlet described by Thissen and Steinberg (1988). Lower left: Trace lines for the number correct on questions following a passage on a reading comprehension test, using parameter estimates obtained by Thissen, Steinberg, and Mooney (1989). Lower right: Trace lines for judge-scored constructed-response item M075101 from the 1996 administration of the NAEP mathematics assessment.

model. The corresponding values of the parameter vectors a and c are shown in Table 3.1.

The curves in the upper left panel of Figure 3.1 artificially illustrate a maximally ordered, centered set of item responses: As seen in the leftmost two columns of Table 3.1 (for Item 1) the values of a_k increase by 1.0 as k increases; as we will see in a subsequent section, that produces an ordered variant of the nominal model. All of the values of c_k are identically 0.0, so the trace lines all cross at that value of θ. The upper right panel of Figure 3.1

TABLE 3.1 Original Nominal Model Parameter Values for the Trace Lines Shown in Figure 3.1

Response Category (k)	Item 1		Item 2		Item 3		Item 4	
	a	c	a	c	a	c	a	c
0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0
1	1.0	0.0	0.0	–0.9	0.2	0.5	0.95	1.2
2	2.0	0.0	1.1	–0.7	0.7	1.8	1.90	0.2
3	3.0	0.0	2.7	0.7	1.3	3.0	2.85	–1.4
4					2.2	3.3	3.80	–2.7

shows trace lines that correspond to parameter estimates (marked Item 2 in Table 3.1) obtained by Thissen and Steinberg (1988) (and subsequently by Hoskens and Boeck (1997); see Baker and Kim (2004) for the details of maximum marginal likelihood parameter estimation) for a testlet comprising two items from Bergan and Stone's (1985) data obtained with a test of preschool mathematics proficiency. The two items required the child to identify the numerals 3 and 4; the curves are marked 0 for neither identified, 1 for 3 identified but not 4, 2 for 4 identified but not 3, and 3 for both identified correctly. This is an example of a testlet with semiordered responses: The 0 and 1 curves are proportional because their a_k estimates are identical, indicating that, except for an overall difference in probability of endorsement, they have the same relation to proficiency: Both may be taken as incorrect. If a child can identify 4 but not 3 (the 2 curve), that indicates a moderate, possibly developing, degree of mathematical proficiency, and both correct (the 3 curve) increases as θ increases.

The lower left panel of Figure 3.1 shows trace lines that correspond to parameter estimates (marked Item 3 in Table 3.1) obtained by Thissen, Steinberg, and Mooney (1989) fitting the nominal model to the number-correct score for the questions following each of four passages on a reading comprehension test. Going from left to right, the model indicates that the responses are increasingly ordered for this number-correct scored testlet: Summed scores of 0 and 1 have nearly the same trace lines, because 0 (of 4) and 1 (of 4) are both scores that can be obtained with nearly equal probability by guessing on five-alternative multiple-choice items. After that, the trace lines look increasingly like those of a graded model. The lower right panel of Figure 3.1 is for a set of graded responses: It shows the curves that correspond to the parameter estimates for an extended constructed response mathematics item administered as part of the National Assessment of Educational Progress (NAEP) (Allen, Carlson, & Zelenak, 1999). The judged scores (from 0 to 4) were fitted with Muraki's (1992, 1997) generalized partial credit (GPC) model, which is a constrained version of the nominal model. In Table 3.1, the parameters for this item (Item 4 in the two rightmost columns) have been converted into values of a_k and c_k for comparability with the other items' parameters. The GPC model is an alternative to Samejima's (1969, 1997) graded model for such ordered responses; the two models generally yield very similar trace lines for the same data. In subsequent sections of this chapter we will discuss the relation between the GPC and nominal models in more detail.

Derivations of the Model

There are several lines of reasoning that lead to Equation 3.3 as an item response model. In this section we describe three kinds of theoretical argument that lead to the nominal model as the result, because they exist, and because different lines of reasoning appeal to persons with different backgrounds.

As Statistical Mechanics

Certainly the simplest development of the nominal model is essentially atheoretical, treating the problem as abstract statistical model creation. To do this, we specify only the most basic facts: that we have categorical item responses in several (>2) categories, that we believe those item responses depend on some latent variable (θ) that varies among respondents, and that the mutual dependence of the item responses on that latent variable explains their observed covariance. Then "simple" mathematical functions are used to complete the model.

First, we assume that the dependence of some response process (value) for each person, for each item response alternative, is a linear function of theta

$$z_k = a_k\theta + c_k \tag{3.7}$$

with unknown slope and intercept parameters a_k and c_k. Such a set of straight lines for a five-category item is shown in the left panel of Figure 3.2, using the parameters for Item 3 from Table 3.1.

To change those straight lines (z_k) into a model that yields probabilities (between 0 and 1) for each response, as functions of θ, we use the so-called multivariate logistic link function

$$\frac{\exp(z_k)}{\sum_i \exp(z_i)} \tag{3.8}$$

This function (Equation 3.8) is often used in statistical models to transform a linear model into a probability model for categorical data. It can be characterized as simple mathematical mechanics: Exponentiation of the values of z_k makes them all positive, and then division of each of those positive line values by the sum of all of them is guaranteed to transform the straight lines in the left panel of Figure 3.2 into curves such as those shown in the right panel of Figure 3.2. The curves are all between 0 and 1, and sum to 1

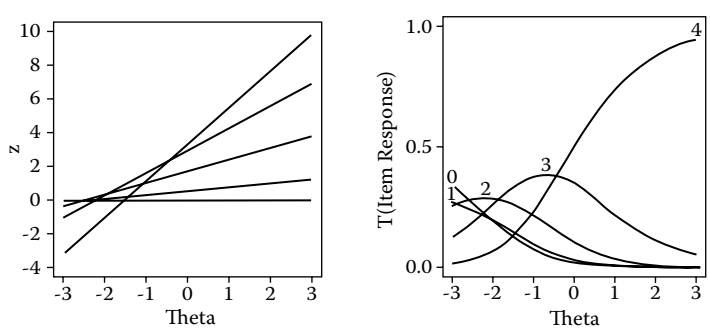

FIGURE 3.2 Left panel: Linear regressions of the response process z_k on θ for five response alternatives. Right panel: Multivariate logistic transformed curves corresponding to the five lines in the left panel.

at all values of θ, as required. (The curves in the right panel of Figure 3.2 are those from the lower left panel of Figure 3.1. de Ayala (1992) has presented a similar graphic as his Figure 1.)

For purely statistically trained analysts, with no background in psychological theory development, this is a sufficient line of reasoning to use the nominal model for data analysis. Researchers trained in psychology may desire a more elaborated theoretical rationale, of which two are offered in the two subsequent sections.

However, it is of interest to note at this point that the development in this section, specifically Equation 3.7, invites the questions: Why linear? Why not some higher-order polynomial, like quadratic? Indeed, quadratic functions of θ have been suggested or used for special purposes as variants of the nominal model: Upon hearing a description of the multiple-choice model (Thissen & Steinberg, 1984) D. B. Rubin (personal communication, December 15, 1982) suggested that an alternative to that model would be a nominal model with quadratic functions replacing Equation 3.7. Ramsay (1995) uses a quadratic term in Equation 3.7 for the correct response alternative for multiple-choice items when the multivariate logistic is used to provide "smooth" information curves for the nonparametric trace lines in the TestGraf system. Sympson (1983) also suggested the use of quadratic, and even higher-order, polynomials in a more complex model that never came into implementation or usage.

Nevertheless, setting aside multiple-choice items, for most uses of the nominal model the linear functions in Equation 3.7 are sufficient.

Relations With Thurstone Models

Relationship to Other Models: The term *Thurstone models* in polytomous IRT typically refers to models where response category thresholds characterize all responses above versus below a given threshold. In contrast, Rasch type models only characterize responses in adjacent categories. However, the Thurstone case V model, which is related to the development of the nominal categories model, is a very different type of Thurstone model–one without thresholds–highlighting the nominal categories model's unique place among polytomous IRT models.

The original development of the nominal categories model by Bock (1972) was based on an extension of Thurstone's (1927) case V model for binary choices, generalized to become a model for the first choice among three or more alternatives. Thurstone's model for choice made use of the concept of a response process that followed a normal distribution, one value (*process* in Thurstone's language) for each object. The idea was that the object or alternative selected was that with the larger value. In practice, a "comparatal" process is computed as the difference between the two response processes, and the first object is selected if the value of the comparatal process is greater than zero.

Bock and Jones (1968) describe many variants and extensions of Thurstone's models for choice, including generalizations to the first choice from among several objects. The obvious generalization of Thurstone's binary

choice model to create a model for the first choice from among three or more objects would use a multivariate normal distribution of $m-1$ comparatal processes for object or alternative j, each representing a comparison of object j with one of the others of m objects. Then the probability of selection of alternative j would be computed as a multiple integral over that $(m-1)$-dimensional normal density, computing a value known as an orthant probability. However, multivariate normal orthant probabilities are notoriously difficult to compute, even for simplified special cases. Bock and Jones suggest substitution of the multivariate logistic distribution, showing that the bivariate logistic yields probabilities similar to those obtained from a bivariate normal (these would be used for the first choice of three objects). The substitution of the logistic here is analogous with the substitution of the logistic function for the normal ogive in the two-parameter logistic IRT model (Birnbaum, 1968). Of course, the multivariate logistic distribution function is Equation 3.1.

In the appendix to this chapter, Bock provides an updated and detailed description of the theoretical development of the nominal categories model as an approximation to the multivariate generalization of Thurstone's model for choice. In addition, the appendix describes the development of the model that is obtained by considering first choices among three or more objects as an "extreme value" problem, citing the extension of Dubey's (1969) derivation of the logistic distribution to the multivariate case that has been used and studied by Bock (1970), McFadden (1974), and Malik and Abraham (1973). This latter development also ties the nominal categories model to the so-called Bradley-Terry-Luce (BTL) model for choice (Bradley & Terry, 1952; Luce & Suppes, 1965).

Thus, from the point of view of mathematical models for choice, the nominal categories model is both an approximation to Thurstone (normal) models for the choice of one of three or more alternatives, and the multivariate version of the BTL model.

The Probability of a Response in One of Two Categories

Another derivation of the nominal model involves its implications for the conditional probability of a response in one category (say k) given that the response is in one of two categories (k or k'). This derivation is analogous in some respects to the development of Samejima's (1969, 1997) graded model, which is built up from the idea that several conventional binary item response models may be concatenated to construct a model for multiple responses. In the case of the graded model, accumulation is used to transform the multiple category model into a series of dichotomous models: The conventional normal ogive or logistic model is used to describe the probability that a response is in category k or higher, and then those cumulative models are subtracted to produce the model for the probability the response is in a particular category. This development of the graded model rests, in turn, on the theoretical development of the normal ogive model as a model for the psychological response process, as articulated by Lord and Novick (1968, pp. 370–373), and then on Birnbaum's (1968) reiteration for test theory of Berkson's (1944, 1953) suggestion that the logistic function could usefully be substituted for

the normal ogive. (See Thissen and Orlando (2001, pp. 84–89) for a summary of the argument by Lord and Novick and the story behind the logistic substitution.)

The nominal model may be derived in a parallel fashion, assuming that the *conditional* probability of a response in one category (say k), given that the response is in one of two categories (k or k'), can be modeled with the two-parameter logistic (2PL). The algebra for this derivation "frontwards" (from the 2PL for the conditional responses to the nominal model for all of the responses) is algebraically challenging as test theory goes, but it is sufficient to do it "backwards," and that is what is presented here. (We note in passing that Masters (1982) did this derivation frontwards for the simpler route from the Rasch or one-parameter logistic (1PL) to the partial credit model.)

If one begins with the nominal model as stated in Equation 3.3, and writes the conditional probability for a response in category k given that the response is in one of categories k or k',

$$T(k|k,k') = \frac{T(k)}{T(k') + T(k)} \tag{3.9}$$

then only a modest amount of algebra (cancel the identical denominators, and then more cancellation to change the three exponential terms into one) is required to show that this conditional probability is, in fact, a two-parameter logistic function:

$$T(k|k,k',) = \frac{1}{\left[1 + \exp\left(-a_k^c \theta + c_k^c\right)\right]} \tag{3.10}$$

with

$$c_k^c = c_{k'} - c_k \tag{3.11}$$

and

$$a_k^c = a_k - a_{k'} \tag{3.12}$$

Placing interpretation on the algebra, what this means is that the nominal model assumes that if we selected the subsample of respondents who selected either alternative k or k', setting aside respondents who made other choices, and analyzed the resulting dichotomous item in that subset of the data, we would use the 2PL model for the probability of response k in that subset of the data. This choice, like the choice of the normal ogive or logistic model for the cumulative probabilities in the graded model, then rests on

the theoretical development of the normal ogive model as a psychological response process model as articulated by Lord and Novick (1968), and Birnbaum's (1968) argument for the substitution of the logistic. The difference between the two ways of dividing multiple responses into a series of dichotomies (cumulative vs. conditional) has been discussed by Agresti (2002).

An interesting and important feature of the nominal model is obtained by specializing the conditional probability for any pair of responses to adjacent response categories (k or k – 1; *adjacent* is meaningful if the responses are actually ordered); the same two-parameter logistic is obtained:

$$T(k|k,k-1)=\frac{1}{\left[1+\exp\left(-a_k^c\theta+c_k^c\right)\right]} \tag{3.13}$$

with

$$c_k^c=c_{(k-1)}-c_k \tag{3.14}$$

and

$$a_k^c=a_k-a_{(k-1)} \tag{3.15}$$

It is worth noting at this point that the threshold b_k^c for the slope-threshold form of the conditional 2PL curve,

$$T(k|k,k-1)=\frac{1}{\left[1+\exp\left(-a_k^c\left[\theta-b_k^c\right]\right)\right]} \tag{3.16}$$

is

$$b_k^c=\frac{c_k^c}{a_k^c}=\frac{c_{k-1}-c_k}{a_k-a_{k-1}} \tag{3.17}$$

which is also the crossing point of the trace lines for categories k and $k-1$ (de Ayala, 1993; Bock, 1997). These values are featured in some parameterizations of the nominal model for ordered data.

This fact defines the concept of order for nominal response categories: Response k is "higher" than response $k-1$ if and only if $a_k > a_{k-1}$, which means that a^c is positive, and so the conditional probability of selecting response k (given that it is one of the two) increases as θ increases. Basically this means that item analysis with the nominal model *tells the data analyst the order of the item responses*. We have already made use of this fact in discussion of order and the a_k parameters in Figure 3.1 and Table 3.1 in the introductory section.

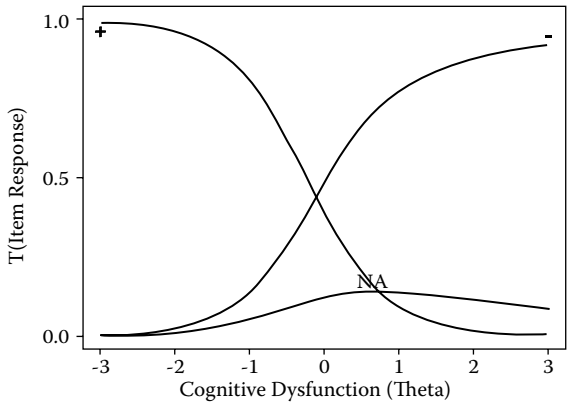

FIGURE 3.3 Trace lines corresponding to item parameters obtained by Huber (1993) in his analysis of the item "Count down from 20 by 3s" on the Short Portable Mental Status Questionnaire (SPMSQ).

Two additional examples serve to illustrate the use of the nominal model to determine the order of response categories, and the way the model may be used to provide trace lines that can be used to compute IRT scale scores (see Thissen, Nelson, Rosa, and McLeod, 2001) using items with purely nominal response alternatives.

Figure 3.3 shows the trace lines corresponding to item parameters obtained by Huber (1993) in his analysis of the item "Count down from 20 by 3s" on the Short Portable Mental Status Questionnaire (SPMSQ), a brief diagnostic instrument used to detect dementia. For this item, administered to a sample of aging individuals, three response categories were recorded: correct, incorrect (scored positively for this "cognitive dysfunction" scale), and refusal (NA). Common practice scoring the SPMSQ in clinical and research applications was to score NA as incorrect, based on a belief that respondents who refused to attempt the task probably could not do it. Huber fitted the three response categories with the nominal model and obtained the parameters $\mathbf{a}' = [0.0, 1.56, 1.92]$ and $\mathbf{c}' = [0.0, -0.52, 0.85]$; the corresponding curves are shown in Figure 3.3. As expected, the a_k parameter for NA is much closer to the a_k parameter for the incorrect response, and the curve for NA is nearly proportional to the − curve in Figure 3.3. This analysis lends a degree of justification to the practice of scoring NA as incorrect. However, if the IRT model is used to compute scale scores, those scale scores reflect the *relative* evidence of failure provided by the NA response more precisely.

The SPMSQ also includes items that many item analysts would expect to be locally dependent. One example involves a pair of questions that require the respondent to state his or her age, and then his or her date of birth. Huber (1993) combined those two items into a testlet with four response categories: both correct (++), age correct and date of birth incorrect (+−), age incorrect and date of birth correct (−+), and both incorrect (−−). Figure 3.4 shows the

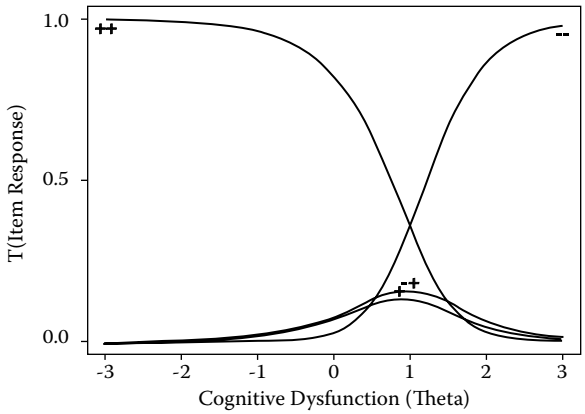

FIGURE 3.4 Nominal model trace lines for the four response categories for Huber's (1993) SPMSQ_ testlet scored as reporting both age and date of birth correctly (++), age correctly and date of birth incorrectly (+–), age incorrectly and date of birth correctly (–+), and both incorrectly (––).

nominal model trace lines for the four response categories for that testlet. While one may confidently expect that the –– response reflects the highest degree of dysfunction and the ++ response the lowest degree of dysfunction, there is a real question about the scoring value of the +– and –+ responses. The nominal model analysis indicates that the trace lines for +– and –+ are almost exactly the same, intermediate between good and poor performance. Thus, after the analysis with the nominal model one may conclude that this testlet yields four response categories that collapse into three ordered scoring categories: ++, [+– or –+], and ––.

Alternative Parameterizations, With Uses

Thissen and Steinberg (1986) showed that a number of other item response models may be obtained as versions of the nominal model by imposing constraints on the nominal model's parameters, and further that the canonical parameters of those other models may be made the αs and γs estimated for the nominal model with appropriate choices of $\mathbf{T}$ matrices. Among those other models are Masters' (1982) *partial credit* (PC) model (see also Masters and Wright, 1997) and Andrich's (1978) *rating scale* (RS) model (see also Andersen (1997) for relations with proposals by Rasch (1961) and Andersen (1977)). Thissen and Steinberg (1986) also mentioned in passing that a version of the nominal model like the PC model, but with discrimination parameters that vary over items, is also within the parameter space of the nominal model. That latter model was independently developed and used in the 1980s by Muraki (1992) and called the *generalized partial credit* (GPC) model, and by Yen (1993) and called the *two-parameter partial credit* (2PPC) model.

More on Ordered Versions of the Nominal Model—Rating Scale and (Generalized) Partial Credit Models

Notational Difference: Remember this model was presented slightly differently in Chapter 2:

$$P_{ik}(\theta_j) = \frac{\exp\left[\sum_{v=0}^{k} a_i(\theta_j - b_{iv})\right]}{\sum_{h=0}^{K-1} \exp\left[\sum_{v=0}^{h} a_i(\theta_j - b_{iv})\right]}$$

Muraki (1992, 1997) has used several parameterizations to describe the GPC model, among them

$$T(k) = \frac{\exp \sum_{j=0}^{k} 1.7a(\theta - b + d_j)}{\sum_{i=0}^{m-1} \exp \sum_{j=0}^{i} 1.7a(\theta - b + d_j)} \tag{3.18}$$

with the constraint that

$$\sum_{i=1}^{m-1} d_i = 0 \tag{3.19}$$

and alternatively

$$T(k) = \frac{\exp[1.7a[T_k(\theta - b) + K_k]]}{\sum_{i=0}^{m-1} \exp[1.7a[T_k(\theta - b) + K_k]]} \tag{3.20}$$

in which

$$K_k = \sum_{i=1}^{k} d_i \tag{3.21}$$

Muraki's parameterization of the GPC model is closely related to Masters' (1982) specification of the PC model:

Notational Difference: Here the authors use θ to refer to the latent variable of interest where Masters (see Equations 5.22 and 5.23 in Chapter 5) and Andrich (see Equations 6.24 and 6.25 in Chapter 6) typically refer to the latent variable using β. This θ/β notational difference will be seen in other chapters and is common in IRT literature.

$$T(k) = \frac{\exp \sum_{j=0}^{k} (\theta - \delta_j)}{\sum_{i=0}^{m-1} \exp \sum_{j=0}^{i} (\theta - \delta_j)} \tag{3.22}$$

with the constraint

$$\sum_{j=0}^{0}(\theta-\delta_j)=0 \tag{3.23}$$

Andrich's (1978) RS model is

$$T(k)=\frac{\exp\sum_{j=0}^{k}(\theta-(\delta+\tau_k))}{\sum_{i=0}^{m-1}\exp\sum_{j=0}^{i}(\theta-(\delta+\tau_j))} \tag{3.24}$$

with the constraints

$$\sum_{j=0}^{0}[\theta-(\delta+\tau_j)]=0 \tag{3.25}$$

and

$$\sum_{j=1}^{m-1}\tau_j=0 \tag{3.26}$$

Thissen and Steinberg (1986)

Thissen and Steinberg (1986) described the use of alternative **T** matrices in the formulation of the nominal model. For example, when formulated for marginal estimation following Thissen (1982), Masters' (1982) PC model and Andrich's (1978) RS model use a single slope parameter that is the coefficient for a linear basis function:

$$\underset{m\times 1}{\mathbf{T}_{\mathbf{a(PC)}}}=\begin{bmatrix}0\\1\\2\\\vdots\\m-1\end{bmatrix} \tag{3.27}$$

Masters' (1982) PC model used a parameterization for the threshold parameters that can be duplicated, up to proportionality, with this **T** matrix for the *c*s:

$$\underset{m\times(m-1)}{\mathbf{T}_{\mathbf{c(PC)}}}=\begin{bmatrix}0&0&\cdots&0\\-1&0&\ldots&0\\-1&-1&\ldots&0\\\vdots&\vdots&&\vdots\\-1&-1&\cdots&-1\end{bmatrix} \tag{3.28}$$

Terminology Note: The authors use the term *threshold* here, whereas in other chapters these parameters are sometimes referred to as *step* or *boundary* parameters.

Andrich's RS model separated an overall item location parameter from a set of parameters describing the category boundaries for the item response scale; the latter were constrained equal across items, and may be obtained, again up to proportionality, with

$$\underset{m\times(m-1)}{\mathbf{T}_{\mathbf{c(RS\text{-}C)}}} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ -1 & -1 & 0 & \cdots & 0 \\ -2 & -1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -(m-2) & -1 & -1 & \cdots & -1 \\ -(m-1) & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{3.29}$$

Andrich (1978, 1985) and Thissen and Steinberg (1986) described the use of a polynomial basis for the cs as an alternative to $\mathbf{T}_{\mathbf{c(RS\text{-}C)}}$ that "smooths" the category boundaries; the overall item location parameter is the coefficient of the first (linear) column, and the coefficients associated with the other columns describe the response category boundaries:

$$\underset{m\times(m-1)}{\mathbf{T}_{\mathbf{c(RS\text{-}P)}}} = \begin{bmatrix} 0 & 0^2 & 0^3 & \cdots \\ -1 & -1^2 & -1^3 & \cdots \\ -2 & -2^2 & -2^3 & \cdots \\ \vdots & \vdots & \vdots & \\ -(m-2) & -(m-2)^2 & -(m-2)^3 & \cdots \\ -(m-1) & -(m-1)^2 & -(m-1)^3 & \cdots \end{bmatrix} \tag{3.30}$$

Polynomial contrasts were used by Thissen et al. (1989) to obtain the trace lines for summed score testlets for a passage-based reading comprehension test; the trace lines for one of those testlets are shown as the lower left panel of Figure 3.1 and the right panel of Figure 3.2. The polynomial contrast set included only the linear term for the a_ks and the linear and quadratic terms for the c_ks for that testlet; that was found to be a sufficient number of terms to fit the data. This example illustrates the fact that, although the nominal model may appear to have many estimated parameters, in many situations a reduction of rank of the $\mathbf{T}$ matrix may result in much more efficient estimation.

A New Parameterization for the Nominal Model

After three decades of experience with the nominal model and its applications, a revision to the parameterization of the model would serve several purposes: Such a revision could be used first of all to facilitate the extension of the nominal model to become a multidimensional IRT (MIRT) model, a first for purely nominal responses. In addition, a revision could make the model easier to explain. Further, by retaining features that have actually been used in data analysis, and discarding suggestions (such as many alternative $\mathbf{T}$ matrices) that have rarely or never been used in practice, the implementation of estimation algorithms for the model in software could become more straightforward.

Thus, while the previous sections of this chapter have described the nominal model as it has been, and as it has been used, this section presents a new parameterization that we expect will be implemented in the next generation of software for IRT parameter estimation. This is a look into the future.

Desiderata

The development of the new parameterization for the nominal model was guided by several goals, combining a new insight with experience gained over the last 30 years of applications of the model:

1. The dominating insight is that a kind of multidimensional nominal model can be created by separating the a parameterization into a single overall (mutliplicative) slope or discrimination parameter, that is then expanded into vector form to correspond to vector θ, and a set of $m-2$ contrasts among the a parameters that represent what Muraki (1992) calls the scoring functions for the responses. This change has the added benefit that, for the first time, the newly reparameterized nominal model has a single discrimination parameter comparable to those of other IRT models. That eases explanation of results of item analysis with the model.
2. In the process of accomplishing Goal 1, it is desirable to parameterize the model in such a way that the scoring function may be (smoothly) made linear $(0,1,2,\cdots,m-1)$ so that the multiplicative overall slope parameter becomes the slope parameter for the GPC model, which, constrained equally across items, also yields the PC and RS models. In addition, with this scoring function the overall slope parameter may meaningfully be set equal to the (also equal) slope for a set of 2PL items to mimic Rasch family mixed models.
3. We have also found it useful at times in the past 20 years to use models between the highly constrained GPC model and the full-rank nominal model, as suggested by Thissen and Steinberg (1986), most often by using polynomial bases for the a and c parameters and reducing the number of estimated coefficients below full rank to obtain "smoothly changing" values of the a and c parameters across response categories. It is desirable to retain that option.

4. With other sets of data, we have found it useful to set equal subsets of the a or c parameters within an item, modeling distinct response categories as equivalent for scoring (the a parameters are equal) or altogether equivalent (both the a and c parameters are equal).

Obtaining Goals 3 and 4 requires two distinct parameterizations, both expressed as sets of **T** matrices; Goals 1 and 2 are maintained in both parameterizations.

The New Parameterization

The new parameterization is

$$T\left(u = k|\theta; a_i^*, \mathbf{a}_\mathbf{k}^\mathbf{s}, \mathbf{c}_\mathbf{k}\right) = T(k) = \frac{\exp(z_k)}{\sum_i \exp(z_i)} \tag{3.31}$$

in which

$$z_{ik} = a_i^* a_{k+1}^s \theta + c_{k+1} \tag{3.32}$$

and a^* is the overall slope parameter, a_{k+1}^s is the scoring function for response k, and c_{k+1} is the intercept parameter as in the original model. The equating following restrictions for identification,

$$a_1^s = 0, \quad a_m^s = m-1, \quad \text{and} \quad c_1 = 0 \tag{3.33}$$

are implemented by reparameterizing, and estimating the parameters α and γ:

$$\mathbf{a}^\mathbf{s} = \mathbf{T}\boldsymbol{\alpha} \quad \text{and} \quad \mathbf{c} = \mathbf{T}\boldsymbol{\gamma} \tag{3.34}$$

The Fourier Version for Linear Effects and Smoothing

To accomplish Goals 1 to 3, we use a Fourier basis as the **T** matrix, augmented with a linear column:

$$\underset{m\times(m-1)}{\mathbf{T}_\mathbf{F}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & f_{22} & \cdots & f_{2(m-1)} \\ 2 & f_{32} & \cdots & f_{3(m-1)} \\ \vdots & \vdots & & \vdots \\ m-1 & 0 & \cdots & 0 \end{bmatrix} \tag{3.35}$$

in which f_{ki} is

$$f_{ki} = \sin[\pi(i-1)(k-1)/(m-1)] \tag{3.36}$$

and $\alpha_1 = 1$. Figure 3.5 shows graphs of the linear and Fourier functions for four categories (left panel) and six categories (right panel). The Fourier-based terms functionally replace quadratic and higher-order polynomial terms that

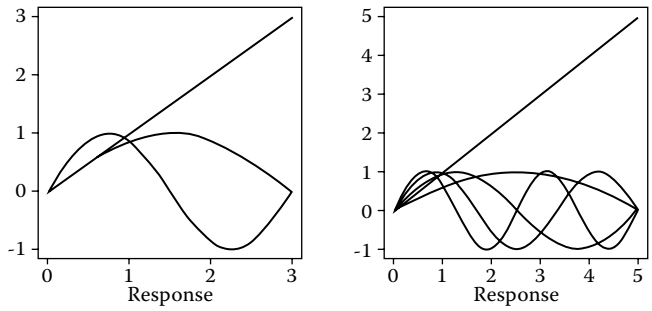

FIGURE 3.5 Graphs of the linear and Fourier basis functions for the new nominal model parameterization, for four categories (left panel) and six categories (right panel); the values of $\boldsymbol{T}$ at integral values on the *Response* axis are the elements of the $\boldsymbol{T}$ matrix of Equations 3.35 and 3.36.

we have often used to smooth sequences of a_k and c_k parameters with a more numerically stable, symmetrical orthogonal basis.

The new parameterization, using the Fourier $\mathbf{T}$ matrix, provides several useful variants of the nominal model: When $a^*, \{\alpha_2, \ldots, \alpha_{m-1}\}$, and γ are estimated parameters, this is the full-rank nominal model. If $\{\alpha_2, \ldots, \alpha_{m-1}\}$ are restricted to be equal to zero, this is a reparameterized version of the GPC model. The Fourier basis provides a way to create models between the GPC and nominal model, as were used by Thissen et al. (1989), Wainer, Thissen, and Sireci (1991), and others.

Useful Derived Parameters

When the linear-Fourier basis $\mathbf{T_F}$ is used for both

$$\mathbf{a}^s = \mathbf{T_F}\boldsymbol{\alpha} \quad \text{and} \quad \mathbf{c} = \mathbf{T_F}\boldsymbol{\gamma} \tag{3.37}$$

with $\alpha_1 = 1$ and $\alpha_2, \ldots, \alpha_{m-1} = 0$, then the parameters of the GPC model

$$T^i(k) = \frac{\exp \sum_{j=0}^{k} 1.7a(\theta - b + d_j)}{\sum_{i=0}^{m-1} \exp \sum_{j=0}^{i} 1.7a(\theta - b + d_j)} \tag{3.38}$$

may be computed as

$$a = \frac{a_i^*}{1.7} \tag{3.39}$$

$$b = \frac{-c_m}{a_i^*(m-1)} = \frac{-\gamma_1}{a_i^*} \tag{3.40}$$

and

$$d_k = \frac{c_k - c_{k-1}}{a_i^*} - \frac{c_m}{m-1} \tag{3.41}$$

for $k = 2,\ldots,m-1$ (noting that $d_1 = 0$ and $c_1 = 0$ as constraints for identification). (Childs and Chen (1999) provided formulae to convert the parameters of the original nominal model into those of the GPC model, but they used the **T** matrices in the computations, which is not essential in the simpler methods given here.)

Also note that if it desired to constrain the GPC parameters d_k to be equal across a set of items, that is accomplished by setting the parameter sets $\gamma_2,\ldots,\gamma_{m-1}$ equal across those items. This kind of equality constraint really only makes sense if the overall slope parameter a^* is also set equal across those items, in which case $b_i = \frac{-\gamma_{i,1}}{a^*}$ reflects the overall difference in difficulty, which still varies over items i. (Another way to put this is that the linear-Fourier basis separates the parameter space into a (first) component for $b_i = \frac{-\gamma_{i,1}}{a_i^*}$ and a remainder that parameterizes the "spacing" among the thresholds or crossover points of the curves.)

The alternative parameterization of the GPC

$$T(k) = \frac{\exp[1.7a[T_k(\theta - b) + K_k]}{\sum_{i=0}^{m-1} \exp[1.7a[T_i(\theta - b) + K_i]} \tag{3.42}$$

in which

$$K_k = \sum_{i=1}^{k} d_i \tag{3.43}$$

simply substitutes K_k parameters that may be computed from the values of d_i. Note that the multiplication of the parameter b by the scoring function T_k provides another explanation of the fact that with the linear-Fourier basis $b_i = \frac{-\gamma_{i,1}}{a^*}$.

To provide translations of the parameters for Rasch family models, some accommodation must be made between the conventions that the scale of the latent variable is usually set for more general models by specifying the θ is distributed with mean zero and variance one, versus many implementations of Rasch family models with the specification that some item's difficulty is zero, or the average difficulty is zero, and the slope is one, leaving the mean and variance of the θ distribution unspecified, and estimated.

If we follow the approach taken by Thissen (1982) that a version of Rasch family models may be obtained with the specification that θ is distributed with mean zero and variance one, estimating a single common slope parameter (a^* in this case) for all items, and all items' difficulty parameters, then the $\boldsymbol{\delta}$ parameters of Masters' PC model are

$$\delta_k = b - d_k \tag{3.44}$$

(in terms of the parameters of Muraki's GPC model) up to a linear transformation of scale, and the δ and τ parameters of Andrich's RS model are

$$\delta = b \tag{3.45}$$

and

$$\tau_k = -d_k \tag{3.46}$$

again up to a linear transformation of scale.

The Identity-Based **T** *Matrix for Equality Constraints*

To accomplish Goals 1, 2, and 4, involving equality constraints, we use **T** matrices for $\mathbf{a}^s$ as of the form

$$\underset{m\times(m-1)}{\mathbf{T}_{\mathbf{Ia}}} = \begin{bmatrix} 0 & \underline{\mathbf{0}}'_{m-2} \\ \underline{\mathbf{0}}_{m-2} & \mathbf{I}_{m-2} \\ m-1 & \underline{\mathbf{0}}'_{m-2} \end{bmatrix} \tag{3.47}$$

with the constraint that $\alpha_1 = 1$. If it is desirable to impose equality constraints in addition on the c_s, we use the following **T** matrix:

$$\underset{m\times(m-1)}{\mathbf{T}_{\mathbf{Ic}}} = \begin{bmatrix} \underline{\mathbf{0}}'_{m-1} \\ \mathbf{I}_{m-1} \end{bmatrix} \tag{3.48}$$

This arrangement provides for the following variants of the nominal model, among others: When a^*, $\{\alpha_2,...,\alpha_{m-1}\}$, and γ are estimated parameters, this is again the full-rank nominal model. If $\alpha_i = i$ for $\{\alpha_2,...,\alpha_{m-1}\}$, this is a reparameterized version of the generalized partial credit model.

The restriction $a_1^s = a_2^s$ is imposed by setting $\alpha_2 = 0$. The restriction $a_{m-1}^s = a_m^s$ is imposed by setting $\alpha_{(m-1)} = m-1$. For the other values of a^s the restriction $a_{k'}^s = a_k^s$ is imposed by setting $\alpha_{k'} = \alpha_k$.

Illustrations

Table 3.2 shows the values of the new nominal model parameters for the items with trace lines in Figure 3.1 and the original parameters in Table 3.1. Note that the scoring parameters in $\mathbf{a}^s$ for Items 1 and 4 are $[0,1,2,...,m-1]$, indicating that the nominal model for those two items is one for strictly ordered responses. In addition, we observe that the lower discrimination of Item 3 (with trace lines shown in the lower left panel of Figure 3.1) is now clearly indicated by the relatively lower value of $a*$; the discrimination

TABLE 3.2 Item Parameters for the New Parameterization of the Nominal Model, for the Same Items With the Original Model Parameters in Table 3.1

Parameter		Item 1		Item 2		Item 3		Item 4	
a*		1.0		0.9		0.55		0.95	
a_1^s	c_1	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0
a_2^s	c_2	1.0	0.0	0.0	–0.9	0.36	0.5	1.00	1.2
a_3^s	c_3	2.0	0.0	1.2	–0.7	1.27	1.8	2.00	0.2
a_4^s	c_4	3.0	0.0	3.0	0.7	2.36	3.0	3.00	–1.4
a_5^s	c_5					4.00	3.3	4.00	–2.7

parameter for Item 3 is only 0.55, relative to values between 0.9 and 1.0 for the other three items. The values of the c parameters are unchanged from Table 3.1. If the item analyst wishes to convert the parameters for Item 3 in Table 3.2 to those previously used for the GPC model, Equations 3.3 to 3.41 may be used.

Multidimensionality and the Nominal Model

The new parameterization of the nominal model is designed to facilitate multidimensional item factor analysis (or MIRT analysis) for items with nominal responses, something that has not heretofore been available (Cai, Bock, & Thissen, in preparation). A MIRT model has a vector-valued $\boldsymbol{\theta}$—two or more dimensions in the latent variable space that are used to explain the covariation among the item responses. Making use of the separation of the new nominal model parameterization of overall item discrimination parameter (a^*) from the scoring functions (in $\mathbf{a}^s$), the multidimensional nominal model has a vector of discrimination parameters a^*, one value indicating the slope in each direction of the $\boldsymbol{\theta}$-space. This vector of discrimination parameters taken together indicates the direction of highest discrimination of the item, which may be along any of the θ axes or between them.

The parameters in $\mathbf{a}^s$ remain unchanged: Those represent the scoring functions of the response categories and are assumed to be the same in all directions in the $\boldsymbol{\theta}$-space. So the model remains nominal in the sense that the scoring functions may be estimated from the data. The intercept parameter c also remains unchanged, taking the place of the standard unitary intercept parameter in a MIRT model.

Assembled in notation, the nominal MIRT model is

$$T(u = k|\boldsymbol{\theta}; \mathbf{a}^*, \mathbf{a}^s, \mathbf{c}) = T(k) = \frac{\exp(z_k)}{\sum_i \exp(z_i)} \tag{3.49}$$

modified from Equation 3.3 with vector $\mathbf{a}^*$ and vector $\boldsymbol{\theta}$, in which

$$z_k = \mathbf{a}'^* a_k^s \boldsymbol{\theta} + c_k \tag{3.50}$$

This is a nominal response model in the sense that, for any direction in the $\boldsymbol{\theta}$ space, a cross section of the trace surfaces may take the variety of shapes provided by the unidimensional nominal model. Software to estimate the parameters of this model is currently under development. When completed this model will permit the empirical determination of response alternative order in the context of multidimensional $\boldsymbol{\theta}$. If an ordered version of the model is used, with scoring functions $[0,1,2,...,m-1]$, this model is equivalent to the multidimensional partial credit model described by Yao and Schwarz (2006).

Conclusion

Reasonable questions may be raised about why the new parameterization of the nominal model has been designed as described in the preceding section; we try to answer some of the more obvious of those questions here:

Why is the linear term of the $\mathbf{T}$ *matrix scaled between zero and* $m-1$*, as opposed to some other norming convention?* It is planned that the implementation of estimation for this new version of the nominal model will be in general purpose computer software that, among other features, can "mix models," for example, for binary and multiple-category models. We also assume that the software can fix parameters to any specified value, or set equal any subset of the parameters. Some users may want to use Rasch family (Masters and Wright, 1984) models, mixing the original Rasch (1960) model for the dichotomous items and the PC or RS models for the polytomous items. To accomplish a close approximation of that in a marginal maximum likelihood estimation system, with a $N(0,1)$ population distribution setting scale for the latent variable, a common slope (equal across items) must be specified for all items (Thissen, 1982). For the dichotomous items that scope parameter is for the items scored 0,1; for the polytomous items it is for item scores $0,1,\ldots,(m-1)$. Thus, scaling the linear component of the scoring function with unit steps facilitates the imposition of the equality constraints needed for mixed Rasch family analysis. It also permits meaningful equality constraints between discrimination parameters for different item response models that are not in the Rasch family.

In the MIRT version of the model, the $\mathbf{a}^*$ parameters may be rescaled after estimation is complete, to obtain values that have the properties of factor loadings, much as has been done for some time for the dichotomous model in the software TESTFACT (du Toit, 2003).

Why does the user need to prespecify both the lowest and highest response category (to set up the $\mathbf{T}$ *matrix) for a nominal model?* This is not as onerous as it may first appear: When fitting the full-rank nominal model, one does not have to correctly specify highest and lowest response categories. If the data indicate another order, estimated values of a_k^s may be less than zero or exceed $m-1$, indicating the empirical scoring order. It is only necessary that the

item analyst prespecify two categories that are *differently* related to θ, such that one is relatively lower and the other relatively higher—but even which one is which may be incorrect, and that will appear as a negative value of a_i^*. Presumably, when fitting a restricted (ordered) version of the model, the user would have already fitted the unrestricted nominal model to determine or check the empirical order of the response categories, or the user would have confidence from some other source of information about the order.

Why not parameterize the model in slope-threshold form, instead of slope-intercept form? Aren't threshold parameters easier to interpret in IRT? While we fully understand the attraction, in terms of interpretability, for threshold-style parameters in IRT models, there are several good reasons to parameterize with intercepts for estimation. The first (oldest historically) reason is that the slope-intercept parameterization is a much more numerically stable arrangement for estimating the parameters of logistic models, due to a closer approximation of the likelihood to normality and less error correlation among the parameters. A second reason is that the threshold parameterization does not generalize to the multidimensional case in any event; there is no way in a MIRT model to "split" the threshold among dimensions, rendering a threshold parameterization more or less meaningless. We note here that, for models for which it makes sense, we can always convert the intercept parameters into the corresponding item location and threshold values for reporting, and in preceding sections we have given formulas for doing so for the GPC model.

Why not use polynomial contrasts to obtain intermediate models, as proposed by Thissen and Steinberg (1986) and implemented in MULTILOG (du Toit, 2003), instead of the Fourier basis? An equally compelling question is to ask: *Why polynomials?* The purpose of either basis is to provide smooth trends in the a_s or c_s across a set of response categories. Theory is not sufficient at this time to specify a particular mathematic formulation for smoothness across categories in the nominal model. The Fourier basis accomplishes that goal as well as polynomials, and is naturally orthogonal, which (slightly) simplifies the implementation of the estimation algorithm.

In this chapter we have reviewed the development of Bock's (1972) nominal model, described its relation with other commonly used item response models, illustrated some of its unique uses, and provided a revised parameterization for the model that we expect will render it more useful for future applications in item analysis and test scoring. As IRT has come to be used in more varying contexts, expanding its domain of application from its origins in educational measurement into social and personality psychology, and the measurement of health outcomes and quality of life, the need to provide item analysis for items with polytomous responses with unknown scoring order has increased. The reparameterized nominal model provides a useful response to that challenge. Combined with the development of multidimensional nominal item analysis (Cai et al., in preparation), the nominal model represents a powerful component among the methods of IRT.

Appendix 1: Background of the Nominal Categories Model

R. Darrell Bock

The first step in the direction of the nominal model was an extension of Thurstone's (1927) method of paired comparisons to first choices among three or more objects. The objects can be anything for which subjects could be expected to have preferences—opinions on public issues, competing consumer products, candidates in an election, and so on. The observations for a set of m objects consist of the number of subjects who prefer object j to object k and the number who prefer k to j. Any given subject does not necessarily have to respond to all pairs. Thurstone proposed a statistical model for choice in which differences in the locations of the objects on a hypothetical scale of preference value predict the observed proportions of choice in all $m(m-1)/2$ distinct pairs. He assumed that a subject's response to the task of choosing between the objects depended upon a subjective variable for, say, object j,

$$v_j = \mu_j + \varepsilon_j \tag{3.51}$$

where, in the population of respondents, ε_j is a random deviation distributed normally with mean 0 and variance σ^2. He called this variable a response process and assumed that the subject chooses the object with the larger process. Although the distribution of ν_j might have different standard deviations for each object and nonzero correlations between objects, this greatly complicates the estimation of differences between the means. Thurstone therefore turned his attention to the case V model in which the standard deviations were assumed equal and all correlations assumed zero in all comparisons. With this simplification, the so-called comparatal process

$$v_{jk} = v_j - v_k \tag{3.52}$$

has mean $\mu_j - \mu_k$ and standard deviation $\sqrt{2}\sigma$, and the comparatal processes v_{jk}, v_{jl} for object j have constant correlation ½. Thurstone's solution to the estimation problem was to convert the response proportions to normal deviates and estimate the location differences by unweighted least squares, which requires only m^2 additions and m divisions. With modern computing machinery, solutions with better properties (e.g., weighted least squares or maximum likelihood) are now accessible (see Bock & Jones, 1968, Section 6.4.1). From the estimated locations, the expected proportions for each comparison are given by the cumulative normal distribution function, $\Phi(y)$, at $y = (\mu_j - \mu_k)$. These proportions can be used in chi-square tests of the goodness of fit of the paired comparisons model (Bock, 1956) (Bock & Jones, 1968, section 6.7.1).

Extension to First Choices

The natural extension of the paired comparison case V solution to what might be called the "method of first choices," that is, the choice of one preferred object in a set of m objects is simply to assume the $m-1$ comparatal processes for object j,

$$v_{jk} = v_j - v_k, \quad k = 1, 2, \ldots, m; \quad k \neq j \tag{3.53}$$

is distributed $(m-1)$-variate normal with means $\mu_j - \mu_k$, constant variance $2\sigma^2$, and constant correlation ρ_{jk} equal to ½ (Bock, 1956; 1975, Section 8.1.3). Expected probabilities of first choice for a given object then correspond to the $(m-1)$-fold multiple integral of the $(m-1)$-variate normal density function in the orthant from minus infinity up to the limits equal to the comparatal means.

For general multivariate normal distributions of high dimensionality, evaluation of orthant probabilities is computationally challenging even with modern equipment. Computing formulae and tables exit for the bivariate case (National Bureau of Standards, 1956) and the trivariate case (Steck, 1958), but beyond that, Monte Carlo approximation of the positive orthant probabilities appears to be the only recourse at the present time. Fortunately, much simpler procedures based upon a multivariate logistic distribution are now available for estimating probabilities of first choice. By way of introduction, the following section gives essential results for the univariate and bivariate logistic distributions.

The Univariate Logistic Distribution

Applied to the case V paired comparisons model, the univariate logistic distribution function can be expressed either in terms of the comparatal process $z = u_{jk}$:

$$\Psi(z) = \frac{1}{1 + e^{-z}} \tag{3.54}$$

or in terms of the separate processes $z_1 = v_j$ and $z_2 = v_k$:

$$\Psi(z_1) = \frac{e^{z_1}}{e^{z_1} + e^{z_2}} \tag{3.55}$$

under the constraint $z_1 + z_2 = 0$. Then $z_1 = -z_2$ and

$$\Psi(z_2) = 1 - \Psi(z_1) = \frac{e^{z_2}}{e^{z_1} + e^{z_2}} \tag{3.56}$$

In either case the distribution is symmetric with mean $z = 0$, where $\Psi(z) = \frac{1}{2}$, and variance $\frac{\pi^2}{3}$. The deviate z is called a logit, and the pair z_1, z_2 could be called a binomial logit.

The corresponding density function can be expressed in terms of the distribution function:

$$\psi(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \Psi(z)[1-\Psi(z)] \tag{3.57}$$

Although $\psi(z)$ is heavier in the tails than $\phi(z), \Phi(z)$ closely resembles $\Psi(1.7z)$. Using the scale factor 1.7 in place of the variance matching factor 1.81379 will bring the logistic probabilities closer to the normal over the full range of the distribution, with a maximum absolute difference less than 0.01 (Johnson, Kotz, & Balakrishnan, 1995, p. 119).

An advantage of the logistic distribution over the normal is that the deviate corresponding to an observed proportion, P, is simply the log odds,

$$z(P) = \log \frac{P}{1-P} \tag{3.58}$$

For that reason, logit linear functions are frequently used in analysis of γ binomially distributed data (see Anscombe, 1956).

Inasmuch as the prediction of first choices may be viewed as an extreme value problem, it is of interest that Dubey (1969) derived the logistic distribution from an extreme value distribution of the double exponential type with mixing variable γ. Then the cumulative extreme value distribution function, conditional on γ, is

$$F(x|\gamma) = \exp[-\gamma \exp(-x)] \tag{3.59}$$

where γ has the exponential density function $g(\gamma) = \exp(-\gamma)$. The corresponding extreme value density function is

$$f(x|\gamma) = \gamma \exp(-x)[-\gamma \exp(-\gamma \exp(-x))], \quad \gamma > 0 \tag{3.60}$$

Integrating the conditional distribution function over the range of γ gives the distribution function of x:

$$F(x) = \int_0^\infty F(x|\gamma) g(\gamma) d\gamma = [1 + \exp(-x)]^{-1} \tag{3.61}$$

which we recognize as the logistic distribution.

A Bivariate Logistic Distribution

The natural extension of the logistic distribution to the bivariate case is

$$\Psi(x_1, x_2) = [1 + e^{-x_1} + e^{-x_2}]^{-1} \tag{3.62}$$

with marginal distributions $\Psi(x_1)$ and $\Psi(x_2)$. The density function is

$$\psi(x_1, x_2) = 2\Psi^3(x_1, x_2)e^{-x_1 - x_2} \tag{3.63}$$

and regression equations and corresponding conditional variances are

$$E(x_1|x_2) = 1 + \log\Psi(z_2) \tag{3.64}$$

$$E(x_2|x_1) = 1 + \log\Psi(z_1) \tag{3.65}$$

$$V(x_1|x_2) = V(x_2|x_1) = \frac{\pi^2}{3} - 1 \tag{3.66}$$

This distribution is the simplest of three bivariate logistic distributions studied in detail by Gumbel (1961). It is similar to the bivariate normal distribution in having univariate logistic distributions as margins, but unlike the normal, the bivariate logistic density is asymmetric and the regression lines are curved (see Figure 3.6). Nevertheless, the distribution function gives probability values reasonably close to bivariate normal values when the 1.7 scale correction is used (see Bock and Jones (1968, Section 9.1.1) for some comparisons of bivariate normal and bivariate logistic probabilities).

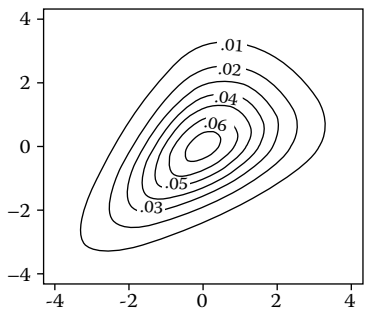

FIGURE 3.6 Contours of the bivariate logistic density. The horizontal and vertical axes are x_1 and x_2 respectively, in Equation 3.64.

A Multivariate Logistic Distribution

The natural extension of the bivariate logistic distribution to higher dimensions is

$$\Psi(\mathbf{z}) = \frac{e^{z_k}}{e^{z_1} + e^{z_2} + \ldots + e^{z_m}}, \quad k = 1, 2, \ldots, m \tag{3.67}$$

where the elements of the vector $\mathbf{z} = [z_1, z_2, \ldots, z_m]'$ are constrained to sum to zero. This vector is referred to as a multinomial logit.

Although this extension of the logistic distribution to dimensions greater than two has been applied at least since 1967 (Bock, 1970; McFadden, 1974), its first detailed study was by Malik and Abraham (1973). They derived the m-variate logistic distribution from the m-fold product of independent univariate marginal conditional distributions of the Dubey (1969) extreme value distribution with mixing variable γ. Integrating over γ gives

$$F(\mathbf{X}) = \int_{-\infty}^{\infty} \prod_{k=1}^{m} F(x_j|\gamma) g(\gamma) d\gamma = \left[1 + \sum_{k=1}^{m} e^{-x_k} \right]^{-1}, \quad m = n = 1 \tag{3.68}$$

The corresponding density function is

$$f(\mathbf{X}) = m! \exp\left[-\sum_{k=1}^{m} x_k \right] \left[1 + \sum_{k=1}^{m} e^{-x_k} \right]^{-m-1} \tag{3.69}$$

McFadden (1974) arrived at the same result by essentially the same method, although he does not cite Dubey (1969). Gumbel's bivariate distribution (above) is included for $n = 2$, and margins of all orders up to $n - 1$ are multivariate logistic and all univariate margins have mean zero and variance $\frac{\pi^2}{3}$. No comparison of probabilities for high-dimensional normal and logistic distributions has as yet been attempted.

Estimating Binomial and Multinomial Response Relations

If we substitute functions of external variables for normal or logistic deviates, we can study the relationships of these variables to the probabilities of first choice among the objects presented. In the two-category case, we refer to these as binomial response relations, and with more than two categories, as multinomial response relations. The analytical problem becomes one of estimating the coefficients of these functions rather than the logit itself. If the relationship is less than perfect, some goodness of fit will be lost relative to direct estimation of the logit (which is equivalent to estimating the category expected probabilities). The difference in the Pearson or likelihood ratio chi-square provides a test of statistical significance of the loss. Examples of weighted least squares estimation of binomial response relations in paired comparison data when the external variables represent a factorial or response surface design on the objects are shown in Section 7.3 of Bock and Jones (1968). Examples of maximum likelihood estimation of multinomial response relations appear in Bock (1970), McFadden (1974), and Chapter 8 of Bock (1975).

An earlier application of maximum likelihood in estimating binomial response relations appears in Bradley and Terry (1952). They assume the model

$$\frac{\pi_j}{\pi_j + \pi_k} \tag{3.70}$$

for the probability that object j is preferred to object k, but they estimated π_j and π_k directly rather than exponentiating in order to avoid introducing a Lagrange multiplier to constrain the estimates to sum to unity.

Luce and Suppes (1965) generalized the Bradley-Terry model to multinomial data,

$$\frac{\pi_j}{\pi_1 + \pi_2 + \ldots + \pi_m} \tag{3.71}$$

but did not make the exponential transformation to the multinomial logit and did not apply the model in estimating multinomial response relations.

Binomial and Multinomial Response Relations in the Context of IRT

In item response theory we deal with data arising from two-stage sampling: in the first stage we sample respondents from some identified population, and in the second stage we sample responses of each respondent to some number of items, usually items from some form of psychological or educational test. Thus, there are two sources of random variation in the data—between respondents and between item responses. When the response is scored dichotomously, right/wrong or yes/no, for example, the logistic distribution for binomial data applies. If the scoring is polytomous, as when the respondent is choosing among several alternatives, for instance, in a multiple-choice test with recording of each choice, the logistic distribution for multinomial data applies. If the respondent's level of performance is graded polytomously in ordered categories, the multivariate logistic can still apply, but its parameterization must be specialized to reflect the assumed order of the categories.

In IRT the "external" variable is not an observable quantity, but rather an unobservable latent variable, usually designated by θ, that measures the respondent's ability or other propensity. The binomial or multinomial logit is expressed as linear functions of θ containing parameters specific to each item. We refer to the functions that depend on θ as item response models. Item response models now in use (see Bock & Moustaki, 2007) include, for item j, the two-parameter logistic model, based on the binomial logistic distribution,

$$\Psi(\theta) = [1 + \exp(a_j\theta + c_j)]^{-1} \tag{3.72}$$

and the nominal categories model, based on the multinomial logistic distribution,

$$\Psi(\theta) = \frac{\exp(a_{jk}\theta + c_{jk})}{\sum_{l=1}^{n} \exp(a_{jl}\theta + c_{jl})} \tag{3.73}$$

under the constraints $\sum_{l=1}^{n} a_{jl} = 0$ and $\sum_{l=1}^{n} c_{jl} = 0$.

In empirical applications, the parameters of the item response models must be estimated in large samples of the two-stage data. Estimation of these parameters is complicated, however, by the presence of the propensity variable θ, which is random in the first-stage sample. Because there are potentially different values of this variable for every respondent, there is no way to achieve convergence in probability as number of respondents increases. We therefore proceed in the estimation by integrating over an assumed or empirically derived distribution of the latent variable. If the first-stage sample is large enough to justify treating the parameter estimates so obtained as fixed values, we can then use Bayes or maximum likelihood estimation to locate each respondent on the propensity dimension, with a level of precision dependent on the number of items.

The special merit of the nominal categories item response model is that no assumption about the order or other structure of the categories is required. Given that the propensity variable is one-dimensional and an ordering of the categories is implicit in the data and is revealed by the order of the coefficients a_{jk} in the nominal model (see Bock & Moustaki, 2007).

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 technical report* (NCES 1999-452). Washington, DC: National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

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

Andersen, E. B. (1997). The rating scale model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 67–84). New York: Springer.

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

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurment. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco: Jossey-Bass.

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika*, *35*, 246–254.

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

Bergan, J. R., & Stone, C. A. (1985). Latent class models for knowledge domains. *Psychological Bulletin*, *98*, 166–184.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, *39*, 357–375.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, *48*, 565–599.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1956). A generalization of the law of comparative judgment applied to a problem in the prediction of choice [Abstract]. *American Psychologist*, *11*, 442.

Bock, R. D. (1970). Estimating multinomial response relations. In E. A. R. C. Bose (Ed.), *Contribution to statistics and probability* (pp. 111–132). Chapel Hill, NC: University of North Carolina Press.

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

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

Bock, R. D. (1997). The nominal categories model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33–50). New York: Springer.

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

Bock, R. D., & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 469–513). Amsterdam: Elsevier.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. I. Method of paired comparisons. *Biometrika*, *39*, 324–345.

Childs, R. A., & Chen, W.-H. (1999). Software note: Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement*, *23*, 371–379.

de Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, *16*, 327–343.

de Ayala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development*, *25*, 172–189.

Dubey, S. D. (1969). A new derivation of the logistic distribution. *Naval Research Logistics Quarterly*, *16*, 37–40.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT*. Lincolnwood, IL: Scientific Software International.

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

Hoskens, M., & Boeck, P. D. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*, 261–277.

Huber, M. (1993). *An item response theoretical approach to scoring the Short Portable Mental Status Questionnaire for assessing cognitive status of the elderly.* Unpublished master's thesis, Department of Psychology, University of North Carolina, Chapel Hill

Johnson, N. L., Kotz, N., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.

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

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce & R. R. Bush (Eds.), *Handbook of mathematical psychology* (Vol. 3 pp. 249–410). New York: Wiley.

Malik, H., & Abraham, B. (1973). Multivariate logistic distributions. *Annals of Statistics*, *1*, 588–590.

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

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

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–122). New York: Springer.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (pp. 105–142). New York: Academic Press.

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

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer.

National Bureau of Standards (1956). Tables of the bivariate normal distribution function and related functions. *Applied Mathematic Series, Number 50.*

Ramsay, J. O. (1995). *Testgraf: A program for the graphical analysis of multiple-choice test and questionnaire data* (Technical Report). Montreal: McGill University (Psychology Department).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Annual Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321–333). Berkeley: University of California Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded cores. *Psychometric Monograph*, No. 18.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report 79-4). Knoxville: University of Tennessee (Department of Psychology).

Samejima, F. (1988). Comprehensive latent trait theory. *Behaviormetrika*, *15*, 1–24.

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

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

Steck, G. P. (1958). A table for computing trivariate normal probabilities. *Annals of Mathematical Statistics, 29*, 780–800.

Sympson, J. B. (1983, June). *A new IRT model for calibrating multiple choice items*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175–186.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (Chap. 4, pp. 141–186). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (Chap. 3, pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika*, *49*, 501–519.

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

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin*, *104*, 385–395.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247–260.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 278–286.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wainer, H., Thissen, D., & Sireci, S. G. (1991). DIFferential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, *28*, 197–219.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*, 469–492.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–214.