Chapter 4

The General Graded Response Model

Fumiko Samejima
University of Tennessee

Personal Reflections on the Origins of the Model

When item response theory (IRT) originated and was developed in psychology and sociology in the 1940s, 1950s, and the first half of the 1960s, the theory only dealt with dichotomous responses, where there are only two item score categories, for example, correct and incorrect in ability measurement, true and false in personality measurement. As a graduate student I was very much impressed by Fred Lord’s (1952) Psychometric monograph “A Theory of Mental Test Scores” and could foresee great potential for latent trait models. It seemed that the first thing to be done was to expand IRT to enable it to deal with ordered, multicategory responses and enhance its applicability, not only in psychology, sociology, and education, but also in many other social and natural science areas. That opportunity came when I was invited to spend one year as visiting research psychologist in the Psychometric Research Group of the Educational Testing Service (ETS), Princeton, New Jersey, in 1966. The essential outcomes of the research conducted during my first year in the United States were published in Samejima (1969).

A subsequent invitation to work in the psychometric laboratory at the University of North Carolina at Chapel Hill allowed continuation of the initial work. The essential outcomes of the research conducted in 1967–1968 were published in Samejima (1972). This second monograph is as important as the first, and the two monographs combined propose the fundamental tenets of the general graded response model framework.
In recent years, more and more researchers have started citing these two *Psychometrika* monographs in their research. In this chapter I will try to correct common misunderstandings among researchers, as well as introduce and explain further developments in the general graded response model.

**Rationale**

In the present chapter, uni-dimensional latent trait models are almost exclusively discussed, where the latent trait assumes any real number. The general graded response model is a comprehensive mathematical model that provides the general structure of latent trait models that deal with cases in which item \( g \), is the smallest observable unit for measuring the latent trait \( \theta \) and, subsequently, one of the graded item scores, or ordered polytomous item scores, \( x_g = 0, 1, 2, \ldots, m_g \), \( (m_g \geq 1) \), is assigned to each response. The highest score, \( m_g \), can be any positive integer, and the general graded response model does not require all items in a test or questionnaire to have the same values of \( m_g \). This is a great advantage, and it makes it possible to mix dichotomous response items with those whose \( m_g \)'s are greater than unity. Models that belong to the general framework discussed in this chapter include the normal ogive model, the logistic model, the graded response model expanded from the logistic positive exponent family of models (Samejima, 2008), the acceleration model, and the models expanded from Bock's nominal response model. Thus the framework described here applies for any of these models.

Graded response model (GRM) was proposed by Samejima (1969, 1972), to provide a general theoretical framework to deal with the graded item scores, 0, 1, 2, \ldots, \( m_g \), in the item response theory (IRT), whereas in the original IRT the item scores were limited to 0, 1. As is explained later in this chapter, the logistic model is a specific model that belongs to GRM. Because the logistic model was applied for empirical data in early years, as exemplified by Roche, Wainer, and Thissen (1975); however, researchers started treating the logistic model as if it were the GRM. Reading this chapter, the reader will realize that GRM is a very comprehensive concept that includes normal ogive model, logistic model, expanded model from the logistic positive exponent family of models, BCK-SMJ model, acceleration model, etc. Correct terminology is important; otherwise, correct research will become impossible.

The latent trait can be any construct that is hypothesized to be behind observable items, such as the way in which ability is behind performance on problem-solving questions, general attitude toward war is represented by responses to peace/war-oriented statements, maturity of human bodies is represented by experts’ evaluations of x-ray films, and so on.

Throughout this paper, the latent trait is denoted by \( \theta \), which assumes any real number in \(( -\infty, \infty )\), except for the case where the multidimensional latent space \( \Theta \) is considered.
The general graded response model framework (Samejima, 1997, 2004) is based on the following five functions for each graded item score $x_g$:

1. Processing function (PRF), $M_{x_g}(\theta) \ (x_g = 0, 1, 2, \ldots, m_g, m_g + 1)$. This is a joint conditional probability, given $\theta$, and given that the individual has passed the preceding process. Specifically $M_{x_g}(\theta) = 1$ for $x_g = 0$, and $M_{x_g}(\theta) = 0$ for $x_g = m_g + 1$ for all $\theta$, respectively, since there is no process preceding $x_g = 0$ and $m_g + 1$ is a nonexistent, imaginary graded score no one can attain.

2. Cumulative operating characteristic (COC), $P_{x_g}^*(\theta) \ (x_g = 0, 1, 2, \ldots, m_g, m_g + 1)$, defined by

$$P_{x_g}^*(\theta) = \text{prob}[X_g \geq x_g | \theta] = \prod_{u=x_g}^{m_g} M_u(\theta). \quad (4.1)$$

This is the conditional probability, given $\theta$, that the individual gets the graded score $x_g$ or greater. In particular, $P_{x_g}^*(\theta) = 1$ for $x_g = 0$ and $P_{x_g}^*(\theta) = 0$ for $x_g = m_g + 1$ for the entire range of $\theta$, since everyone obtains a score of 0 or greater, and no one gets a score of $m_g + 1$ or greater. Also from Equation 4.1, $P_{x_g}^*(\theta) = M_{x_g}(\theta)$ when $x_g = 0$ for the entire range of $\theta$.

Terminology Note: Elsewhere in this book and in the literature this function is called a category boundary function or a threshold function. In another difference, we and other authors in this book have indexed this and other types of polytomous response functions using $i$ for items and $k$ for category response [e.g., $p_k^*$] while the Author uses $g$ for items, following Lord & Novick (1968, Chap. 16). Note that in addition to using different letters to index item and category components, Samejima also indexes the category response ($x$) prior to the item index ($g$), where other authors index the item prior to the category response.

3. Operating characteristic (OC), $P_{x_g}(\theta) \ (x_g = 0, 1, 2, \ldots, m_g)$ defined by

$$P_{x_g}(\theta) \equiv \text{prob}[X_g = x_g | \theta] = P_{x_g}^*(\theta) - P_{x_g+1}^*(\theta). \quad (4.2)$$

This is the conditional probability, given $\theta$, that the individual obtains a specific graded score $x_g$.

Note that when $m_g = 1$, from Equations 4.1 and 4.2, both $P_{x_g}^*(\theta)$ and $P_{x_g}(\theta)$ for $x_g = 1$ become the item characteristic function (ICF; Lord & Novick, 1968, Chap. 16) for a dichotomous item. Thus, a specific graded response model, defined in this way, models dichotomous item responses as a special case, and so the general graded response model framework also applies to dichotomous response items.

4. Basic function (BSF), $A_{x_g}(\theta) \ (x_g = 0, 1, 2, \ldots, m_g)$ defined by

$$A_{x_g}(\theta) \equiv \frac{\partial}{\partial \theta} \log P_{x_g}(\theta) = [P_{x_g}(\theta)]^{-1} \frac{\partial}{\partial \theta} P_{x_g}(\theta). \quad (4.3)$$
It is obvious that the basic function exists as long as \( P_{xg}(\theta) \) is positive for the entire range of \( \theta \), and is differentiable with respect to \( \theta \).

5. Item response information function (IRIF), \( I_{xg}(\theta) \) \((x_g = 0,1,2,\ldots,m_g)\), is defined by

\[
I_{xg}(\theta) \equiv -\frac{\partial^2}{\partial \theta^2} \log P_{xg}(\theta) = -\frac{\partial}{\partial \theta} A_{xg}(\theta) \tag{4.4}
\]

\[
= -[P_{xg}(\theta)]^{-2} \left[ P_{xg}(\theta) \frac{\partial^2}{\partial \theta^2} P_{xg}(\theta) - \left\{ \frac{\partial}{\partial \theta} P_{xg}(\theta) \right\}^2 \right]
\]

\[
= -[P_{xg}(\theta)]^{-2} \left[ \frac{\partial^2}{\partial \theta^2} P_{xg}(\theta) \right] + [P_{xg}(\theta)]^{-2} \left[ \frac{\partial}{\partial \theta} P_{xg}(\theta) \right]^2
\]

\[
= [P_{xg}(\theta)]^{-2} \left[ \frac{\partial}{\partial \theta} P_{xg}(\theta) \right]^2 - [P_{xg}(\theta)]^{-2} \left[ \frac{\partial^2}{\partial \theta^2} P_{xg}(\theta) \right].
\]

One can conclude that the item response information function exists as far as \( P_{xg}(\theta) \) is positive for the entire range of \( \theta \), and is twice differentiable with respect to \( \theta \).

Thissen and Steinberg (1986) called the normal ogive model and logistic model for graded responses the difference models, and models for graded responses expanded from Bock’s (1972) nominal model divide-by-total models. The naming may be a little misleading, however, because the general framework for graded response models that has been introduced above accommodates both of Thissen’s two categories. From Equation 4.2 we see that for any \( x_g (= 0,1,2,\ldots,m_g) \),

\[
P_{x_1}^{\ast}(\theta) \geq P_{x_1+1}^{\ast}(\theta), \tag{4.5}
\]

for the entire range of \( \theta \) in order to satisfy the definition of the operating characteristic \( P_{x_1}(\theta) \) because it is a conditional probability.

---

**Item Information Function**

Samejima (1973b) defined the item information function, \( I^g(\theta) \), for the general graded response item \( g \) as the conditional expectation of the IRIF, given \( \theta \), that was defined by Equation 4.4. Thus it can be written

\[
I^g(\theta) \equiv E[I_{xg}(\theta)|\theta] = \sum_{x_g=0}^{m_g} I_{xg}(\theta)P_{xg}(\theta). \tag{4.6}
\]

Note that this item information function (IIF) of the graded item \( g \) includes Birnbaum’s (1968) item information function on the dichotomous responses.
as a special case. To simplify the notation, let the item characteristic function (ICF) for dichotomous responses (Lord & Novick, 1968, Chapter 16) be represented as

\[
P_g(\theta) \equiv \text{prob}[X_g = 1|\theta] = P_{x_g}(\theta; x_g = m_g),
\]

(4.7)

where \(m_g = 1\), and

\[
Q_g(\theta) \equiv \text{prob}[X_g = 0|\theta] = P_{x_g}(\theta; x_g = 0) = 1 - P_g(\theta).
\]

(4.8)

Let \(P'_g(\theta)\) and \(Q'_g(\theta)\) denote their first derivatives with respect to \(\theta\), respectively. Due to the complementary relationship of Equations 4.7 and 4.8 we can see that

\[
Q'_g(\theta) = -P'_g(\theta)
\]

(4.9)

and from Equation 4.9,

\[
Q''_g(\theta) = -P''_g(\theta),
\]

(4.10)

where \(P''(\theta)\) and \(Q''(\theta)\) denote the second derivatives of \(P_g(\theta)\) and \(Q_g(\theta)\) with respect to \(\theta\), respectively.

From Equations 4.4 and 4.7 to 4.10 we can now rewrite our IRIF as

\[
I_{u_g}(\theta) = \begin{cases} 
[Q_g(\theta)]^{-2}[-P'_g(\theta)]^2 - [Q_g(\theta)]^{-1}[-P''_g(\theta)] & u_g = 0 \\
[P_g(\theta)]^{-2}[P'_g(\theta)]^2 - [P_g(\theta)]^{-1}[P''_g(\theta)] & u_g = 1.
\end{cases}
\]

(4.11)

Thus from Equations 4.6 and 4.11 for the IIF a dichotomous response item can be written as

\[
I_{x_g}(\theta) = I_{u_g}(\theta; u_g = 0)Q_g(\theta) + I_{u_g}(\theta; u_g = 1)P_g(\theta)
\]

(4.12)

\[
= [Q_g(\theta)]^{-2}[-P'_g(\theta)]^2 - [Q_g(\theta)]^{-1}[-P''_g(\theta)] + [P_g(\theta)]^{-2}[P'_g(\theta)]^2 - [P''_g(\theta)]
\]

\[
= [P'_g(\theta)]^2[Q_g(\theta)]^{-1} + [P_g(\theta)]^{-1}
\]

\[
= [P'_g(\theta)]^2[P_g(\theta) + Q_g(\theta)][P_g(\theta)Q_g(\theta)]^{-1}
\]

\[
= [P'_g(\theta)]^2[P_g(\theta)Q_g(\theta)]^{-1}.
\]

The last expression of Equation 4.12 equals Birnbaum’s (1968) IIF for the dichotomous response item (p. 454).
Expansion of Latent Trait Models for Dichotomous Responses to Those for Graded Responses

It is noted from Equation 4.1 that the definition of the COC, \( P_x (\theta) \), of the graded response score \( x_g \) becomes the ICF, \( P_g (\theta) \), that is defined by Equation 4.7, if \( X_g \) is replaced by the binary item score \( U_g \) and \( m_g \) is 1. This implies that expansion of the general dichotomous response model to the general graded response model can be done straightforwardly, with the restriction of Equation 4.5.

For example, suppose that the final grade of a mathematics course that all math majors are required to pass is based on five letter grades, A, B, C, D, and F. For these graded responses, \( m_g = 4 \). When we reclassify all math majors into pass and fail, there are, in general, \( m_g \) different ways to set the borderline of pass and fail between (1) A and B, (2) B and C, (3) C and D, and (4) D and F. It is noted that Way 1 is the strictest of passing math majors, Way 4 is the most generous, Way 2 is moderately strict, and Way 3 is moderately generous.

Now the course grade has been changed to a set of two grade categories from five, in four different ways, and in each case the item characteristic function \( P_g (\theta) \) that is defined by Equation 4.7 can be specified. Note that these four ICFs equal the COCs, that is, \( P_{x_g} (\theta) \), for the letter grades A, B, C, and D, respectively.

Figure 4.1 illustrates these \( m_g = 4 \) ICFs. Because Way 1 is the strictest of passing math majors and Way 4 is the most generous, it is natural that their ICFs are located at the right-most and left-most parts in Figure 4.1.
respectively, and the other two ICFs are positioned and ordered between the two with respect to their levels of generosity. These curves also satisfy Equation 4.5, as is obvious from the nature of recategorizations. Thus it is clear from the definitions of pass and fail and Equation 4.2 that the OCs for A, B, C, D, and F are given as the differences of the two adjacent curves, given $\theta$, as indicated in Figure 4.1. Note that these curves do not have to be identical in shape or point symmetric, but should only satisfy Equation 4.5.

**Terminology Note:** As the author points out, it is not necessary for the curves in Figure 4.1 to have identical shapes. This is a key distinction between the heterogeneous and the homogenous models. In the homogeneous case the COC forms are always parallel, whereas in the heterogeneous case they are not necessarily parallel.

The above explanation may be the easiest way to understand the transition from models of dichotomous responses to those of graded responses. Because of this close relationship between the ICFs for dichotomous responses and the COCs for graded responses, in the following sections specific mathematical models that belong to the general graded response model will be represented by their COCs in most cases.

**Unique Maximum Condition**

**Terminology Note:** The unique maximum condition is an important concept in the graded response model framework and will be referred to throughout this chapter. In essence this condition requires that for a given model and for a given likelihood function of the specific response pattern, there exists a single maximum point that can be used as the estimate of the latent trait.

In IRT, the test score is practically useless in estimating the individual’s latent trait level even though asymptotically there is a one-to-one correspondence between the test score and the latent trait $\theta$. (Note that no tests have infinitely many items.)

The main reason is that the use of the test score will reduce the local accuracy of latent trait estimation (cf. Samejima, 1969, Chap. 6, pp. 43–45; 1996b), unless there exists a test score or any summary of the response pattern that is a sufficient statistic for the model, such as Rasch’s (1960) model or the logistic model (Birnbaum, 1968) for dichotomous responses. In general, the direct use of the response pattern or the sequence of item scores, $x_k$'s, rather than a single test score, therefore, is strongly encouraged for estimating individuals’ latent traits.

Although the proposal of the logistic model as a substitute for the normal ogive model was a big contribution in the 1960s, in these days when most researchers have access to electronic computers, there is little need for the substitution of a model by another that has a sufficient statistic.
Let \( v \) be a specified response pattern, or a sequence of specified graded item scores, that includes a sequence of specified binary item scores as a special case, such that \( v' = (x_1, x_2, \ldots, x_n) \) for a set of \( n \) items. Because of local independence (Lord & Novick, 1968, Chap. 16), it can be written

\[
L_v(\theta) = P_v(\theta) = \prod_{x \in v} P_{x}(\theta),
\]

(4.13)

where \( L_v(\theta) \) is the likelihood function for the specific response pattern \( V = v \), and \( P_v(\theta) \) denotes the conditional probability of the response pattern \( v \), given \( \theta \). Using Equations 4.3 and 4.13, the likelihood equation is given by

\[
\frac{\partial}{\partial \theta} \log L_v(\theta) = \sum_{x \in v} \frac{\partial}{\partial \theta} \log P_{x}(\theta) = \sum_{x \in v} A_{x}(\theta) \equiv 0. \tag{4.14}
\]

Thus, there exists a sufficient condition that a specific graded response model provides a unique maximum for any likelihood function (i.e., for each and every response pattern) and the condition is that both of the following requirements are met:

1. The basic function \( A_{x}(\theta) \) of each and every graded score \( x \) of each item \( g \) is strictly decreasing in \( \theta \).
2. Its upper and lower asymptotes are nonnegative and nonpositive, respectively.

For brevity, this condition is called the unique maximum condition (Samejima, 1997, 2004).

On the dichotomous response level, such frequently used models as the normal ogive model, logistic model, and all models that belong to the logistic positive exponent family of models, satisfy the unique maximum condition. A notable exception is the three-parameter logistic model (3PL; Birnbaum, 1968). In that model, for \( x = 1 \), the unique maximum condition is not satisfied, and it is quite possible that for some response patterns the unique MLE does not exist (for details see Samejima, 1973b).

An algorithm for writing all the basic functions and finding the solutions of Equation 4.14 for all possible response patterns is easy and straightforward for most models that satisfy the unique maximum condition, so the unique local or terminal maximum likelihood estimate (MLE) of \( \theta \) can be found easily without depending on the existence of a sufficient statistic.

It should be noted that, when the set of \( n \) items do not follow a single model, but they follow several different models, as long as all of these models satisfy the unique maximum condition, a unique local or terminal maximum of the likelihood function of each and every possible response pattern is also assured to exist.
Criteria for Evaluating Specific Graded Response Models

Samejima (1996a) proposed five different criteria to evaluate a latent trait model from a substantive point of view:

1. The principle behind the model and the set of accompanying assumptions agree with the psychological nature that underlies the data.
2. Additivity 1, that is, if the existing graded response categories get finer (e.g., pass and fail are changed to A, B, C, D, and F), their OCs can still be specified in the same model.
3. Additivity 2, that is, following a combination of two or more adjacent response categories (e.g., A, B, C, D, and F to pass and fail), the OCs of the newly combined categories can still be specified in the same mathematical form. (If additivities 1 and 2 hold, the model can be naturally expanded to a continuous response model.)
4. The model satisfies the unique maximum condition (discussed above).
5. Modal points of the OCs of the \( m_g + 1 \) graded response categories are ordered in accordance with the graded item scores, \( x_g = 0, 1, 2, ..., m_g \).

Of these five criteria, the first is related to the data to which the model is applied, but the other four can be used strictly mathematically.

Response Pattern and Test Information Functions

Because the specific response pattern \( v \) is the basis of ability estimation, it is necessary to consider the information function provided by \( v \). Samejima (1973b) defined the response pattern information function (RPIF), \( I_v(\theta) \), as

\[
I_v(\theta) \equiv -\frac{\partial^2}{\partial \theta^2} \log P_v(\theta).
\]  

Equation 4.15 is analogous to the definition of the IRIF, \( I_{xg}(\theta) \), given earlier by Equation 4.4. Using Equations 4.13, 4.14, and 4.4, this can be changed to

\[
I_v(\theta) = -\sum_{x \in v} \frac{\partial^2}{\partial \theta^2} \log P_{x}(\theta) = \sum_{x \in v} I_{xg}(\theta)
\]  

indicating that the RPIF can be obtained as the sum of all IRIFs for \( x \in v \).

The test information function (TIF), \( I(\theta) \), in the general graded response model is defined by the conditional expectation, given \( \theta \), of the RPIF, \( I_v(\theta) \), as analogous to the relationship between the IIF and IRIFs. Thus

\[
I(\theta) \equiv E[I_v(\theta)|\theta] = \sum_v I_v(\theta)P_v(\theta).
\]  

Since it can be written that

\[
P_{xg}(\theta) = \sum_{x \in v} P_v(\theta),
\]
we obtain from Equations 4.17 to 4.18 and 4.6

\[ I(\theta) = \sum_{\gamma=1}^{n} \sum_{x_{\gamma}=0}^{m_{\gamma}} I_{x_{\gamma}}(\theta)P_{x_{\gamma}}(\theta) = \sum_{\gamma=1}^{n} I_{x_{\gamma}}(\theta) = \sum_{\gamma=0}^{n} I_{x_{\gamma}}(\theta). \] (4.19)

Note that this outcome, that the test information function equals the sum total of the item information functions, is true only if the individual’s ability estimation is based on that individual’s response pattern and not its aggregate, such as a test score, unless it is a simple sufficient statistic, as is the case with the Rasch model. Otherwise, the test information function assumes a value less than the sum total of the item information functions (Samejima, 1996b).

The final outcome of Equation 4.19, that the TIF equals the sum total of IIFs over all items in the test, questionnaire, and so on, is the same as the outcome of the general dichotomous response model (cf. Birnbaum, 1968, Chap. 20).

It should be noted that, because of the simplicity of the above outcome, that is, the test information function equals the sum total of the item information functions, researchers tend to take it for granted. It is necessary, however, that the reader understands how this outcome was obtained based on the definitions of the TIF and the RPIF, in order to apply IRT properly and innovatively.

### Latent Trait Models in the Homogeneous Case

All specific latent trait models that belong to the general graded response framework can be categorized into the homogeneous case and the heterogeneous case. To give some examples, such models as the normal ogive model and logistic model belong to the former, and the graded response model expanded from the logistic positive exponent family of models (Samejima, 2008), acceleration model (Samejima, 1995), and graded response models expanded from Bock’s (1972) nominal response model belong to the latter.

**Terminology Note:** As mentioned earlier, the essential difference between the homogeneous case and the heterogeneous case is whether or not the shapes of COCs vary from one score category to the next. In the homogeneous case the COCs are parallel, whereas in the heterogeneous case they are not parallel.

### Rationale Behind the Models in the Homogeneous Case

Lord set a hypothetical relation between dichotomous item score \( u_{g} \) and latent trait \( \theta \) that leads to the normal ogive model (cf. Lord & Novick, 1968, Section 16.6). He assumes a continuous variable \( Y_{g} \) behind the item score \( u_{g} \) and the critical value \( \gamma_{g} \), as well as the following:

1. An individual \( a \) will get \( u_{a} = 1 \) (e.g., pass) if \( Y_{g} \geq \gamma_{g} \), and if \( Y_{g} < \gamma_{g} \), the individual will obtain \( u_{a} = 0 \) (e.g., fail). This assumption may be reasonable if the reader thinks of the fact that within a group of individuals who get credit for solving problem \( g \) there are diversities of different levels of ability; that is, some individuals may solve it very easily while some others may barely make it after having struggled a lot, and so on.
2. The regression (conditional expectation) of $Y'_g$ on $\theta$ is linear.
3. The conditional distribution of $Y'_g$, given $\theta$, is normal.
4. The variance of these conditional distributions is the same for all $\theta$.

The figure that Lord used for the normal ogive model for dichotomous responses (Lord & Novick, 1968, Figure 16.6.1) is illustrated in Figure 4.2. In this figure, two critical values, $\gamma_{g0}$ and $\gamma_{g1}$ ($\gamma_{g0} < \gamma_{g1}$), are used instead of a single $\gamma_g$, as is also shown in Figure 4.2, and Hypothesis 1 is changed to Hypothesis 1*: Individual $a$ will get $x_g = 2$ (e.g., honor pass) if $Y'_g \geq \gamma_{g1}$, $x_g = 1$ (e.g., pass) if $\gamma_{g0} \leq Y'_g < \gamma_{g1}$, and $x_g = 0$ (e.g., fail) if $Y'_g < \gamma_{g0}$. As is obvious from Equations 4.1 and 4.2, the shaded area for the interval $[\gamma_{g1}, \infty)$ indicates the OC for $x_g = 2$ of item $g$; for $[\gamma_{g0}, \gamma_{g1})$, the OC for $x_g = 1$; and for $(-\infty, \gamma_{g0})$, the OC for $x_g = 0$ at each of the two levels of $\theta$ in Figure 4.2. The above example leads to the normal ogive model for graded responses when $m_g = 2$. By increasing the number of the critical values, $\gamma'_g$s, however, a similar rationale can be applied for any positive integer for $m_g$.

It should also be noted that Hypotheses 3 and 4 can be replaced by any other conditional density functions, symmetric or asymmetric, in so far as their shapes are identical at all the fixed values of $\theta$. All those models are said to belong to the homogeneous case. Thus, a model that belongs to the homogeneous case does not imply that its COCs are point symmetric for $x_g = 1, 2, ..., m_g$, nor do its OCs provide symmetric curves for $x_g = 1, 2, ..., m_g - 1$, although both the normal ogive and logistic models do so.
From the above definition and observations, it is clear that any graded response model that belongs to the homogeneous case satisfies additivities 1 and 2, which were introduced earlier as criteria for evaluating mathematical models for graded responses.

From Equation 4.1 it can be seen that a common feature of the models in the homogeneous case is that the cumulative operating characteristics, $P_x^*(\theta)$'s, for $x = 1, 2, ..., m$ are identical in shape except for the positions on the $\theta$ dimension, which are ordered in accordance with the graded score $x$.

**Normal Ogive Model (NMLOG)**

The rationale behind the normal ogive model was provided earlier as an example of the rationale behind any model that belongs to the homogeneous case. In the normal ogive model, the COC is specified by

$$P_x^*(\theta) = \left[2\pi\right]^{1/2} \int_{-\infty}^{a_x(\theta-b_x)} \exp\left[-\frac{z^2}{2}\right] dz, \quad (4.20)$$

where $a_x$ denotes the item discrimination parameter and $b_x$ is the item response difficulty parameter, the latter of which satisfies

$$-\infty = b_0 < b_1 < ... < b_m < b_{m+1} = \infty. \quad (4.21)$$

Figures 4.3a and b illustrate the OCs in the normal ogive model, for two different items, both with $m = 5$, but having different $a_x$ and $b_x$'s, that is, for the item in Figure 4.3a $a_x = 1.0$ and $b_x = -1.50, -0.50, 0.00, 0.75, 1.25$, while for the item in Figure 4.3b $a_x = 2.0$ and $b_x = -2.00, -1.00, 0.00, 1.00, 2.00$.

It is noted that for both items the OCs for $x = 0$ and $x = 5 (=m)$ are strictly decreasing and increasing in $\theta$, respectively, with unity and zero as the two asymptotes in the former, and with zero and unity as the two asymptotes in the latter. They are also point symmetric, meaning if each curve is rotated by $180^\circ$ around the point, $\theta = b_1$ and $P_0(\theta) = 0.5$ when $x = 0$, and $\theta = b_5$ and $P_5(\theta) = 0.5$ when $x = 5$, then the rotated upper half of the curve overlaps the original lower half of the curve, and vice versa. It is also noted that in both figures the OCs for $x = 1, 2, 3, 4$ are all unimodal and symmetric.

These two sets of OCs provide substantially different impressions, because in Figure 4.3a the four bell-shaped curves have variously different heights. They are determined by the distance, $b_{x+1} - b_x$. In this figure, the modal point of the curve for $x = 1$ equals $b_2 - b_1 = 1.00$, and the maximal OC is higher than any others, because $b_3 - b_2 = 0.50$, $b_4 - b_3 = 0.75$, and $b_5 - b_4 = 0.50$. Thus the second highest modal point belongs to $x = 3$, and the lowest is shared by $x = 2$ and $x = 4$. For the item in Figure 4.3b, it is noted that those distances ($b_{x+1} - b_x$) are uniformly 1.00; thus the heights of the four bell-shape curves are all equal. The height of a bell-shaped curve
The General Graded Response Model

also depends on the value of $a_g$. In Figure 4.3b, the common height of the four bell-shaped curves is higher than the one for $x_g = 1$ in Figure 4.3a, and this comes from the larger value of $a_g (= 2.0)$ for the item in Figure 4.3b than that of the item in Figure 4.3a for which $a_g = 1.0$. It should also be noted that in each of the two examples the modal point of the OCs is ordered in accordance with the item score, $x_g = 1, 2, 3,$ and $4.$

The above characteristics of the NMLOG are also shared by the logistic model that will be introduced in the following section.

It has been observed (Samejima, 1969, 1972) that the BSFs of the $x_g$'s are all strictly decreasing in $\theta$, with $0$ and $-\infty$ as the two asymptotes for $x_g = 0$, with $\infty$ and $-\infty$ for $0 < x_g < m_g$, and with $\infty$ and $0$ for $x_g = m_g$, respectively, indicating that the model satisfies the unique maximum condition discussed above. The IRIFs for all $0 \leq x_g \leq m_g$ are positive for the entire range of $\theta$. The processing functions (PRFs) are all strictly increasing in $\theta$ for all $0 < x_g \leq m_g$, with zero and unity as the two asymptotes. (For more details, cf. Samejima, 1969, 1972.)

FIGURE 4.3 Two examples of six-category operating characteristics when categories are (a) not equally spaced and (b) equally spaced.
Logistic Model (LGST)

In the logistic model, the cumulative operating characteristic is specified by

\[
P^*_s(\theta) = \left[1 + \exp\left(-Da_s(\theta - b_{x_s})\right)\right]^{-1}, \tag{4.22}
\]

where \(a_s\) denotes the item discrimination parameter and \(b_x\) is the item response difficulty parameter that satisfies the inequality presented in Equation 4.21, as is the case with the normal ogive model. \(D\) is a scaling factor usually set equal to 1.702 or 1.7 so that Equation 4.22 provides a very close curve to Equation 4.20, that is, the COC in the normal ogive model when the same values of item discrimination parameter \(a_s\) and item response difficulty parameters \(b_{x_s}\)'s are used. As is expected, the set of COCs and the set of OCs are similar to the corresponding sets in the normal ogive model, illustrated in Figure 4.3a and b.

Notable differences are found in its PRFs and BSFs, however (Samejima, 1969, 1972, 1997). Although the PRFs are strictly increasing in \(\theta\) for all \(0 < x_s \leq m_s\) and their upper asymptotes are all unity, as is the case with the NMLOG, their lower asymptotes equal \(\exp[-Da_s(b_{x_s} - b_{x_{s-1}})]\) (cf. Samejima, 1972, p. 43), which is positive except for \(x_s = 1\), where it is zero. This indicates that, unlike in the NMLOG, in the LGST for all \(1 < x_s \leq m_s\) the lower asymptotes are positive, and moreover, the closer the item difficulty parameter \(b_{x_s}\) is to that of the preceding item score, \(b_{x_{s-1}}\), the better the chances are of passing the current step \(x_s\), giving favor to individuals of lower levels of ability. This fact is worth taking into consideration when model selection is considered. (A comparison of LGST processing functions with those in the NMLOG is illustrated in Samejima (1972, Figure 5-2-1, p. 43).)

Although the BSFs are strictly decreasing in \(\theta\) for all \(0 \leq x_s \leq m_s\), unlike in the normal ogive model, its two asymptotes for \(x_s = 0\) are zero and a finite value, \(Da_s\), those for \(x_s = m_s\) are \(Da_s\) and zero, and for all other intermediate \(x_s\)'s their asymptotes are finite values, \(Da_s\) and \(-Da_s\), respectively. The unique maximum condition is also satisfied (Samejima, 1969, 1972), however, and the IRIFs for all \(0 \leq x_s \leq m_s\) are positive for the entire range of \(\theta\).
An Example of the Application of the Logistic Model to Medical Science Research

It was personally delightful when, as early as 1975, Roche, Wainer, and Thissen applied the logistic model for graded responses in medical science research in the book *Skeletal Maturity*. The research is a fine combination of medical expertise and a latent trait model for graded responses. It is obvious that every child grows up to become an adolescent and then an adult, and its skeletal maturity progresses with age. But there are many individual differences in the speed of that process, and a child's chronological age is not an accurate indicator of his or her skeletal maturity. For example, if you take a look at a group of sixth graders, in spite of the closeness of their chronological ages, some of them are already over 6 feet tall and look like young adults, while others still look like small children. Measuring the skeletal maturity of each child accurately is important because certain surgeries have to be conducted when a child or adolescent’s skeletal maturity has reached a certain level, to give an example.

In Roche et al. (1975), x-ray films of the left knee joint that were taken from different angles were mostly used as items, or skeletal maturity indicators. The items were grouped into three categories: femur (12), tibia (16), and fibula (6). The reference group of subjects for the skeletal maturity scale consists of 273 girls and 279 boys of various ages. A graded item score was assigned to each of those subjects for each item following medical experts’ evaluations of the x-ray film. The reader is strongly encouraged to read the entire Roche et al. text to learn more about this valuable research and to see how the LGST, a specific example of the graded response model framework, has been applied in practice.

Further Observations of the Normal Ogive and Logistic Models

Both the normal ogive and logistic models satisfy the unique maximum condition, and the additivities 1 and 2 criteria (discussed above), and the modal points of their OCs are arranged in accordance with the item scores and they can be naturally expanded to respective continuous response models (cf. Samejima, 1973a).

It is clear that in the NMLOG and LGST for graded responses (and also in many other models in the homogeneous case) the COC can be expressed as

$$P_{a}(\theta) = \int_{-\infty}^{a} \psi_{g}(z) \, dz,$$

(4.23)

where \(\psi_{g}(z)\) is replaced by the standard normal and logistic density functions, respectively.
It can be seen from Equation 4.23 that these models for graded responses can be expanded to their respected models for continuous responses. Replacing $x_i$ by $z_i$ in Equation 4.23, the operating density characteristic $H_{z}(\theta)$ (Samejima, 1973a) for a continuous response $z_i$ is defined by

$$H_{z}(\theta) \equiv \lim_{\Delta z \to 0} \frac{P'_z(\theta) - P'_{z+\Delta z}(\theta)}{\Delta z} = a_g \psi_{z_g}(a_g(\theta - b_{z_g})) \frac{d}{d\theta} - b_{z_g}$$

where $a_g$ is the item discrimination parameter and $b_{z_g}$ is the item response difficulty parameter, the latter of which is a continuous, strictly increasing, and differentiable function of $z_g$.

In the normal ogive model for continuous responses, there exists a sufficient statistic, $t(v)$, such that

$$t(v) = a_g^2 b_{z_g}$$

and the MLE of $\theta$ is provided by dividing $t(v)$ by the sum total of $a_g^2$ over all $n$ items.

When the latent space $\Theta$ is multidimensional, that is,

$$\Theta' = \{\theta_1, \theta_2, ..., \theta_j, ..., \theta_r\}$$

in the NMLOG the sufficient statistic becomes a vector of order $r$, that is,

$$t(v) = \sum_{z_g} a_g a'_g b_{z_g}$$

(4.25)

where the bold letters indicate vectors of order $r$, and the MLE of $\theta$ is given by the inverse of the matrix $\sum_{z_g} a_g a'_g$ postmultiplied by $t(v)$. It is noted that Equation 4.24 is a special case of Equation 4.25 when $r = 1$ (for details, the reader is directed to Samejima (1974)).

For graded response data, when $m_g$ is very large, a continuous response model may be more appropriately applied instead of a graded response model, as is often done in applying statistic methods. (Note that the test score is a set of finite values, and yet it is sometimes treated as a continuous variable, for example.) In such a case, if the normal ogive model fits our data, the MLE of $\theta$ will be obtained more easily, taking advantage of the sufficient statistic when the latent space is multidimensional, as well as unidimensional.

It has been observed (Samejima, 2000) that the normal ogive model for dichotomous responses provides some contradictory outcomes in the orders of MLEs of $\theta$, because of the point-symmetric nature of its ICF that is characterized by the right-hand side of Equation 4.20, with the replacement of the item response difficulty parameter $b_{z_g}$ by the item difficulty parameter $b_{z}$. To illustrate this fact, Table 1 of Samejima (2000) presents all 32 ($= 2^5$) response patterns of five hypothetical dichotomous items following the NMLOG, with $a_g = 1.0$ and $b_{z_g} = -3.0, -1.5, 0.0, 1.5, 3.0$, respectively, that are arranged in the ascending order of the MLEs. When the model is changed to the LGST, because all five items share the same item discrimination parameter, $a_g = 1.0$, the simple number correct test score becomes a...
The General Graded Response Model

sufficient statistic, and a subset of response patterns that have the same number of \( \mu = 1 \) shares the same value of MLE.

The following are part of all 32 response patterns listed in that table and their corresponding MLEs in the NMLOG:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Corresponding MLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 (10000)</td>
<td>-2.284</td>
</tr>
<tr>
<td>7 (00001)</td>
<td>-0.866</td>
</tr>
<tr>
<td>26 (01111)</td>
<td>0.866</td>
</tr>
<tr>
<td>31 (11110)</td>
<td>2.284</td>
</tr>
</tbody>
</table>

It is noted that the first two response patterns share the same subresponse pattern, 000, for Items 2 to 4, and the second two share the same subresponse pattern 111 for the same three items. In the first pair of response patterns, the subresponse patterns for Items 1 and 5 are 10 and 01, respectively, while in the second pair they are 01 and 10. Because the only difference in each pair of response patterns is this subresponse pattern of items 1 and 5, it is contradictory that in the first pair (Patterns 2 and 7) success in answering the most difficult item is more credited (-0.866 > -2.284) in the normal ogive model, while in the second pair (Patterns 26 and 24) success in answering the easiest item is more credited (2.284 > 0.866).

Observations like that above provided the motivation for proposing a family of models, the logistic positive exponent family (LPEF; Samejima, 2000), for dichotomous responses, which arrange their MLEs consistently following one principle concerning penalties or credits for failing or succeeding in answering easier or more difficult items. This was later expanded to a graded response model (LPEFG; Samejima, 2008) that will be introduced later in this chapter.

In spite of some shortcomings of the normal ogive and logistic models for dichotomous responses, they are useful models as working hypotheses, and effectively used, for example, in on-line item calibration in computerized adaptive testing (cf. Samejima, 2001).

---

Models in the Heterogeneous Case

The heterogeneous case consists of all specific latent trait models for graded responses that do not belong to the homogeneous case. In each of those models, the COCs, for \( x = 1,2,...,m \), are not all identical in shape, unlike those models in the homogeneous case, and yet the relationship in Equation 4.5 holds for every pair of adjacent \( x_j \). That is, even though adjacent functions are not parallel, they never cross.

Two subcategories are conceivable for specific graded response models in the heterogeneous case. One is a subgroup of those models that can be naturally expanded to continuous response models. In this section, the graded response model (LPEFG), which was expanded from the logistic positive exponent family (LPEF) of models for dichotomous responses, and the acceleration model, which was specifically developed for elaborate
cognitive diagnosis, are described and discussed. The other subcategory contains those models that are *discrete in nature*, which are represented by those models expanded from Bock’s (1972) nominal response model (BCK-SMJ).

**Models Expanded from the Logistic Positive Exponent Family of Models**

**Logistic Positive Exponent Family of Models for Dichotomous Responses (LPEF)**

This family previously appeared in Samejima’s (1969) *Psychometrika* monograph, using the normal ogive function instead of the logistic function in ICFs, although at that time it was premature for readers and practically impossible to pursue the topic and publish in refereed journals.

As was exemplified earlier by the NMLOG, if the ICF is point symmetric, there is no systematic principle of ordering the values of MLE obtained on response patterns, as is seen in the example of the normal ogive model. The logistic model, where a simple sufficient statistic $\sum_{u,v} a_{u,v} u v$ (Birnbaum, 1968) exists, is an exception, and there the ordering depends solely on the discrimination parameters, $a$, without being affected by the difficulty parameters, $b$.

The strong motivation for the LPEF was to identify a model that arranged the values of MLE, that is, all possible response patterns consistently following a single principle, that is, penalizing or crediting incorrect or correct responses, respectively. This motivation was combined with an idea to perceive Birnbaum’s logistic model as a transition model in a family of models. After all, the fact that in the logistic model MLEs are determined from the sufficient statistic (Birnbaum, 1968) that disregards, totally, the difficulty parameters, $b$, and is solely determined by the discrimination parameters, $a$, is not easily acceptable to this researcher’s intuition. This led to the family of models called the logistic positive exponent family (LPEF) (Samejima, 2000), where ICFs are defined by

$$P_g(\theta) = [\Psi_g(\theta)]^g \quad 0 < \xi_g < \infty,$$

where the third parameter, $\xi_g$, is called the *acceleration* parameter, and

$$\Psi_g(\theta) = [1 + \exp(-Da_g(\theta - b_g))]^{-1}$$

the right-hand side of which is identical to the logistic ICF (Birnbaum, 1968), where the scaling factor $D$ is usually set equal to 1.702. Note that Equation 4.26 also becomes the logistic ICF when $\xi_g = 1$, that is, a point-symmetric curve, and otherwise, it provides point-asymmetric curves, having a long tail on lower levels of $\theta$ as $\xi_g$ ($< 1$) gets smaller. Samejima (2000) explains that when $\xi_g < 1$ the model arranges the values of MLE following the principle that penalizes the failure in solving as easier item, and when $\xi_g < 1$, following the principle that gives credit for solving a more difficult item, provided that the discrimination parameters assume the same value for all items (cf. Samejima, 2000). Thus Birnbaum’s logistic model can be considered to represent the transition between the two opposing principles.
Figure 4.4 illustrates the ICFs of models that belong to the LPEF, with the common item parameters $a_g = 1$ and $b_g = 0$, where the values of $\xi_g$ are 0.3, 0.5, 0.8, 1.0, 1.5, 2.0, and 3.0, respectively.

The characteristics of the LPEF are as follows:

1. If $0 < \xi_g < 1$, then the principle arranging the MLEs of $\theta$ is that failure in answering an easier item correctly is penalized (success in answering an easier item correctly is credited).
2. If $1 < \xi_g < \infty$, then the principle arranging the MLEs of $\theta$ is that success in answering a more difficult item is credited (failure in answering a more difficult item correctly is penalized).
3. When $\xi_g = 1$, both of the above principles degenerate, and neither of the two principles works.

The reader is directed to Samejima (2000) for detailed explanations and observations of the LPEF for dichotomous responses. It is especially important to understand the role of the model/item feature function, $S_g(\theta)$, which is defined in that article in Equation 7, specified by Equation 28 for the LPEF, and illustrated in Figures 4(a) to (c) (pp. 331–332).

It should be noted that the item parameters, $a_g$ and $b_g$, in the LPEF should not be considered as the discrimination and difficulty parameters. (The same is also true with the 3PL.) Actually, the original meaning of the difficulty parameter is the value of $\theta$ at which $P_g(\theta) = 0.5$. These values are indicated in Figure 4.4, where they are strictly increasing with $\xi_g$, not constant for all items. Also, the original meaning of the discrimination parameter is a parameter proportional to the slope of $P_g(\theta)$ at the level of $\theta$ where $P_g(\theta) = 0.5$, and it is also strictly increasing with $\xi_g$, not a constant value for all seven items.
Graded Response Model Expanded from the Logistic Positive Exponent Family of Models (LPEFG)

It can also be seen in Figure 4.4 that whenever \( \xi_g < \xi_h \) for a pair of arbitrary items, \( g \) and \( h \), there is a relationship that \( P_g(\theta) > P_h(\theta) \) for the entire range of \( \theta \). The reason is because for any value \( 0 < R < 1, R' > R \) holds for any \( s < t \).

The ICFs that are illustrated in Figure 4.4 can be used, therefore, for an example of a set of COCs for graded item scores for a single graded response item with \( m_g = 7 \) that satisfies Equation 4.5. More formally, the LPEFG is characterized by

\[
M_{x_g}(\theta) = [\Psi_g(\theta)]^{\xi_g - \xi_{x_g - 1}} \quad x_g = 0, 1, 2, ..., m_g, \quad (4.28)
\]

where \( \Psi_g(\theta) \) is given by Equation 4.27 and

\[
\xi_{x_g} \equiv 0 = \xi_0 < \xi_1 < ... < \xi_x < ... < \xi_{m_g} < \xi_{m_g + 1} \equiv \infty, \quad (4.29)
\]

which leads to

\[
P_{x_g}^*(\theta) = [\Psi_g(\theta)]^{\xi_{x_g}}, \quad (4.30)
\]

due to Equations 4.28 and 4.1. From Equations 4.2 and 4.28 through 4.30 the OC in the LPEFG is defined by

\[
P_{x_g}(\theta) = [\Psi_g(\theta)]^{\xi_{x_g}} - [\Psi_g(\theta)]^{\xi_{x_g + 1}}, \quad (4.31)
\]

and all the other functions, \( A_{x_g}(\theta), I_{x_g}(\theta), I_{x_g}(\theta), \) and \( I(\theta) \), can be obtained by replacing \( P_{x_g}(\theta) \) by the right-hand side of Equation 4.22 and evaluating its derivatives in Equations 4.3, 4.4, 4.6, and 4.19, and substituting these outcomes into Equations 4.2 through 4.4 (details in Samejima, 2008).

Figure 4.5a and b presents the OCs and BSFs (per Equation 4.3) of an example of the LPEFG, with \( m_g = 5, a_g = 1, b_g = 0, \) and \( \xi_{x_g} = 0, 0.3, 0.8, 1.6, 3.1, \) and 6.1, respectively. Note that, for \( x_g = 0 \) and \( x_g = m_g \), OCs are strictly decreasing and increasing in \( \theta \), respectively, and for all the other graded item scores they are unimodal, and the BSFs are all strictly decreasing in \( \theta \), with the upper asymptotes zero for \( x_g = 0 \) and \( Da_{x_g} \xi_{x_g} \) for \( x_g = 1, 2, 3, 4, 5 \), respectively, with the lower asymptotes \(-Da_{x_g} \) for \( x_g = 0, 1, 2, 3, 4 \) and zero for \( x_g = 5 \), indicating the satisfaction of the unique maximum condition. The set of BSFs in Figure 4.5b is quite different from that of NMLOG or LGST (see Samejima, 1969, 1972) because the upper limit of the BSF is largely controlled by the item response parameter \( \xi_{x_g} \). Because of this fact, it
FIGURE 4.5 Operating characteristics (a) and basic functions (b) of an example six-category item modeled with the LPEFG. In (a), the modal points of OCs are produced according to $x_g$'s, i.e., the lowest is $-\infty$ for $x_g = 0$, and the highest is $\infty$ for $x_g = 5$. In (b), the asymptotes when $\theta$ approaches $-\infty$ are ordered, i.e., lowest for $x_g = 0$ and highest for $x_g = 5$. 

$\text{Latent Trait } \theta$

(a)

$\text{Latent Trait}$

(b)
can also be seen that the amount of information shown in the IRIF becomes larger as the item score $x_{\xi}$ gets larger. The value of $\theta$ at which each curve in Figure 4.5b crosses the $\theta$-dimension for each $x_{\xi}$ indicates the modal point of the corresponding OC that is seen in Figure 4.5a, and these modal points are ordered in accordance with the $x_{\xi}$'s, with the terminal maximum at negative infinity and positive infinity for $x_{\xi} = 0$ and $x_{\xi} = m_{\xi}$ ($= 5$), respectively.

Another set of PRFs, COCs, OCs, BSFs, and IRIFs for the LPEFG with different parameter values can be found in Samejima (2008).

It was noted above that the LPEFG satisfies the unique maximum condition, as illustrated in Figure 4.5b. In addition, both additivity and expandability to a continuous response model (discussed above) are intrinsic in the LPEFG (Samejima, 2008).

**LPEFG as a Substantive Mathematical Model**

It is noted that, unlike the normal ogive or the logistic model, LPEFG is a substantive mathematical model, in the sense that the principle and nature of the model support, consistently, certain psychological phenomena. To give an example, for answering a relatively difficult problem-solving question, we must successfully follow a sequence of cognitive processes. The individual’s performance can be evaluated by the number of processes in the sequence that he or she has successfully cleared, and one of the graded item scores, 0 through $m_{\xi}$, is assigned. It is reasonable to assume that passing up to each successive cognitive process becomes progressively more difficult, represented by the item response parameter $\xi_{\xi}$. Concrete examples of problem solving, for which the LPEFG is likely to fit, are various geometric proofs (Samejima, 2008).

Usually, there is more than one way of proving a given geometry theorem. Notably, it is said that there are 362 different ways to prove the Pythagoras theorem! It would make an interesting project for a researcher to choose a geometry theorem having several proofs, collect data, categorize subjects into subgroups, each of which consists of those who choose one of the different proofs, assign graded item scores to represent the degrees of attainment for each subgroup, and apply LPEFG for the data of each subgroup. It is most likely that separate proofs will have different values of $m_{\xi}$, and it would be interesting to observe the empirical outcomes.

Readers will be able to think of other substantive examples. It would be most interesting to see applications of the LPEFG to data collected for such examples to find out if the model works well. Any such feedback would be appreciated.

**Relationship to Other Chapters:** Huang and Mislevy do something similar to what is suggested here with responses to a physical mechanics exam. However, they use the polytomous Rasch model to investigate response strategies rather than the LPEFG, and take a slightly different approach given that Rasch models do not model processing functions.
Acceleration Model (ACLR)

**Greater Opportunities for Applying Mathematical Models for Cognitive Psychology Data**

For any research in the social sciences, mathematical models and methodologies are important, if one aims at truly scientific accomplishments. Because of the intangible nature of the social sciences, however, there still is a long way to go if the levels of scientific attainments in natural sciences are one’s goal.

Nonetheless, the research environment for behavioral science has been improved, especially during the past few decades. One of the big reasons for the improvement is advancement in computer technologies. To give an example, in cognitive psychology it used to be typical that a researcher invited a subject to an experimental room and gave him or her instructions, which the subject followed and responded to accordingly. Because of its time-consuming nature, it was very usual that research was based on a very small group of subjects and quantitative analysis of the research data was practically impossible.

With the rapid advancement of computer technologies, microcomputers have become much more capable, smaller, and much less expensive. It is quite possible to replace the old procedure by computer software that accommodates all experimental procedures, including instructions, response formats, and data collections. The software is easy to copy, and identical software can be installed onto multiple laptops of the same type, each of which can be taken by well-trained instructors to different geographical areas to collect data for dozens of subjects each. Thus, data can be collected with a sample size of several hundred relatively easily, in a well-controlled experimental environment. Sampling can also be made closer to random sampling.

In return, the need for mathematical models for cognitive processes has become greater, and one must propose mathematical models with the above perspective.

**Acceleration Model**

Samejima (1995) proposed the acceleration model that belongs to the heterogeneous case with such a future need in mind. In general, cognitive diagnosis is complicated, so naturally models for cognitive diagnosis must be more complicated than many other mathematical models that are applied, for example, to test or questionnaire data.

In the acceleration model, the PRF is defined by

\[ M_{x_g} (\theta) = \left[ \Psi'_{x_g} (\theta) \right]^{\xi_{x_g}}, \quad (4.32) \]

where \( \xi_{x_g} > 0 \) is also called the acceleration parameter in this model, and \( \Psi_{x_g} (\theta) \) is a member of the family of functions satisfying

\[ \xi_{x_g} = 1 - \left[ \Psi'_{x_g} (\theta) \Psi''_{x_g} (\theta) \right] \left[ \Psi'_{x_g} (\theta) \right]^2 \quad (4.33) \]
that includes the logistic function such that

\[ \Psi_{x_g}(\theta) = \left[ 1 + \exp \left( -D\sigma_{x_g} (\theta - b_{x_g}) \right) \right]^{-1}. \]  

(4.34)

In Samejima (1995) Equation 4.33 is mostly used in Equation 4.32. The COC in this model is provided by

\[ P_{x_g}^*(\theta) = \left[ \Psi_{x_g}(\theta) \right]^{\sum_{x_g} x_g}. \]  

(4.35)

and all the other functions, such as OC, BSF, and IRIF in this model, are given by substituting Equation 4.35 into those formulas of the general graded response model, Equations 4.2 through 4.4, respectively.

It should be noted that on the left-hand side of Equation 4.34 \( \Psi_{x_g}(\theta) \) is used instead of \( \Psi_g(\theta) \) in Equation 4.27, with \( a_{x_g} \) and \( b_{x_g} \) replacing \( a_g \) and \( b_g \), respectively, on the right-hand side. This indicates that in the acceleration model the logistic function is defined separately for each graded score \( x_g \) while in the LPEFG it is common for all the graded item scores of item \( g \). This difference makes the acceleration model more complicated than the LPEFG for the purpose of using it for cognitive diagnosis of more complicated sequences of cognitive processes.

It is also noted that if \( \Psi_{x_g}(\theta) \) in Equation 4.34 is replaced by \( \Psi_g(\theta) \) in Equation 4.27, and we define \( \xi_{x_g} \equiv \xi_{x_g} - \xi_{x_g + 1} \) and \( \xi_{x_g} \equiv 0 \), then the LPEFG can be considered as a special, simplified case of the acceleration model. The model is described in detail in Samejima (1995). It may be wise to collect data to which the LPEFG substantively fits, and analyze them first, building on that experience to analyze more elaborate cognitive data using the acceleration model.

**Bock’s Nominal Model Expanded to a Graded Response Model (BCK-SMJ)**

Bock’s (1972) nominal response model is a valuable model for nominal response items in that it discloses the implicit order of the nominal response categories. Samejima (1972) proposed a graded response model expanded from Bock’s nominal model. When a model fits data that implicitly have ordered response categories, it is easy to expand the model to a graded response model for the explicit graded item scores.

Samejima did not pursue BCK-SMJ much further, however, because an intrinsic restriction was observed in the expanded model. Later, Masters (1982) proposed a special case of BCK-SMJ as the partial credit model and Muraki (1992) proposed BCK-SMJ itself as a generalized partial credit model without realizing that the model had already been proposed in 1972. Many researchers have applied those models. Practitioners using IRT in their research should only use either model when their research data are within the limit of the previously identified restriction, however.
The OC in the BCK-SMJ is given by

\[ P_{x_x}(\theta) = \exp \left[ \alpha_{x_x} \theta + \beta_{x_x} \right] \left[ \sum_{u \in x_x} \exp \left[ \alpha_u \theta + \beta_u \right] \right]^{-1} \]  \hspace{1cm} (4.36)

with

\[ 0 < \alpha_0 < \alpha_1 < ... < \alpha_{m_x} < \infty. \]

It is noted that the denominator of Equation 4.36 is common for all \( x_g \)'s. This makes the conditional ratio, given \( \theta \), of any pair of the OCs for \( x_s = s \) and \( x_t = t \) (\( s \neq t \)) such that

\[ P_s(\theta) P_t(\theta)^{-1} = \exp[\alpha_s - \alpha_t] \theta \exp[\beta_s - \beta_t] \]  \hspace{1cm} (4.37)

indicating the invariance of this conditional ratio, which characterizes Bock's nominal response model. The same characteristic, however, becomes a restriction for the BCK-SMJ model. When \( s \) and \( t \) are two arbitrary adjacent graded item scores, for example, the combined graded response category will have the OC

\[ P_{s+t}(\theta) = \left[ \exp[\alpha_s \theta + \beta_s] + \exp[\alpha_t \theta + \beta_t] \right] \left[ \sum_{u \in x_x} \exp[\alpha_u \theta + \beta_u] \right]^{-1}. \]  \hspace{1cm} (4.38)

It is obvious that Equation 4.38 does not belong to Equation 4.36, and thus additivity 2 does not hold for the BCK-SMJ. It can also be seen that additivity 1 does not hold for the model either. Thus BCK-SMJ is discrete in nature, and cannot be naturally expanded to a continuous response model, unlike the normal ogive model, logistic model, and LPEFG. It should be applied strictly for data that are collected for a fixed set of graded response categories, where no recategorizations are legitimate. This is a strong restriction.

A summary of the characteristics of the five specific graded response models discussed above, with respect to the four evaluation criteria that were discussed earlier, is given in Table 4.1.

<table>
<thead>
<tr>
<th>TABLE 4.1</th>
<th>Summary of the Characteristics of the Specific Graded Response Model With Respect to the Four Evaluation Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NMLOG</td>
</tr>
<tr>
<td>Additivity 1</td>
<td>Yes</td>
</tr>
<tr>
<td>Additivity 2</td>
<td>Yes</td>
</tr>
<tr>
<td>Expands to CRM</td>
<td>Yes</td>
</tr>
<tr>
<td>Satisfies unique maximum condition</td>
<td>Yes</td>
</tr>
<tr>
<td>Ordered modal points</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The Importance of Nonparametric Estimation

Failure in Parametric Estimation of Item Parameters in the Three-Parameter Logistic Model

Quite often researchers, using simulated data for multiple-choice items, adopt software for a parametric estimation of the three-parameter logistic (3PL) model (Birnbaum, 1968), where the ICF is defined as

\[ P_{\theta} (\theta) = \epsilon_{\theta} + (1 - \epsilon_{\theta})[1 + \exp\{-Da_{\theta} (\theta - b_{\theta})\}]^{-1} \]  

(4.39)

and fail to recover the values of three parameters, \( a_{\theta}, b_{\theta}, \) and \( \epsilon_{\theta} \), within the range of error. Sometimes all the item parameter estimates are outrageously different from their true values. This is a predictable result because in most cases simulated data are based on a mound-shaped ability distribution with very low densities for very high and very low levels of \( \theta \). In Equation 4.39 estimating the third parameter, \( \epsilon_{\theta} \), which is the lower asymptote of the ICF, will naturally be inaccurate. This inaccuracy will also affect, negatively, the accuracies in estimating the other two parameters. Moreover, even if the ability distribution has large densities on lower levels of \( \theta \), when \( \theta \) is treated as an individual parameter and if an EM algorithm is used to estimate both the individual parameter and item parameters, then the more hypothetical individuals at lower levels of \( \theta \) are included, the larger the amount of estimation error of individual parameters will occur, influencing accuracy in estimating \( \epsilon_{\theta} \) negatively and, consequently, in \( a_{\theta} \) and \( b_{\theta} \). Thus such an attempt is doomed to fail.

Even without making an additional effort to increase the number of subjects at lower levels of the latent trait \( \theta \) to more accurately recover the three item parameters, which will not in any event be successful, if the true curve and the estimated curve with outrageously wrong estimated parameter values are plotted together, the fit of the curve with the estimated parameter values to the true curve is usually quite good for the interval of \( \theta \) at which densities of ability distribution are high. We could say that, although the parametric estimation method aims at the recovery of item parameters, it actually recovers the shape of the true curve for that interval of \( \theta \), as a well-developed nonparametric estimation method does, not the item parameters themselves.

Nonparametric Estimation of OCCs

From a truly scientific standpoint, parametric estimation of OCCs is not acceptable unless there is evidence to justify the adoption of the model in question, because if the model does not fit the nature of our data, it "molds" the research data into a wrong mathematical form and the outcomes of research will become meaningless and misleading.

Thus, well-developed nonparametric estimation methods that will discover the shapes of OCCs will be valuable. Lord developed such a nonparametric estimation method, and applied it for estimating the ICFs of Scholastic
Aptitude Test items (Lord, 1980: Figure 2.31 on page 16, for example outcomes). The method is appropriate for a large set of data, represented by widely used tests that are developed and administered by the Educational Testing Service, American College Testing, and Law School Admission Council, for example, but it is not appropriate for data of relatively small sizes, such as those collected in a college or university environment. The nonparametric methods that were developed by Levine (1984), Ramsay (1991), and Samejima (1998, 2001) will be more appropriate to use for data collected on a relatively small number of individuals.

Figure 4.6 exemplifies the outcomes obtained by Samejima's (1998, 2001) simple sum procedure (SSP) and differential weight procedure (DWP) of the conditional probability density function (pdf) approach, based on the simulated data of 1,202 hypothetical examinees in computerized adaptive testing (CAT). The outcome of DWP1 (thin, solid line) was obtained by using the outcome of SSP (dashed line) as the differential weight function, while the fourth curve is the result of DWP using the true curve (thick, solid line) as the differential weight function. The DWP_True is called the criterion operating characteristic (dashed line), indicating the limit of the closeness of an estimated curve to the true curve; if they are not close enough, either the procedures in the method of estimation should be improved, or the sample size should be increased.

It should be noted that the nonmonotonicity of the true curve is detected by both the SSP and DWP1 in Figure 4.6. Even if the true curve is nonmonotonic, which is quite possible, especially for the item characteristic function of a multiple-choice test item (Samejima, 1979), such detection

**FIGURE 4.6** A non-monotonic ICF (TRUE), its two nonparametric estimates (SSP, DWP1), and the criterion ICF (DWP_true).
cannot be made by a parametric estimation. If, for example, the true curve in Figure 4.6 is the ICF of a multiple-choice test item and a parametric estimation method such as the 3PL is used, the estimated three item parameters will provide, at best, an estimated curve with a monotonic tail.

There is no reason to throw away an item whose ICF is nonmonotonic, as illustrated in Figure 4.6. It is noted that approximately for the interval of \( \theta \) (0.0, 1.5) the amount of item information at each value of \( \theta \) is large, so there is no reason why we should not take advantage of it. On the other hand, on levels lower than this interval of \( \theta \) the nonmonotonicity of the curve will make the IRIFs negative, so this part of the curve should not be used. Samejima (1973b) pointed out that the IRIF of the 3PL for \( u_f = 1 \) assumes negative value, and for that reason, 3PL does not satisfy the unique maximum condition. Using an item whose ICF is nonmonotonic, as illustrated by Figure 4.6, is especially easy in CAT (Samejima, 2001), but a similar method can be used in a paper-and-pencil test or questionnaire.

In Figure 4.6 it can be seen that (1) the outcome of DWP1 is a little closer to the criterion operating characteristic than that of SSP, (2) the outcomes of SSP and DWP1 are both very close to the criterion operating characteristic, and (3) the criterion operating characteristic is very close to the true curve. For more details, the reader is directed to Samejima (1998, 2001).

Samejima (1994) also used the SSP on empirical data, for estimating the conditional probability, given \( \theta \), of each distractor of the multiple-choice items of the Level 11 Vocabulary Test of the Iowa Test of Basic Skills, and called those functions of the incorrect answers plausibility functions. It turned out that quite a few items proved to possess plausibility functions that have differential information, and the use of those functions in addition to the ICFs proved to be promising for increasing the accuracy of ability estimation.

In the example of Figure 4.6, nonparametric estimation of the ICFs for dichotomous items, or that of the COCs for graded responses, was considered. We could estimate PRFs or OCs first, however. Equations 4.1 and 4.2 can be changed to

\[
M_{x_\theta} (\theta) = P_{x_\theta} (\theta) \left[ P_{x_{\theta-1}} (\theta) \right]^{-1} \quad \text{for} \quad x_\theta = 1, 2, ..., m_\theta \quad (4.40)
\]

and

\[
P_{x_\theta} (\theta) = \sum_{x_{\theta-1} = m_{\theta-1}} P_{x_{\theta-1}} (\theta) \quad \text{for} \quad x_\theta = 1, 2, ..., m_\theta \quad (4.41)
\]

respectively. Nonparametric estimation can be performed to discover the PRFs, first, and using those outcomes the COCs and then the OCs can be obtained through Equations 4.1 and 4.2. An alternative way is to estimate the COCs first, and using Equation 4.40, the PRFs can be obtained, and then the OCs through Equation 4.2. It is possible to estimate the OCs first,
and then using the outcomes, the COCs can be obtained through Equation 4.41, and then the PRFs through Equation 4.40. Note, however, this last method may include substantial amounts of error, because it is quite possible that some graded score may include only a small number of individuals, unless the total sample size is large enough.

In any case, after the shapes of those functions are nonparametrically estimated, it is wise to parameterize the nonparametrically discovered functions selecting a parametric model that is legitimate in principle and agrees with the nature of the data. Otherwise, it is difficult to proceed in research using functions with no mathematical forms.

**Limitation of Curve Fittings in Model Validation and Selection**

While the goodness of fit of curves is important, it has its limitations, especially when a model belongs to the heterogeneous case where it has less restrictions and more freedom for innovation. Samejima (1996a, 1997) demonstrated two sets of OCs that belong to the two specific graded response models of quite different principles, the ACLR and BCK-SMJ, which are nevertheless practically identical to each other. This means that if the OCs that are discovered as the outcomes of a nonparametric estimation method fit the OCs of the ACLR, they should also fit those of the BCK-SMJ.

Curve fitting alone cannot be a good enough criterion for model validation. In model selection, in addition to curve fitting, the most important consideration should be how well the principle behind each model agrees with the nature of the research data. Furthermore, considerations should be made of whether each of the other four criteria listed in Table 4.1 fits the model, as well as the research data. For example, if we know that the results of our research may be compared with other research on the same or similar contents, mathematical models that lack additivity should be avoided. If continuous responses are used, a model should be chosen that can be expanded naturally from a graded response model in order to make future comparisons possible with the outcomes of other research in which graded responses are used.

An effort to select a substantive model is by far the most important criterion, and curve fitting can be used as an additional criterion, to see if those curves in a substantive model provide at least reasonably good fit to the data.

**Conclusion**

IRT has developed so much in the past few decades that it is hard to write even just the essential elements of the general graded response model framework as a handbook chapter. Many important and useful topics have been omitted. An attempt has been made to include useful hints for researchers and practitioners in applying IRT within this chapter, including suggested readings. But even with reference to the original work cited in this chapter, it may be difficult to identify ways to apply the models.
Face-to-face workshops may be a useful way to supplement the written and cited material in this chapter. Such opportunities would make interactive communications and deeper understanding possible.

References


This page intentionally left blank