5
Graded Response Model

Fumiko Samejima

Introduction

The graded response model represents a family of mathematical models that deals with ordered polytomous categories. These ordered categories include rating such as letter grading, A, B, C, D, and F, used in the evaluation of students' performance; strongly disagree, disagree, agree, and strongly agree, used in attitude surveys; or partial credit given in accordance with an examinee's degree of attainment in solving a problem.

Presentation of the Model

Let θ be the latent trait, or ability, which represents a hypothetical construct underlying certain human behavior, such as problem-solving ability, and this ability is assumed to take on any real number. Let i denote an item, which is the smallest unit for measuring θ. Let U_i be a random variable to denote the graded item response to item i, and u_i $(= 0, 1, \ldots, m_i)$ denote the actual responses. The category response function, $P_{u_i}(\theta)$, is the probability with which an examinee with ability θ receives a score u_i, that is,

$$P_{u_i}(\theta) \equiv \text{Prob}[U_i = u_i \mid \theta].$$

$P_{u_i}(\theta)$ will be assumed to be five times differentiable with respect to θ. For convenience, u_i will be used both for a specific discrete response and for the event $U_i = u_i$, and a similar usage is applied for other symbols.

For a set of n items, a response pattern, denoted by V, indicates a sequence of U_i for $i = 1, 2, \ldots, n$, and its realization v, can be written as

$$v = \{u_1, u_2, \ldots, u_i, \ldots, u_n\}'.$$

It is assumed that local independence (Lord and Novick, 1968) holds, so that within any group of examinees with the same value of ability θ the distributions of the item responses are independent of each other. Thus, the conditional probability, $P_v(\theta)$, given θ, for the response pattern v can

W. J. van der Linden et al. (eds.), *Handbook of Modern Item Response Theory*

be written as

$$P_v(\theta) \equiv \text{Prob}[V = v \mid \theta] = \prod_{u_i \in v} P_{u_i}(\theta), \tag{1}$$

which is also the likelihood function, $L(v \mid \theta)$, for $V = v$.

Samejima (1969) proposed the first graded response models: the normal-ogive model and the logistic model for graded response data (i.e., ordered polytomous categories). Later she proposed a broader framework for graded response models, distinguishing the *homogeneous* case, to which the normal-ogive and logistic models belong, and the *heterogeneous* case (Samejima, 1972).

General Graded Response Model

Suppose, for example, that a cognitive process, like problem solving, contains a finite number of steps. The graded item score u_i should be assigned to the examinees who successfully complete up to step u_i but fail to complete the step $(u_i + 1)$. Let $M_{u_i}(\theta)$ be the *processing function* (Samejima, 1995) of the graded item score u_i, which is the probability with which the examinee completes the step u_i successfully, under the joint conditions that (a) the examinee's ability level is θ and (b) the steps up to $(u_i - 1)$ have already been completed successfully. Let $(m_i + 1)$ be the next graded item score above m_i. Since everyone can at least obtain the item score 0, and no one is able to obtain the item score $(m_i + 1)$, it is reasonable to set

$$M_{u_i}(\theta) = \begin{cases} 1 & \text{for } u_i = 0 \\ 0 & \text{for } u_i = m_i + 1, \end{cases} \tag{2}$$

for all θ. For each of the other u_i's, it is assumed that $P_{u_i}(\theta)$ is nondecreasing in θ. This assumption is reasonable considering that each item has some direct and positive relation to the ability measured. Thus, the category response function, $P_{u_i}(\theta)$, of the graded item score u_i is given by

$$P_{u_i}(\theta) = \left[\prod_{s \leq u_i} M_s(\theta) \right] [1 - M_{(u_i+1)}(\theta)]. \tag{3}$$

This provides the fundamental framework for the general graded response model.

Let $P^*_{u_i}(\theta)$ denote the conditional probability with which the examinee of ability θ completes the cognitive process successfully up to step u_i, or further. Then,

$$P^*_{u_i}(\theta) = \prod_{s \leq u_i} M_s(\theta). \tag{4}$$

This function is called the *cumulative category response function* (Samejima, 1995), although cumulation is actually the usual convention. From

Eqs. (3) and (4), the category response function can also be expressed as

$$P_{u_i}(\theta) = P^*_{u_i}(\theta) - P^*_{(u_i+1)}(\theta). \tag{5}$$

Note that $P^*_{u_i}(\theta)$ becomes the response function, $P_i(\theta)$, of the positive response to item i, when the graded item score u_i is changed to the binary score, assigning 0 to all scores less than u_i and 1 to those score categories greater than or equal to u_i. It is obvious from Eqs. (2) and (4) that $P^*_{u_i}(\theta)$ is also nondecreasing in θ, and assumes unity for $u_i = 0$ and zero for $u_i = m_i + 1$ for the entire range of θ.

The general graded response model includes many specific mathematical models. In an effort to select a model interpretable within a particular psychological framework, it will be wise to examine whether the model has certain desirable features. Among others, a model should be examined as to whether (a) the principle behind the model and the set of assumptions agree with the psychological reality in question; (b) it satisfies the unique maximum condition (Samejima, 1969, 1972); (c) it provides the ordered modal points of the category response functions in accordance with the item scores; (d) additivity of the category response functions (Samejima, 1995) holds; and (e) the model can be naturally generalized to a continuous response model.

The unique maximum condition is satisfied if the *basic function*, $A_{u_i}(\theta)$, defined by

$$A_{u_i}(\theta) \equiv \frac{\partial}{\partial\theta} \log P_{u_i}(\theta) = \sum_{s \le u_i} \frac{\partial}{\partial\theta} \log M_s(\theta) + \frac{\partial}{\partial\theta} \log[1 - M_{(u_i+1)}(\theta)], \tag{6}$$

is strictly decreasing in θ and its upper asymptote is nonnegative and its lower asymptote is nonpositive (Samejima, 1969, 1972). Satisfaction of this condition ensures that the likelihood function of *any* response pattern consisting of such response categories has a unique local or terminal maximum. Using this basic function, a sufficient, though not necessary, condition for the strict orderliness of the modal points of the category response functions is that $A_{(u_i-1)}(\theta) < A_{u_i}(\theta)$ for all θ for $u_i = 1, 2, \ldots, m_i$.

Additivity of a model will hold if the category response functions still belong to the same mathematical model under finer recategorizations and combinations of two or more categories together, implying that the unique maximum condition will be satisfied by the resulting category response functions if it is satisfied by those of the original u_i's. Graded item scores, or partial credits, are more or less incidental. For example, sometimes letter grades, A, B, C, D, and F, are combined to pass–fail grades. Also, with the advancement of computer technologies, more abundant information can be obtained from an examinee's performance in computerized experiments, and thus finer recategorizations of the whole cognitive process are possible. Additivity of response categories and generalizability to a continuous model are, therefore, important criteria in evaluating models.

Homogeneous Case

The homogeneous case of the graded response model represents a family of models in which $P^*_{u_i}(\theta)$'s for $u_i = 1, 2, \ldots, m_i$ are identical in shape, and these m_i functions are positioned alongside the ability continuum in accordance with the item score u_i. Thus, it is obvious that additivity of the category response functions always holds for mathematical models that belong to the homogeneous case.

The *asymptotic basic function* $\tilde{A}_{u_i}(\theta)$ has been defined in the homogeneous case for $u_i = 1, 2, \ldots, m_i$ by

$$\tilde{A}_{u_i}(\theta) \equiv \lim_{\lambda_{(u_i+1)} \to \lambda_{u_i}} A_{u_i}(\theta) = \frac{\partial}{\partial \theta} \log \left[\frac{\partial}{\partial \theta} P^*_{u_i}(\theta) \right] - \frac{\frac{\partial^2}{\partial \theta^2} M_1(\theta - \lambda_{u_i})}{\frac{\partial}{\partial \theta} M_1(\theta - \lambda_{u_i})}, \quad (7)$$

with λ_{u_i} being zero for $u_i = 1$ and increases with u_i, which is identical in shape for all $u_i = 1, 2, \ldots, m_i$ except for the positions alongside the dimension θ (Samejima, 1969, 1972). Using this asymptotic basic function, Samejima (1972, 1995) demonstrated that, in the homogeneous case, the unique maximum condition is simplified: (a) the lower and upper asymptotes of $M_{u_i}(\theta)$ for $u_i = 1$ are zero and unity, respectively, (b) $\frac{\partial}{\partial \theta} \tilde{A}_{u_i}(\theta) < 0$ for the entire range of θ for an arbitrarily selected u_i, and (c) for this specific u_i, the upper and lower asymptotes of $\tilde{A}_{u_i}(\theta)$ have some positive and negative values, respectively.

When the unique maximum condition is satisfied, the mathematical model can be represented by

$$P^*_{u_i}(\theta) = \int_{-\infty}^{a_i(\theta - b_{u_i})} \psi(t)\, dt, \quad (8)$$

where the item discrimination parameter a_i is finite and positive, and the difficulty or location parameters, b_{u_i}'s, satisfy

$$-\infty = b_0 < b_1 < b_2 < \ldots < b_{m_i} < b_{m_i+1} = \infty,$$

and $\psi(\cdot)$ denotes a density function that is four times differentiable with respect to θ, and is unimodal with zero as its two asymptotes as θ tends to negative and positive infinities, and with a first derivative that does not assume zero except at the modal point.

In the homogeneous case, satisfaction of the unique maximum condition also implies:

1. A strict orderliness among the modal points of $P_{u_i}(\theta)$'s, for it can be shown that
$$A_{(u_i-1)}(\theta) < \tilde{A}_{u_i}(\theta) < A_{u_i}(\theta)$$
for $u_i = 1, 2, 3, \ldots, m_i$, throughout the whole range of θ (Samejima, 1972, 1995).

2. Additivity of the category response functions, for, even if two or more adjacent graded item scores are combined, or if a response category is more finely recategorized, the $\tilde{A}_{u_i}(\theta)$'s for the remaining u_i's will be unchanged, and those of the newly created response categories will have the same mathematical form as that of the $\tilde{A}_{u_i}(\theta)$'s for the original response categories.

3. A natural expansion of the model to a continuous response model by replacing u_i in Eq. (8) by z_i, which denotes a continuous item response to item i and assumes any real number between 0 and 1, and by defining the *operating density characteristic*

$$H_{z_i}(\theta) = \lim_{\Delta z_i \to 0} \frac{P^*_{z_i}(\theta) - P^*_{(z_i+\Delta z_i)}(\theta)}{\Delta z_i} = a_i \psi(a_i(\theta - b_{z_i})) \left[\frac{db_{z_i}}{dz_i}\right],$$

where b_{z_i} is the difficulty parameter for the continuous response z_i and is a strictly increasing function of z_i (Samejima, 1973). The basic function, $A_{z_i}(\theta)$, in the continuous response model is identical with the asymptotic base function $\tilde{A}_{u_i}(\theta)$ defined by Eq. (7) on the graded response level, with the replacement of u_i by z_i, and thus for these models the unique maximum condition is also satisfied on the continuous response level.

It is wise, therefore, to select a specific model from those which satisfy the unique maximum condition, if the fundamental assumptions in the model agree with the psychological reality reflected in the data.

Samejima (1969, 1972) demonstrated that both the normal-ogive model and the logistic model belong to this class of models. In these models, the category response function is given by

$$P_{u_i}(\theta) = \frac{1}{[2\pi]^{1/2}} \int_{a_i(\theta - b_{u_i+1})}^{a_i(\theta - b_{u_i})} \exp\left[\frac{-t^2}{2}\right] dt, \tag{9}$$

and

$$P_{u_i}(\theta) = \frac{\exp[-Da_i(\theta - b_{u_i+1})] - \exp[-Da_i(\theta - b_{u_i})]}{[1 + \exp[-Da_i(\theta - b_{u_i})]][1 + \exp[-Da_i(\theta - b_{u_i+1})]]}, \tag{10}$$

respectively. In both models, $M_{u_i}(\theta)$ is a strictly increasing function of θ with unity as its upper asymptote; the lower asymptote is zero in the former model, and $\exp[-Da_i(b_{u_i} - b_{u_i} - 1)]$ in the latter (see Fig. 5-2-1 in Samejima, 1972). The upper asymptote of the basic function $A_{u_i}(\theta)$ for $u_i = 1, 2, \ldots, m_i$ and the lower asymptote for $u_i = 0, 1, \ldots, m_i - 1$, are positive and negative infinities in the normal-ogive model, and Da_i and $-Da_i$ in the logistic model (Samejima, 1969, 1972).

Figure 1 illustrates the category response functions in the normal-ogive model for $m_i = 4$, $a_i = 1.0$ and b_{u_i}'s equal to: -2.0, -1.0, 0.7, 2.0 for $u_i = 1$

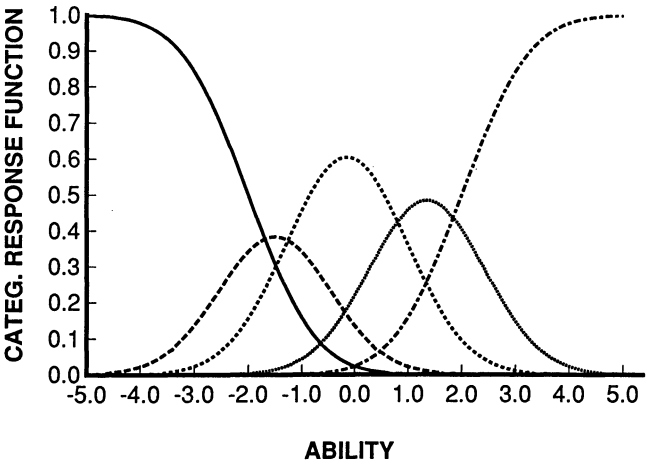

FIGURE 1. Example of a set of category response functions in the normal-ogive model for the item scores, 0, 1, 2, 3, and 4.

2, 3, 4, respectively. The corresponding category response functions in the logistic model with $D = 1.7$ are very similar to those in the normal-ogive model with the identical set of modal points; those curves are a little more peaked, however.

In the normal-ogive model for continuous response data, there exists a simple sufficient statistic, $t(v)$, for a specific response pattern v, and is provided by

$$t(v) = \sum_{z_i \in v} a_i^2 b_{z_i}, \tag{11}$$

so that the maximum likelihood estimate $\hat{\theta}$ is obtained directly from this simple sufficient statistic (Samejima, 1973). This fact suggests that, when m_i is substantially large, the normal-ogive model for continuous response data can be used as an approximation to the model for graded response data. In so doing, the value of m_i should be perhaps nine or greater.

Heterogeneous Case

The heterogeneous case of the graded response model represents all mathematical models that provide a set of cumulative category response functions $P_{u_i}^*(\theta)$'s *not all* of which are identical in shape, that is, those which do not belong to the homogeneous case. One example is Bock's nominal response model (Bock, 1972), represented by

$$P_{h_i}(\theta) = \frac{\exp[\alpha_{h_i}\theta + \beta_{h_i}]}{\sum_{s \in H_i} \exp[\alpha_s \theta + \beta_s]}, \tag{12}$$

where h_i denotes a nominal response to item i and H_i is the set of all h_i's, and α_{h_i} (> 0) and β_{h_i} are item category parameters. Samejima (1972) demonstrated that Bock's nominal response model can be considered as a graded response model in the heterogeneous case, if the nominal response h_i in Eq. (12) is replaced by the graded item response u_i and the parameter α_{u_i} satisfies

$$\alpha_0 \leq \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_{m_i}, \tag{13}$$

where a strict inequality should hold for at least one pair of α values. Two examples are Masters' (1982; this volume) Partial Credit Model (PCM) and Muraki's (1992; this volume) Generalized Partial Credit Model (GPCM). This family of models satisfies the unique maximum condition, and the perfect orderliness of the modal points of the category response functions is realized with strict inequality held between every pair of α_{u_i}'s in Eq. (13). However, it does not have additivity or generalizability to a continuous response model, and thus applicabilities are limited because of the assumption used in Bock's model (Samejima, 1995).

Samejima (1995) proposed the *acceleration model* that belongs to the heterogeneous case of the graded response model. Consider a situation, such as problem solving, that requires a number of subprocesses be completed correctly before attaining the solution. It is assumed that there is more than one step in the process and the steps are observable. Graded item scores, or partial credits, 1 through m_i, are assigned for the successful completion of these separate observable steps. The processing function for each u_i $(= 1, 2, \ldots, m_i)$ is given by

$$M_{u_i}(\theta) = [\Psi_{u_i}(\theta)]^{\xi_{u_i}}, \tag{14}$$

where ξ_{u_i} (> 0) is called the *step acceleration parameter*. The acceleration model is a family of models in which $\Psi_{u_i}(\theta)$ is specified by a strictly increasing, five times differentiable function of θ with zero and unity as its two asymptotes. Here a specific model in this family will be introduced, in which $\Psi_{u_i}(\theta)$ is given by

$$\Psi_{u_i}(\theta) = \frac{1}{1 + \exp[-D\alpha_{u_i}(\theta - \beta_{u_i})]}, \tag{15}$$

where $D = 1.7$ and α_{u_i} (> 0) and β_{u_i} are the discrimination and location parameters, respectively. It is assumed that the process leading to a problem solution consists of a finite number of *clusters*, each containing one or more steps, and within each cluster the parameters α and β in the logistic distribution function are common. Thus, if two or more adjacent u_i's belong to the same cluster, then the parameters α_{u_i}'s and β_{u_i}'s are the same for these u_i's, and, otherwise, at least one parameter is different.

It can be seen from Eqs. (14) and (15) that the roles of the step acceleration parameter, ξ_{u_i}, are to control (a) the general shape of the curve representing $M_{u_i}(\theta)$, (b) the steepness of the curve, and (c) the position

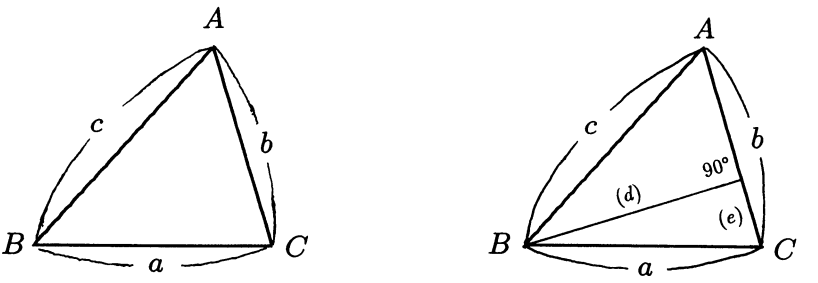

FIGURE 2. Triangle provided for the proof of the law of cosine (left), and the one after the perpendicular line has been drawn by the examinee (right).

of the curve alongside the ability continuum. When $\Psi_{u_i}(\theta)$ is given by Eq. (15), $M_{u_i}(\theta)$ is point-symmetric only when $\xi_{u_i} = 1$, and point-asymmetric otherwise. Let θ_{stp} denote the point of θ at which $M_{u_i}(\theta)$ is steepest. This is provided by

$$\theta_{\text{stp}} = \Psi_{u_i}^{-1}\left[\frac{\xi_{u_1}}{1+\xi_{u_i}}\right],$$

which increases with ξ_{u_i}. This maximum slope at $\theta = \theta_{\text{stp}}$ equals $Da_i[\xi_{u_i}/(1+\xi_{u_1})]^{\xi_{u_i}+1}$, which is also an increasing function of ξ_{u_i}. The value of $M_{u_i}(\theta)$ at $\theta = \theta_{\text{stp}}$ is $[\xi_{u_i}/(1+\xi_{u_i})]^{\xi_{u_i}}$, which decreases with ξ_{u_i}. The general position of the curve representing $M_{u_i}(\theta)$ is shifted to the positive side alongside the continuum as ξ_{u_i} increases, and in the limiting situation when ξ_{u_i} tends to positive infinity $M_{u_i}(\theta)$ approaches 0 for all θ, whereas in the other limiting situation where ξ_{u_i} tends to zero $M_{u_i}(\theta)$ approaches 1 for all θ.

Let w denote a subprocess, which is the smallest unit in the cognitive process. Thus, each step contains one or more w's. Let ξ_w (> 0) represent the *subprocess acceleration* parameter, and then the step acceleration parameter, ξ_{u_i}, for each of $u_i = 1, 2, \ldots, m_i$ is given as the sum of ξ_w's over all $w \in u_i$. The name, *acceleration parameter*, derives from the fact that, within each step, separate subprocesses contribute to accelerate the value of θ at which the discrimination power is maximal (Samejima, 1995).

Consider the following problem-solving example:

Using the triangle shown on the left-hand side of Fig. 2, prove that

$$a^2 = b^2 + c^2 - 2bc\cos A.$$

Five observable steps in solving this particular problem may be as shown below:

$a^2 = d^2 + e^2$ Step 1 [Pythagoras' theorem]

$a^2 = (c\sin A)^2 + (b - c\cos A)^2$ Step 2 [sine and cosine]

$$a^2 = c^2 \sin^2 A + b^2 - 2bc\cos A + c^2 \cos^2 A \quad \text{Step 3 } [(a-b)^2 = a^2 - 2ab_b^2]$$

$$a^2 = b^2 + c^2(\sin^2 A + \cos^2 A) - 2bc\cos A \quad \text{Step 4 } [ab + ac = a(b+c)]$$

$$a^2 = b^2 + c^2 - 2bc\cos A \quad \text{Step 5 } [\sin^2 A + \cos^2 A = 1],$$

where d and e are as shown in the right-hand side diagram of Fig. 2.

Each of the steps contain more than one substep or subprocess except for Step 5. For example, Step 1 includes three subprocesses, that is, (a) thinking of the use of the Pythagoras' theorem, (b) drawing the perpendicular line from B to b (or, alternatively, from C to c), and naming the perpendicular line d and the lower line sigment of be, and (c) applying the Pythagoras' theorem to the right triangle that includes C; Step 2 includes two subprocesses: the evaluation of d as $(c \sin A)$ and that of e as $(b - c\cos A)$.

It should be noted that in Eq. (15) the location parameter, β_{u_i}, does not necessarily increase with u_i. To illustrate an extreme case, assume that in the above problem-solving example Step 1 is the most critical step, and that for examinees who have completed the subprocesses of considering and using the Pythagoras' theorem, all other steps will be completed mechanically. In such a case, location parameter, β_{u_i}, for each of $u_i = 2, 3, 4, 5$ is likely to be substantially lower than β_i. As a result, with reasonably small values of ξ_{u_i} for $u_i = 2, 3, 4, 5$, the item will become practically a dichotomous item, with $P_{u_i}(\theta)$ for each of $u_i = 1, 2, 3, 4$, taking on very small values for all θ, and $P_5(\theta)$ will be close to $M_1(\theta)$. On the other hand, if the steps leading to a solution become progessively more difficult, then it is likely that β_{u_i} increases with u_i.

Note that in Step 2 the sequential order between the two subprocesses is arbitrary. Successful completion of the step does not depend on the order in which the subprocesses are completed. If there are two or more subprocesses within a step whose sequential order is arbitrary, then these subprocesses are said to be *parallel*, as distinct from *serial* subprocesses. It is assumed that for any number of parallel subprocesses the subprocesses acceleration parameters are invariant to shifts of the positions of the subprocesses in the sequence. Thus, the step acceleration parameter, ξ_{u_i} $(= \xi_{w_{u_i 1}} + \xi_{w_{u_i 2}} + \cdots)$, will be unchanged regardless of the sequential order of these parallel subprocesses.

Note that any grading system is arbitrary. If our experimental setting is improved and allows observation of the examinee's performance in more finely graded steps, then m_i will become larger. It is obvious from Eq. (14) and the definition of ξ_{u_i} that the resulting category response functions still belong to the acceleration model which partially satisfies the additivity criterion. It is also obvious that the model can be generalized to a continuous response model as the limiting situation in which the number of steps approaches infinity.

From Eqs. (4) and (14), the cumulative category response function,

$P^*_{u_i}(\theta)$, is given by

$$P^*_{u_i}(\theta) = \prod_{s=0}^{u_i} [\Psi_s(\theta)]^{\xi_s},$$

and from Eqs. (3) and (14), the category response function is obtained such that

$$P_{u_i}(\theta) = \prod_{s=0}^{u_i} [\Psi_s(\theta)]^{\xi_s} \left[1 - [\Psi_{(u_i+1)}(\theta)]^{\xi_{u_i+1}}\right]. \tag{16}$$

The basic function, $A_{u_i}(\theta)$, in this model is obtained from Eqs. (6), (14), and (15), so that

$$A_{u_i}(\theta) = D\left[\sum_{s \le u_i} \xi_s \alpha_s [1 - \Psi_s(\theta)] - \xi_{u_i+1}\alpha_{u_i+1} \frac{[\Psi_{(u_i+1)}(\theta)]^{\xi_{u_i+1}}[1 - \Psi_{(u_i+1)}(\theta)]}{1 - [\Psi_{(u_i+1)}(\theta)]^{\xi_{u_i+1}}}\right], \tag{17}$$

for $u_i = 1, 2, \ldots, m_i - 1$, and for $u_i = 0$ and $u_i = m_i$ the first term and the second term on the right-hand side of Eq. (17) disappear, respectively. The upper and lower asymptotes of this basic function are $D\sum_{s \le u_i} \xi_s \alpha_s$ (> 0) for $u_i = 1, 2, \ldots, m_i$, and $-D\alpha_{u_i+1}$ (< 0) for $u_i = 0, 1, \ldots, m_i - 1$, respectively, and the upper asymptote for $u_i = 0$ and the lower asymptote for $u_i = m_i$ are both zero. It has been demonstrated (Samejima, 1995) that $A_{u_i}(\theta)$ in Eq. (17) is strictly decreasing in θ. Thus, the unique maximum condition is satisfied for $u_i = 0, 1, \ldots, m_i$. It has also been shown (Samejima, 1995) that the orderliness of the modal points of the category response functions usually holds, except for cases in which the unidimensionality assumption is questionable, and additivity of the category response functions practically holds. Figure 3 illustrates the six category response functions by a solid line, with $m_i = 5$ and the parameters:

$\alpha_{u_i} = 1.35119,\ 1.02114,\ 0.85494,\ 1.08080,\ 0.58824;$

$\beta_{u_i} = -0.95996,\ -0.79397,\ -0.01243,\ 1.33119,\ 0.8000;$ and

$\xi_{u_i} = 0.43203,\ 0.53337,\ 0.57260,\ 0.61858,\ 1.00000,$ for $u_i = 1, 2, 3, 4, 5.$

The dashed lines in this figure represent those following Master's PCM or Muraki's GPCM with unity for the item parameters, using $\alpha_{u_i} = 1$, 2, 3, 4, 5, 6 and $\beta_{u_i} = 1.0$, 2.0, 3.0, 3.5, 1.8, 1.0 in Eq. (12) with h_i replaced by u_i for $u_i = 0$, 1, 2, 3, 4, 5, respectively.

Parameter Estimation

In estimating the item parameters of the normal-ogive model in the homogeneous case, if our data are collected for a random sample from an unscreened population of examinees, assuming a normal ability distribution,

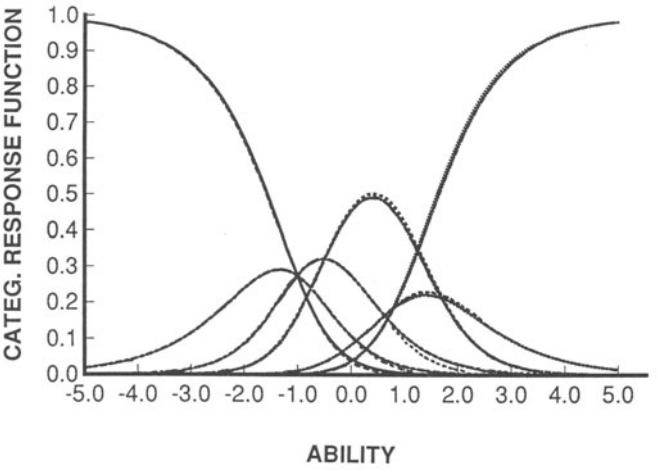

FIGURE 3. Example of a set of six category response functions following the acceleration model, which approximate those following the PCM (dotted line).

we can (a) compute the polychoric correlation coefficient for each pair of items, (b) factor analyze the resulting correlation matrix and confirm the unidimensionality, and (c) estimate a_i and b_{u_i} for $u_i = 1, 2, \ldots, m_i$, using the same method adopted by Lord (1952) and Samejima (1995). In the logistic model, item parameter estimation can be done by using the EM algorithm. Thissen wrote Multilog for this purpose, as well as for Bock's nominal response model, among others, with facilities for fixing parameter values at priors and imposing equality constraints (Thissen, 1991; Thissen and Steinberg, this volume).

Multilog is based on the EM solution of the marginal likelihood equations. Bock and Aitkin (1981) proposed a solution for dichotomous responses, and Multilog uses a direct expansion to handle graded responses in the homogeneous case. Unlike the joint likelihood solution, the marginal likelihood solution treats the parameter θ_j of the examinee j as incidental, and integrates it out from the likelihood function. Thus, the marginal likelihood function, $L(a_i, b_{u_i})$, can be written as

$$L(a_i, b_{u_i}) = \prod_{j=1}^{N} \int_{-\infty}^{\infty} g(\theta_j) P_{v_j}(\theta_j) d\theta_j = \prod_{j=1}^{N} P_{v_j} = \prod_{V} P_v^{r_v}, \qquad (18)$$

where N is the number of examinees in the sample, $g(\theta_j)$ is the ability density function, v_j is the response pattern obtained by an examinee j, $P_{v_j}(\theta)$ given by Eq. (1) for an examinee i equals the joint likelihood function, $L(\theta_j, a_i, b_{u_i})$, P_{v_i} is the marginal probability for the response pattern v_j of an examinee j, r_v is the frequency of a specific response pattern v, and $\prod_v$ indicates the product over all possible response patterns. Actually,

the continuous ability θ is replaced by q discrete latent classes, each characterized by θ_k $(k = 1, 2, \ldots, q)$, and homogeneity is assumed within each class.

Let $\tilde{P}_v$ be an approximation for P_v, which is given by

$$\tilde{P}_v = \sum_{k=1}^{q} P_v(\theta_k) G(\theta_k), \tag{19}$$

where $G(\theta_k)$ is the Gauss–Hermite quadrature weight. The expected frequency, $\bar{r}_{u_i k}$, of the graded response u_i to item i in the latent class k is provided by

$$\bar{r}_{u_i k} = \frac{\sum_v r_v x_{v u_i} P_v(\theta_k) G(\theta_k)}{\tilde{P}_v}, \tag{20}$$

where $x_{v u_i}$ $(= 0,1)$ is the indicator variable that assumes 1 if $u_i \in v$, and 0, otherwise. The expected sample size, $\bar{N}_k$, of the latent class k is given by

$$\bar{N}_k = \frac{\sum_v r_v P_v(\theta_k) G(\theta_k)}{\tilde{P}_v}. \tag{21}$$

Thus, in the E-step of the EM algorithm, $P_v(\theta_k)$ is computed for provisional a_i and b_{u_i}'s, and then $\tilde{P}_v$, $\bar{r}_{u_i k}$, and $\bar{N}_k$ are obtained, using Eqs. (19), (20), and (21), respectively. In the M-step, improved estimates of a_i and b_{u_i}'s are obtained by maximizing the approximated likelihood function, which is given by the last expression of Eq. (18), with P_v replaced by $\tilde{P}_v$. The cycles are continued until the estimates become stable to the required number of places. Muraki and Bock (1993) wrote Parscale, which includes the logistic model for graded response data, based on essentially the same EM algorithm.

Since the acceleration model was recently proposed, there is little software available. Parameter estimation can be done, however, using the following method when Eq. (15) is adopted for $\Psi_{u_i}(\theta)$. Suppose we have used a nonparametric estimation method like Levine's (1984) or Samejima's (1983, 1993, 1996), and estimated category response functions, $P_{u_i}(\theta)$'s, and they have tentatively been parametrized using a very general semiparametric method [see, for example, Ramsay and Wang, (1993)]. From these results, $\hat{M}_{u_i}(\theta)$ and its partial derivative with respect to θ can be obtained by means of Eqs. (4) and (5). Then select three arbitrary probabilities, p_1, p_2, and p_3, which are in an ascending order, and determine θ_1, θ_2, and θ_3 at which $\hat{M}_{u_i}(\theta)$ equals p_1, p_2, and p_3, respectively. From Eqs. (14) and (15), the estimated acceleration parameter $\hat{\xi}_{u_i}$ is obtained as the solution of

$$\frac{\theta_3 - \theta_2}{\theta_2 - \theta_1} = \frac{\log[(p_2)^{-1/\xi_{u_i}} - 1] - \log[(p_3)^{-1/\xi_{u_i}} - 1]}{\log[(p_1)^{-1/\xi_{u_i}} - 1] - \log[(p_2)^{-1/\xi_{u_i}} - 1]}. \tag{22}$$

The estimate, $\hat{\beta}_{u_i}$, is given as the solution of

$$\hat{M}_{u_i}(\beta_{u_i}) = \left[\frac{1}{2}\right]^{\xi_{u_i}}, \tag{23}$$

and from these results the estimate of α_{u_i} is obtained by

$$\hat{\alpha}_{u_i} = \frac{2^{\hat{\xi}_{u_i}+1}}{D\hat{\xi}_{u_i}} \frac{\partial}{\partial\theta}\hat{M}_{u_i}(\theta) \quad \text{at } \theta = \hat{\beta}_{u_i}. \tag{24}$$

Note that this method can be applied for any curve as long as $\frac{\partial}{\partial\theta}\hat{M}_{u_i}(\theta)$ (> 0) is available at $\theta = \hat{\beta}_{u_i}$. Actually, the parameters in the acceleration model used in Fig. 3 were obtained by this method from the $\hat{M}_{u_i}(\theta)$'s in Masters' or Muraki's model as the solutions of Eqs. (22)–(24), setting $p_1 = 0.25$, $p_2 = 0.50$, and $p_3 = 0.75$ in Eq. (22).

Goodness of Fit

For small numbers of items, the observed frequencies for each response pattern may be tabulated; Multilog (Thissen, 1991) computes the expected frequencies for each response pattern and then the likelihood-ratio statistic as a measure of the goodness of fit of the model. Parscale (Muraki and Bock, 1993) adopts the likelihood-ratio chi-square statistic as a measure of fit for each item, and the sum of these chi-square statistics provides the likelihood-ratio chi-square statistic for the whole test. When the polychoric correlation matrix is used for parameter estimation in the normal-ogive model, goodness of fit can be examined by the chi-square statistic using bivariate normal frequency for each of the $(m_i + 1)(m_h + 1)$ cells for a pair of items, i and h, as the theoretical frequency; when most of the $n(n-1)/2$ chi-square values are large, then we must conclude that one or more assumptions are violated; when large chi-square values are clustered around a relatively small number of items, exclusion of these items will usually improve the fit, although the psychological meaningfulness of the remaining items must be checked to be such that the construct has not been distorted or changed (Samejima, 1994). Note that goodness of fit of the curves should not be used as the *sole* criterion in accepting or rejecting a specified model; there are many models that are based on quite different principles and yet produce similar sets of curves.

Example

An interesting application of a graded response model was reported in the medical literature. Roche et al. (1975) applied the logistic model in the homogeneous case, represented by Eq. (10), for skeletal maturity. Children of the same chronological age and sex differ in maturity levels of body systems, and thus the skeletal age must be estimated using a skeletal maturity scale. The items were thirty-four carefully selected radiographically visible maturity indicators of the left knee joint, and the radiographs for

552 normal children were evaluated by experts in terms of their maturities with respect to each maturity indicator into two to five graded categories ($1 \leq m_i \leq 4$). An example of the items can be seen in item FEM-K, lateral capping, graded into three categories, that is, absent, incomplete, and complete. Capping refers to the way in which the epiphysis overlaps the metaphysis as maturation proceeds (Roche et al. 1975). This item is suitable for 7.5 to 16 year-old girls and 8 to 17 year-old boys.

The male and female data were analyzed separately, using Logog (Kolakowski and Bock, 1973), and 3,997 and 3,800 radiographs were involved for the two groups, respectively. For each group, the origin and unit of the skeletal maturity scale were set so that its mean and variance equal the mean and variance of the chronological age in the calibration sample, respectively. The resulting 34 discrimination parameters, a_i's, for the most part, ranged from 0.5381 to 3.8376 for males and from 0.4269 to 6.1887 for females, with two common outliers, item FEM-D (epiphyseal shape) and item TIB-C (metaphyseal shape), for both groups, of which a_i's are 8.1123 and 5.7874 for males and 8.8509 and 7.9769 for females. Overall configurations of a_i's were very similar between the two groups, and the same was also true with the configurations of the difficulty, or threshold, parameters. To give a couple of examples, out of the six items with $m_i = 4$, the item with the lowest b_1 was item FEM-A ([epiphyseal width]/[metaphyseal width]) for both groups, with 0.4927, 1.3556, 2.3016, and 3.2535 for b_{u_i} ($u_i = 1$, 2, 3, 4) for males and 0.4262, 0.9984, 1.7310, 2.2797 for females; and the item with the highest b_4 was item TIB-B ([epiphyseal width]/[epiphyseal height]) for both groups, with 5.0022, 6.3704, 8.2832, and 10.4988 for males and 3.8412, 4.8152, 6.3514, and 8.3020 for females.

The resulting skeletal maturity scale demonstrated assessibility to a high proportion of the participants in the Fels Longitudinal Study for each age group, that is, for most age groups the proportion is approximately 0.9 or greater, except for ages 0.1, 17.0, and 18.0. It was found that the regression of the skeletal age on the chronological age is almost linear, except for drops at the end showing the maximum drop of 16.88 against 18.0 for the female 18-year-old group. The distributions of skeletal age are skewed significantly within some chronological age groups in each sex with variable directions of skewness. The standard deviations of skeletal age increase until about 4.5 years in each sex, then more or less stabilize, and drop at 18 years for males and 17 and 18 years for females, the same tendencies observed in the regression of the skeletal age on the chronological age. There are many other interesting findings reported in Roche et al. (1975).

Discussion

It has been observed that in the homogeneous case, the models have features that satisfy the criteria for good models, while in the heterogeneous

case, fulfillment of these criteria becomes more difficult. The heterogeneous case provides greater varieties in the configuration of the category response functions; and therefore model fit will be better. In certain situations, such as Likert scale attitude surveys, it will be reasonable to apply a model in the homogeneous case, which assumes invariance of the discrimination power of an item for every possible redichotomization of the graded categories. This is supported by the results by Koch (1983). In general, however, it will be wise to start with a nonparametric estimation of the category response functions, and let the data determine which of the cases is more appropriate.

Acknowledgments

The author would like to thank David Thissen for his assistance running Multilog and H. Paul Kelley and Mark Reckase for their assistance with a literature review on the topic of applications of graded response models. Part of the work in preparing this paper was supported by the Office of Naval Research, N00014-90-J-1456.

References

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

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**, 443–459.

Kolakowski, D. and Bock, R.D. (1973). *Maximum Likelihood Item Analysis and Test Scoring: Logistic Model for Multiple Item Responses.* Ann Arbor, MI: National Educational Resources.

Koch, W.R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement* **7**, 15–32.

Levine, M. (1984). *An Introduction to Multilinear Formula Scoring Theory* (Office of Naval Research Report, 84-4). Champaign, IL: Model-Based Measurement Laboratory, Education Building, University of Illinois.

Lord, F.M. (1952). A theory of mental test scores. *Psychometric Monograph*, No. 7.

Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison Wesley.

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

Muraki, E. (1992). A generalized partial credit model: Application of an

EM algorithm. *Applied Psychological Measurement* **16**, 159–176.

Muraki, E. and Bock, R.D. (1993). *Parscale.* Chicago, IL: Scientific Software.

Ramsay, J.O. and Wang, X. (1993). *Hybrid IRT Models.* Paper presented at the Meeting of the Psychometric Society, Berkeley, CA.

Roche, A.F., Wainer, H., and Thissen, D. (1975). *Skeletal Maturity: The Knee Joint As a Biological Indicator.* New York, NY: Plenum Medical Book.

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

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

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

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer and S. Messick (Eds.), *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord* (pp. 159–182). Hillsdale, NJ: Lawrence Erlbaum.

Samejima, F. (1993). Roles of Fisher type information in latent trait models. In H. Bozdogan (Ed.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* (pp. 347–378). Netherlands: Kluwer Academic Publishers.

Samejima, F. (1994). Nonparametric estimation of the plausibility functions of the distractors of vocabulary test items. *Applied Psychological Measurement* **18**, 35–51.

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

Samejima, F. (1996). Rationale and actual procedures of efficient nonparametric approaches for estimating the operating characteristics of discrete item responses. (In press.)

Thissen, D. (1991). *Multilog User's Guide — Version 6.* Chicago, IL: Scientific Software.