Wu, A. D., & Zumbo, B.D. (2007). Thinking About Item Response Theory from a Logistic Regression Perspective: A Focus on Polytomous Models. In Shlomo S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 241-269). Information Age Publishing, Inc.., Greenwich, CT..

Bruno D. Zumbo, Ph.D.
Professor

Associate Member: Department of Statistics, Department of Psychology
Member: Institute of Applied Mathematics

Dept. of ECPS (MERM Program)
University of British Columbia

Phone: 604-822-1931
Fax: (604) 822-3302
e-mail: **bruno.zumbo@ubc.ca**
http://www.educ.ubc.ca/faculty/zumbo/index.html

Real Data Analysis

edited by

Shlomo S. Sawilowsky
Wayne State University

IAP
INFORMATION AGE
PUBLISHING
Charlotte, North Carolina • www.infoagepub.com

Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching

Ronald C. Serlin, Series Editor

Structural Equation Modeling: A Second Course (2005)
edited by Gregory R. Hancock and Ralph O. Mueller

Real Data Analysis (2006)
edited by Shlomo Sawilowsky

Library of Congress Cataloging-in-Publication Data

Real data analysis / edited by Shlomo S. Sawilowsky.
p. cm. — (Quantitative methods in education and the behavioral sciences)
Includes bibliographical references.
ISBN-13: 978-1-59311-564-7 (pbk.)
ISBN-13: 978-1-59311-565-4 (hardcover)
1. Social sciences—Statistics—Methodology. 2. Social
sciences—Research—Methodology. I. Sawilowsky, Shlomo S.
HA29.R32 2007
300.72—dc22

2006033775

ISBN 13: 978-1-59311-564-7 (pbk.)
ISBN 13: 978-1-59311-565-4 (hardcover)
ISBN 10: 1-59311-564-4 (pbk.)
ISBN 10: 1-59311-565-2 (hardcover)

Credits: The back cover image is by Student (William Sealey Gosset), "Probable error of a correlation coefficient," *Biometrika,* 1908, *6*(1), p. 305, copyrighted by Oxford Unversity Press, and reprinted with permission.

Printed in the United States of America

CHAPTER 16

THINKING ABOUT ITEM RESPONSE THEORY FROM A LOGISTIC REGRESSION PERSPECTIVE

A Focus on Polytomous Models

Amery D. Wu and Bruno D. Zumbo

The purpose of this chapter is to describe the conceptual bridge between item response theory (IRT) and logistic regression (LogR) by describing the essential similarities and differences between these two statistical frameworks. In so doing, we foster knowledge translation from psychometrics to those disciplines extensively using LogR (e.g., sociology, health care, and epidemiology) hence increasing the use of IRT. Therefore, the goal of this chapter is to advance the use of item response theory in real data analyses settings. Furthermore, it becomes apparent early on in this chapter that IRT is a special case of LogR, hence one can not only use LogR as a perspective to describe IRT to novices but also as a way of IRT specialists gaining insight into complex models such as polytomous IRT and their assumptions.

Real Data Analysis, pp. 241–269

It should be noted that we are not suggesting that we have built the bridge between LogR and IRT but rather that we are describing this bridge and using it as a way of getting from one vantage point to the other. The chapter is organized in three major sections traveling along the bridge from LogR to IRT. The first section is a brief overview of the family of logistic regression models. The second section describes the bridge between LogR and IRT. In the third section, IRT is described from the vantage point of LogR with particular attention to how these IRT models are constructed and their assumptions. This description of IRT will focus, in particular, on organizing and articulating the variety of polytomous IRT models because polytomous data are commonly found in day-to-day research settings yet polytomous IRT is seldom applied. It becomes apparent throughout that the LogR perspective brings a useful organizing framework and allows one to fully appreciate the range of IRT models and their assumptions.

A BRIEF OVERVIEW OF THE FAMILY OF LOGISTIC REGRESSION MODELS

ZUMBO

The use of LogR has greatly increased during the last decade and become routinely available in statistical packages (Hosmer & Lemeshow, 2000; Peng, Lee, & Ingersoll, 2002), especially in areas like medicine, health science and epidemiology. The goal of LogR is to model categorical outcome variables by regressing on some explanatory variable(s) (Hosmer, & Lemeshow, 2000). LogR is an engine for modeling *categorical* outcome variables that are unlikely to meet the demanding assumptions of least squares regression. LogR only assumes "conditional independence" which means that the error terms are uncorrelated and that a linear relationship between the explanatory variable and the logit outcome variable (as discussed later).

Generally speaking, categorical outcome variables can be classified into two kinds: (1) *nominal*, if the variation between/among the possible outcomes is in the form of "types" such as types of learning strategies, and (2) *ordinal*, if the possible outcomes can be logically "ordered" such as grades. Under each of the nominal or ordinal form, the outcomes variable may take up two or more categories. When there are only two categories, the outcome variable is referred to as binary; and polytomous if there are three or more categories. Because there are only two possible outcomes for a binary variable, the distinction between whether a binary outcome variable is nominal or ordinal is usually regarded as irrelevant. For this reason, the majority of the textbooks and statistical software often organize LogR analyses into three sections: binary LogR, ordinal LogR,

Table 16.1. Classification of LogR and IRT Models

Number of Categories for the Outcome Variable	*LogR Models*	*IRT Models*
2 (binary)	Binary LogR	Binary IRT (1, 2, or 3 PL)
More than 2 (polytomous)		
Ordinal	Ordinal LogR	PC (1PL), RS (1PL), GR (2PL)
Nominal	Multinomial LogR	NR (2PL)

Note: LogR = logistic regression, IRT = item response theory, PL = parameter logistic, PC = partial credit, RS = rating scale, GR = graded response, and NR = nominal response.

and multinominal LogR. When the outcome variable is binary, one can simply apply binary LogR regardless of whether the outcome variable is nominal or ordinal. When the outcome variable is polytomous, one can apply multinomial LogR if the outcome variable is nominal, or ordinal LogR if ordered. Table 16.1 is a summary of our above description of the LogR models—we will return to Table 16.1 when we describe the IRT models and connect them to their corresponding LogR models. The first column of Table 16.1 classifies the three types of LogR by the number of outcome variable categories. Note that polytomous LogR models can also be applied to binary outcome variables and yields the same results. This is because a binary outcome variable can be seen as a special case of polytomous variable with only two categories.

BINARY LOGISTIC REGRESSION

A simple LogR with one explanatory variable and one binary outcome variable can be expressed as a probability function

$$P(u = 1|X) = \frac{\exp[c + \alpha X]}{1 + \exp[c + \alpha X]} \tag{1}$$

where u is a discrete random variable that takes up the sample space of "1" (a.k.a., case; success) or "0" (a.k.a., noncase; failure); X is an explanatory variable; the "exp" denotes the operator[1] that returns e (the base of natural logarithms) raised to a power; c and α are the intercept and the regression coefficient, respectively, for the linear regression in the logit form discussed below. The expression $P(u = 1|X)$ can be read as "the probability of success/case given X, for example, the probability of having lung cancer given that one smokes."

Readers may have noticed that we used somewhat different notation from those of most textbooks; this is because our notation will serve to maintain the consistency in our later discussion of IRT models. To estimate the regression coefficients, LogR makes use of the maximum likelihood method where one maximizes the likelihood function given the data at hand. For mathematical and practical simplicity, however, one actually minimizes the –2 log likelihood function. The Wald statistic and its associated p value are used to test the significance of individual coefficients. The amount of reduction in –2 log likelihood minimized by adding the explanatory variable compared to the base model that includes only the intercept term serves as a model fit statistics. In other words, a perfect fitting model will minimize the starting –2 log likelihood (a.k.a., deviance) to 0. A variety of effect size measures such as Nagelkerke R-square and Pseudo R-square were proposed to mimic "the percentage of variance explained" in linear regression.

Modeling the occurrence of a certain outcome is related to another key feature of LogR: the nonlinear relationship between the "probabilistic" outcome variable and the explanatory variable. The nonlinear relationship for a binary LogR is often characterized by a monotonically increasing S-shaped curve. Note that the modeled variable, $P(u = 1|X)$, in Equation 1 can be transformed by taking the natural logarithm of the odds (i.e., the ratio of probability of outcome being 1 to the probability of not being 1), and yields

$$\text{Logit} = \ln\left[\frac{P(u = 1|X)}{1 - P(u = 1|X)}\right] = c + \alpha X, \qquad (2)$$

or simply

$$\text{Logit} = c + \alpha X, \qquad (3)$$

where α is analogous to the regression coefficient and c to the intercept in linear regression. Two things should be noted here. First, LogR does not model the raw response (i.e., 0 and 1); instead, it models the probability or the logit. Second, the logit is assumed to be linear in its coefficients and is continuous ranging from $-\infty$ to $+\infty$. The transformed logit regression models given in Equation 2 or 3 have many of the desired properties of a linear regression model, and hence, the LogR model is regarded as a type of *generalized linear model*.

POLYTOMOUS LOGISTIC REGRESSION: MULTINOMIAL AND ORDINAL

For binary LogR, there are only two possible outcomes, 0 and 1. Modeling the probability of outcome "1" occurring, *P*, is sufficient because the probability of "0" occurring is simply 1 – *P*. In contrast to binary LogR, multinomial and ordinal LogR involve more than two possible outcomes; therefore, require simultaneously fitting multiple regression curves. As a general rule, for an outcome variable that consists of $J + 1$ categories, J regression analyses will be entailed. Here J is the maximum coding of the outcome categories when the coding of the possible outcomes begins with 0. For example, an outcome variable that is measured on five ordered categories such as strongly disagree, disagree, neutral, agree, and strongly agree and is coded as 0, 1, 2, 3, and 4 will have a maximum coding of J equals to 4. Hence, four regression lines will be fitted for the $J + 1 = 5$ response categories. This systematic notation of "*J*" is capable of providing a lot of information about the specification of a model including the coding for the possible outcomes (i.e., 0, 1, ..., *J*), maximum value for outcome coding (*J*), number of outcome categories ($J + 1$), and number of regression analyses involved (*J*). Hence, this notation will be used throughout the rest of this chapter.

There are numerous ways of specifying the probabilities for the $J + 1$ outcomes occurring (see Agresti, 2002, chapter 7). For nominal outcomes, multinomial LogR involves a *direct* method of specifying $P(u = j|X), j = 0, 1, 2, \ldots, J$, which means that the probability is obtained directly by a divide-by-total procedure such that

$$P(u = j|X) = \frac{\exp[c_j + \alpha_j X]}{\sum_{j=0}^{J} \exp[c_j + \alpha_j X]}, \tag{4}$$

with $c_0 = 0$ and $\alpha_0 = 0$. For ordinal outcomes, the most common method for specifying the probabilities for the $J + 1$ outcomes is an *indirect* (a.k.a., difference) method using the *cumulative logit* (see O'Connell, 2006 for other ordinal LogR models). Namely, the ordered categories are contrasted into J dichotomies such that responding in $u \leq j$ is contrasted with $u > j$. For example, suppose that the outcome variable has four ordered categories coded as 0, 1, 2, and 3, three ($J = 3$) regression analyses will be entailed and are achieved by contrasting the ordered outcomes into three dichotomies: (i) 0 versus 1, 2, and 3 (ii) 0 and 1 versus 2 and 3 (iii) 0, 1, and 2 versus 3. This cumulative contrasting is most widely used and the cumulative probability is written as

$$P(u \leq j|X) = \frac{\exp[c_j + \alpha X]}{1 + \exp[c_j + \alpha X]}, \text{ or} \tag{5}$$

$$Logit = \ln\left[\frac{P(u \leq j|X)}{P(u > j|X)}\right] = c_j + \alpha X. \tag{6}$$

Note that there is no subscript for the α slopes across the cumulative logits. This is because the slopes, most commonly, are assumed to be equal (i.e., parallel) in cumulative logit LogR. This *equal slopes assumption* in (5) and (6) is also referred to as *proportional odds assumption* (Agretsi, 2002, p. 275). The cumulative logit LogR is the default model in SPSS and SAS and is the mostly commonly applied LogR model for ordinal outcomes. Because ordinal LogR models the cumulative probabilities, the probability of a specific category occurring must be written as a difference between two adjacent cumulative probabilities such that

$$P(u = j\,|\,X) = P(u \leq j\,|\,X) - P(u \leq j - 1\,|\,X). \tag{7}$$

Later in our discussion readers will see that one of the ordinal IRT models, the Graded Response model, is built on very similar conceptual frameworks and assumptions in expressing the probabilities of examinees' specific response to test item.

To reiterate our brief introduction on LogR: The goal of LogR is to model the nonlinear probabilistic relationship between a categorical outcome variable and the explanatory variable(s). The logit form of the regression is assumed to be linear in its regression coefficients: intercept and slope, and can be classified as a generalized linear model. There are three major classes of LogR models: binary, ordinal and multinomial. The choice of which model to apply depends on the metric, nature, and number of the categorical outcome variable. For polytomous LogR models, J regression analyses are entailed for the $J + 1$ possible outcomes and the probability of an individual outcome can be obtained by direct or indirect specification. In addition to logit linear and conditional independence, the proportional odds ordinal LogR assumes equal slopes across the cumulative logits. These backbones of LogR foreshadow our discussions on IRT as a special form of LogR in the next section. Readers interested in LogR should consult Hosmer and Lemeshow (2000), Menard (2001), O'Connell (2006) or Peng et al. (2002) for more details.

DESCRIBING THE BRIDGE BETWEEN LOGR AND LOGISTIC IRT

Although IRT and LogR seem to share little in common on the surface, the embryo of using logistic regression groundwork in developing IRT can be traced back to Birnbaum (1968) and Rasch (1960). However, users of neither methods have explicitly described the affiliation between the two methods, and, consequently, there has been a lack of conceptual and organizational framework linking these two popular methods. In addition, we believe, that IRT is used less often in day-to-day research practice because it is often portrayed as a distinct method from what is widely known, such as LogR and regression modeling.

How is IRT connected to and distinct from LogR? In broad strokes, IRT and LogR are both branches of generalized linear models except that the explanatory variable in IRT is a *latent variable*, as opposed to an *observed variable* in LogR. For this reason, IRT is referred to as, to be more precise, a *generalized linear latent model*. Another major distinction, which is also related to the construction of the latent explanatory variable, is that IRT simultaneously model a number of categorical outcome variables. In LogR language, IRT runs multiple regression analyses at the same time. We will further explain these two distinctions in our subsequent discussions. However, at this point, it is more important to foreshadow that IRT is connected to LogR because they share the same framework and mechanism. These commonalities include the purpose, the assumptions, the shapes of the regression curve, the coding system, the specification of the probability, the estimation method, as well as the classification and choice of major IRT models explained hereafter.

IRT is defined as a model-based measurement theory that aims to specify a mathematical function relating the probability of an examinee's response on a test item to an underlying ability (van der Linden & Hambleton, 1996). Often, the choice of the mathematical functional form is a logistic curve. Namely, IRT uses a logistic curve to depict the nonlinear probabilistic relationship, Item Characteristic Curve (ICC, see Figure 16.1), between examinees' item response and their ability. Here, *ability*, often denoted as θ, is a generic term used in IRT literature to represent the underlying characteristic, construct, or trait being measured. Hence, in a regression sense, IRT intends to regress a categorical outcome variable, *item response*, onto the explanatory variable, the examinees' ability. In addition to the conditional independence assumption, the relationship between the logit-transformed item response and ability is assumed to be linear. In this sense, IRT uses the logistic function to model an item response and can be regarded as a type of generalized linear model.

However, note that what makes IRT distinct from a typical LogR is the nature of the explanatory variable. Normally, the explanatory variable in a LogR is an observed variable where data is obtained from the direct observation of the sampling units, whereas the explanatory variable in IRT, θ, is a *continuous latent variable*—an unobserved variable that must be created and estimated. This continuous latent variable θ, in short, is constructed by exploiting the *joint probability distributions* of the examinees' responses to a studied item and the rest of the test items. Therefore, when building an IRT regression model, one must simultaneously estimate the explanatory variable, which is the *person parameter* θ and the regression coefficients, which are the *item parameters.* As a side note, there exists an analogy between least squares regression and normal theory factor analysis. That is, one can simply conceive of factor analysis as multivariate ordinary least squares regressions with the latent continuous explanatory variable(s) (i.e., factor, also called the latent variable) being the predictor(s), which are created by accounting for the inter-correlations among the observed continuous variables. Despite these two major differences, the same principles of classification and choice of models apply to IRT. Binary IRT model is designed for binary item response. When item response is polytomous, nominal response IRT model is the choice if response categories are nominal, otherwise ordinal IRT model.

LOGISTIC IRT AS A SPECIAL CASE OF LOGR

Binary Logistic IRT

In this chapter, the notation used to describe IRT models will follow those of Embretson and Reise (2000). For the purpose of easy illustration, we will restrict the subsequent discussions of LogR and IRT to only one explanatory variable. Often, the number of sufficient item parameters assumed to aptly fit the data classifies IRT models (a.k.a., 1PL, 2PL and 3PL, etc.). In LogR language, the number of (item level) regression coefficients needed to accurately describe the relationship between examinees' ability and the item responses classifies IRT models. The binary logistic IRT model given below includes three parameters (i.e., 3PL) and is in the most general form,

$$P(u_i = 1 | \theta, \alpha_i, \beta_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{\exp[\alpha_1(\theta - \beta_1)]}{1 + \exp[\alpha_i(\theta - \beta_i)]}, \quad (8)$$

where

i = i^{th} item; 0, 1, ..., I
u_i = the item response of an examinee to item i (0 or 1)
θ = the ability level of an examinee

α_i = the discrimination (a.k.a., slope) for item i
β_i = the threshold (a.k.a., difficulty) for item i
γ_i = the lower asymptote (a.k.a., pseudo-chance) for item i.

Note that we have not indexed θ. Instead, we treat it as a random vector with dimension equal to the sample size N. Figure 16.1 shows the ICC for a hypothetical item with item parameters $\alpha = 1.5$, $\beta = 0.5$, and $\gamma=0.1$. One can see that the x-axis is the latent continuous ability θ scaled to a mean of 0 and standard deviation of 1. The probability of $u = 1$ shown on the y-axis given the item parameters is a function of examinees' θ. The discrimination parameter, α, is related to the slope at the point of inflection of the ICC indicating how precise or sensitive an item is in discriminating an examinee with high ability from one with low ability. The pseudo-chance parameter, γ, is located at the point on the y-axis with which the lower asymptote intersects. The pseudo-chance parameter indicates the chance of endorsing or getting an item right with no or little of the ability being measured (e.g., $\theta = -3$). The threshold parameter, β, is the value on the θ continuum with which the vertical line drawn from the inflection point intersects. The threshold value indicates how much ability examinees would need to have a $(1 + \gamma)/2$ chance of endorsing or getting an item right (i.e., 0.5 for 1PL and 2 PL models because $\gamma = 0$).

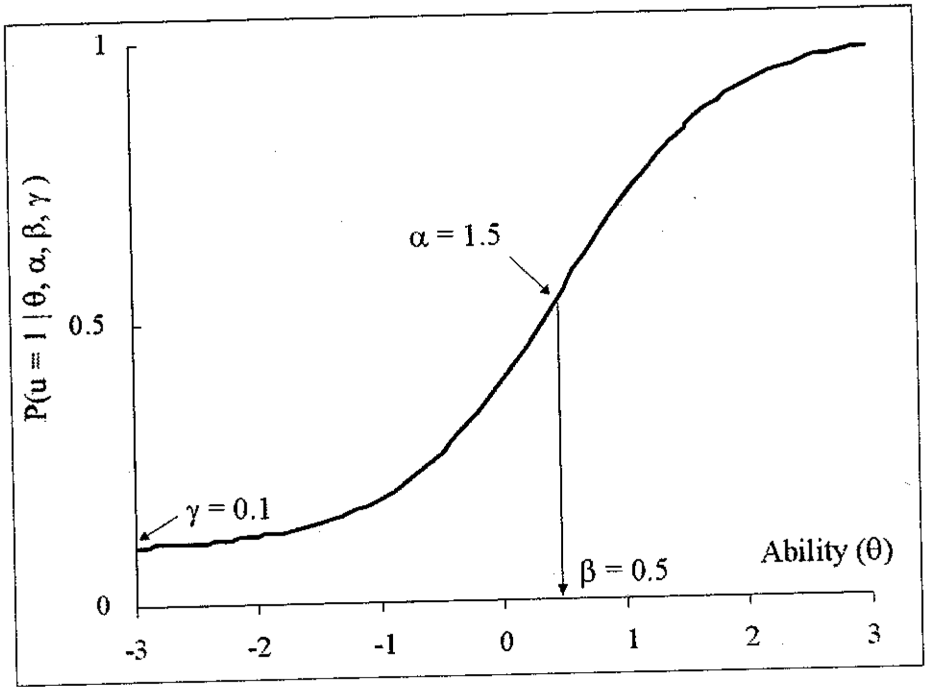

Figure 16.1. Item Characteristic Curve (ICC) for binary item response.

Note that the structure of Equation 8 looks very similar to that of the simple binary LogR given in Equation 1, the only difference is that the IRT model is more elaborate in two ways: First, a latent variable θ is involved, and second, a larger number of coefficients (i.e., item parameters) are specified. The term "$(\theta - \beta_i)$" represents the discrepancy between how much ability an examinee possesses and how much ability an examinee should possess to have a $(1 + \gamma_i)/2$ chance of endorsing or getting item i right (i.e., item threshold). Hence, one can understand Equation 8 as "regressing the probability of getting an item right onto the term $(\theta - \beta_i)$."

One can easily construct the 1PL and 2PL IRT models by removing the unnecessary parameters in Equation 8. To show the connection between LogR and IRT, a 2PL model for item i is shown below by removing the γ_i parameter,

$$P(u = 1|\theta, \alpha, \beta) = \frac{\exp[\alpha(\theta - \beta)]}{1 + \exp[\alpha(\theta - \beta)]}. \tag{9}$$

In Equation 9, if we treat $(\theta - \beta)$ as our explanatory variable by relabeling it with X and add a zero intercept term, c, we get a variation of the basic LogR Equation 1,

$$P(u = 1|X) = \frac{\exp[c + \alpha X]}{1 + \exp[c + \alpha X]}, \tag{10}$$

where $X = (\theta - \beta)$, and $c = 0$.

One can see that the 2PL IRT model in Equation 10 is identical to the binary LogR model in Equation 1. Of course one can also perform the logit transformation to Equation 10 and obtain expressions like Equations 2 and 3 and show that 2PL IRT logit is linear in its coefficients, and hence can be classified as a generalized latent linear model. These simple mathematical manipulations demonstrate that unidimensional IRT (i.e., involving only one θ) indeed is a simple LogR model. Same as LogR, estimation of IRT person and item parameters makes use of maximum likelihood estimation methods, or sometimes Bayes estimation methods, which we will not explain in detail in this chapter. Interested readers are referred to Baker and Kim (2004) for details or Embretson and Reise (2000) and Hambleton, Swaminathan, and Rogers (1991) for a concise introduction.

In summary, IRT uses LogR function to characterize the nonlinear relationship between the probability of a categorical item response and a continuous latent variable θ by estimating the necessary item parameters. Given that IRT is a type of LogR, one way of classifying the IRT models is to follow Table 16.1, which is organized by the number of response categories. Focusing on column three of Table 16.1 one is able to map our subsequent discussions on the commonly used polytomous IRT models.

POLYTOMOUS LOGISTIC IRT

Apparent in the labeling, polytomous IRT models refer to modeling item responses that take up three or more categories. Following our earlier classification of LogR in Table 16.1, there are two divisions of polytomous IRT models classified by the metric nature of the item response: ordinal or nominal. However, to better understand our discussion of the two divisions of polytomous IRT models, we need to preface our discussion by looking at the similarities with and differences between binary and polytomous IRT models and some common features of the various polytomous IRT models.

Same as binary LogR, binary IRT models require only one regression analysis, $P(u = 1)$, to describe the probabilistic relationship. This is because the probability of the only other outcome, $P(u = 0)$, is simply equal to $1 - P(u = 1)$ as in Figure 16.1. In fact, in the logit form of the regression equation as in Equation 2 and 3, it is written as the ratio of $P(u = 1)$ to $P(u = 0)$ to uniformly express the linear relationship. However, for the same reason as we discussed on polytomous LogR, polytomous IRT models involve multiple response categories. Consequently, one has to model multiple relationships for each item. In other words, for each item in a test, a series of multiple Category Characteristic Curves (CCCs) will be modeled (Dodd, 1984). Modeling multiple relationships is often done by some kind of contrasting among the response categories as we discussed in LogR: the maximum code and the number of analyses entailed are equal to J for the $J + 1$ categories when the lowest category is coded as 0. In other words, within each item, it takes J steps (i.e., thresholds) for an examinee to stride from the lowest response category, 0, to the highest, J. Note that the number of response categories does not have to be equal across items. Another important notion is needed to prime our polytomous IRT discussions. For polytomous IRT models, there is a hierarchical structure of parameters involved. *At the test level*, a polytomous IRT will simultaneously model a number of items and their corresponding item parameters may or may not vary across items. *At the item level*, in the meantime, a polytomous IRT will simultaneously model a number of response categories within each item, and their corresponding parameters may or may not vary across categories.

In sum, historically, a variety of polytomous IRT models were developed to appropriately describe the multiple probabilistic relationships varying in these regards: (1) the metric nature of the item response, (2) the methods of contrasting among the $J + 1$ multiple response categories, (3) within an item, how the parameters are assumed across response categories, and (4) within a test, how the parameters are assumed across items. The first three points are analogous to LogR whereas the last point is unique to IRT. Obviously, modeling polytomous IRT is far more complex than the binary models. What are the justifications and payoffs for choosing the more complex polytomous models over the binary models? Ostini and Nering (2005) and van der Ark (2001) listed three major reasons for preference for polytomous items over binary items. First, polytomous responses provide more precise information than binary responses. As a result, fewer items are typically needed to achieve the same degree of reliability. Second, some psychological constructs are often measured on rating scales. Last, certain kinds of item responses (i.e., those that are naturally ordered) are better characterized on an ordinal scale. For these reasons, polytomous IRT is believed to be a statistically more malleable and practically useful for polytomous responses. However, because of its statistical complexities in applications and interpretations, polytomous IRT models are less often discussed than they should have been. Therefore, by using the conceptual framework of LogR, we hope our discussions will disentangle the complexities and elucidate understanding of the polytomous IRT models. Before proceeding, therefore, readers may wish to review the aforementioned similarities and distinctions between LogR and IRT as in Table 16.2, which is a summary of the above description.

The following remarks provide an overview of four commonly applied polytomous IRT models: three IRT models for ordinal responses and one IRT model for nominal responses. Readers are directed to de Ayala (1993), Embretson and Reise (2000), Ostini and Nering (2005), van der Ark (2001), or Van der Linden & Hambleton (1996) for more technical details and alternative models. Also, acknowledging the inconsistency and complexity in the notation and terminology used in the polytomous IRT literature, we attempt to synthesize the discrepancies in the terminology and minimize the number of necessary notations needed to express across various polytomous IRT models. For this notation system to work, the coding of polytomous response categories should follow the coding system we mentioned throughout our earlier discussions.

Table 16.2. Similarities and Distinctions Between LogR and Logistic IRT

Similarities	*LogR*	*IRT*
Purpose	Modeling categorical outcome variable	Modeling categorical item response
Model assumption	Conditional independence Logit linear	Conditional independence Logit linear (1PL, 2PL)
Outcome category coding	0, 1, ..., J for J + 1 categories	0, 1, ..., J for J + 1 categories
Probability modeling	Direct or indirect	Direct or indirect
Regression function	Logistic	Logistic
Estimation method	Maximum likelihood	Maximum likelihood (or Bayes)
Classification	Binary, ordinal, nominal	Binary, ordinal, nominal
Distinctions	*LogR*	*IRT*
Explanatory variable	Observed categorical or continuous	Latent continuous
Number of outcome variables	Modeling one outcome variable at a time	Modeling multivariate items simultaneously
Parameter assumption	Only across response category assumption	Both across item and across response category assumptions

ORDINAL ITEM RESPONSE

Partial Credit Model (PC)

Self-evident in the labeling, the 1PL partial credit model was originally developed by Masters (1982) to model partial credits assigned to examinees, who respond to test items involving multiple steps. For example, an item may instruct examinees to resolve the height of a triangle as the first step and then resolve the area of the triangle as the second step. One partial credit will be assigned to an examinee who only correctly solves the first step; two full credits will be assigned to an examinee who correctly solves both steps; and no credit will be assigned if an examinee does not successfully solve the first step given that one is unlikely to successfully solve the second step without correctly solving the first. Using our coding system, each examinee will receive a score, u = 0, 1, or 2. The maximum value J and the number of logistic regression analyses entailed are equal to 2 for the J + 1 = 3 response categories. In PC models, the probabilistic relation is specified as a *direct* IRT model like the multinomial LogR. As described in LogR introduction, the probability of getting a particular credit (i.e., category) is written directly as an exponential divided by the sum of all the exponentials that can possibly appear in the numerator (Embretson & Reise, 2000, p. 105) as below:

$$P_{ij}(\theta) = \frac{\exp\left[\sum_{j=0}^{u_i}(\theta-\beta_{ij})\right]}{\sum_{j=0}^{J}\left[\exp\sum_{j=0}^{J}(\theta-\beta_{ij})\right]}, \qquad (11)$$

$\sum(\theta-\beta_{ij}) = 0$ when $u_i = 0$, $P_{ij}(\theta)$ = the probability of $u = j$ for item i given θ.

An example will make Equation 11 more accessible. If the i^{th} item is scored on a scale of 0, 1, 2, and 3, the probability of $u = 2, J = 3$ would have a numerator of $\text{Exp}[(\theta-\beta_0) + (\theta-\beta_1) + (\theta-\beta_2)]$ and a denominator of $\text{Exp}[(\theta-\beta_0)] + \text{Exp}[(\theta-\beta_0) + (\theta-\beta_1)] + \text{Exp}[(\theta-\beta_0) + (\theta-\beta_1) + (\theta-\beta_2)] + \text{Exp}[(\theta-\beta_0) + (\theta-\beta_1) + (\theta-\beta_2) + (\theta-\beta_3)\ ]$. In words, for item i, the numerator is the exponential of the cumulative $(\theta-\beta_j)$ up to $j = u$, and the denominator is always the same, which is the sum of all the possible numerators. As one can see in Equation 11, PC models assume that only the step parameters β_{ij} (i.e., threshold) is needed to specify the probabilistic relationship between the multiple response categories and θ, and are interpreted as thresholds for transition from one category to the next. The step parameters are located at the intersection points between two adjacent CCCs indicating where on the θ continuum the response of one category becomes relatively more likely than the previous category (see Figure 16.2). Readers should be cautious not to interpret step parameters as the point on the θ continuum where an examinee has a 50% of responding above a category threshold. Using the example of the area of a triangle, the PC model requires that the steps within an item be completed in sequence, although the steps need not be equally difficult, nor be ordered by the levels of difficulty. When the step difficulties are not ordered (see the example in Embretson & Reise, 2000, p. 109), a *reversal* is said to exist (Dodd & Koch, 1987). In PC models, the slopes (i.e., the discrimination) for the CCCs across the response categories are assumed to be equal and fixed at 1, hence drop out of Equation 11. De Ayala (1993) showed that the Rasch model (binary, 1PL, $\alpha_i = 1$) is simply a special case of the PC model with two response categories. Like the Rasch model, packaged with the equal discrimination assumption, PC models have the advantage of using the total score as a sufficient statistic for estimating examinees' θ score. Because only the β_{ij} parameter is estimated, the sample size required to obtain quality item and category parameters is smaller than the 2PL polytomous IRT models discussed later. However, the assumption of equal slopes (i.e.m discrimination) across items and category parameters, at the same time, restricts the use of PC models in practice. For this reason, Muraki (1992, 1993) proposed the Generalized Partial Credit model (GPC) where item discrimination parameters are allowed to vary across items.

To illustrate a PC model, we used the item parameters of Table 5.5 in Embretson and Reise (2000, p. 108). The item parameters were estimated based on 350 undergraduate students who responded to the 12 items on the neuroticism scale of the Neuroticism Extroversion Openness Five-Factor Inventory (NEO-FFI) (Costa & McCrae, 1992). These items were scored on a 0 to 4 scale (0 = *strongly disagree*; 1 = *disagree*; 2 = *neutral*; 3 = *agree*; 4 = *strongly agree*). Figure 16.2 shows the CCCs and the four threshold parameters for item two: "I feel inferior to others." Note that the threshold parameter β_1 between item categories 0 and 1 is the intersection of the two CCCs and is located at the point −1.763 on the θ continuum. In addition, the four thresholds divide the θ continuum into five intervals and each interval encompasses the θ range where a specific response category is more likely. For example, if a respondent has a θ value between −1.673 and 0.080, as shown in Figure 16.2, he or she will be estimated to endorse "1" more likely than other response categories.

Rating Scale Model (RS)

Extended from the PC models, Andrich (1978a, 1978b) proposed a 1PL polytomous IRT model to accommodate the rating scale type of responses. In performance or achievement tests, it is logical that a test item requires multiple steps for completion, and the step difficulty would differ across steps and across items. However, it is reasonable to assume that the threshold values would remain very similar across items for an attitude rating scale with common anchors such as 0 = strongly disagree, 1 = disagree, 2 = neutral, 3 = agree, and 4 = strongly agree. Namely, a set of ordered *J*-step parameters (i.e., category thresholds) is well suited for all items in a measure. For this reason, a step threshold β_{ij} in the RS model is decomposed into two parts, β_i and δ_j, where $\beta_{ij} = \beta_i + \delta_j$. For each item, there is one β_i threshold parameter that is allowed to vary across items. The set of δ_js are the category level parameters and are fixed to be the same for all items in a test. If one substitutes $\beta_i + \delta_j$ for β_{ij} in Equation 11 for the PC model, one would get the expression as below:

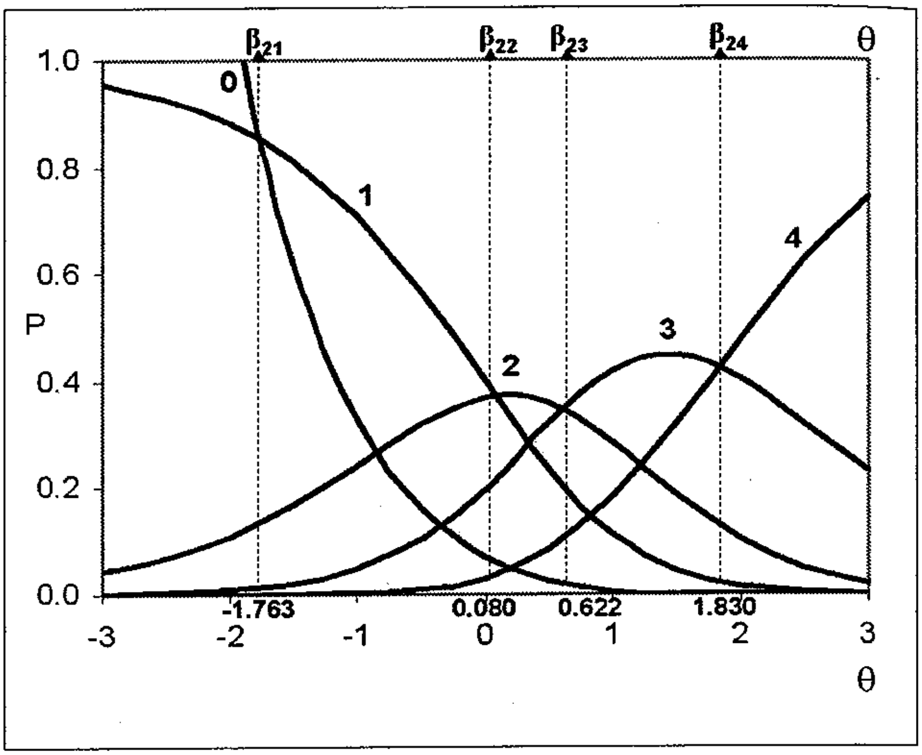

Figure 16.2. Category Characteristic Curves for the PC Model.

$$P_{ij}(\theta) = \frac{\exp\left[\sum_{j=0}^{u_i}[\theta-(\beta_i+\delta_j)]\right]}{\sum_{j=0}^{J}\left[\exp\sum_{j=0}^{J}[\theta-(\beta_i+\delta_j)]\right]}, \tag{12}$$

where $\sum[\theta-(\beta_i+\delta_j)] = 0$ when $u_i = 0$.

The interpretation for the parameters β_{ij} is the same as that of the PC model; They are thresholds for transition from one category to the next and are located at the intersection points between two adjacent CCCs, indicating where on the θ continuum the response of one category becomes relatively more likely than the previous category. Figure 16.3 illustrates the CCCs for a RS model using our earlier example item 2. We see that, for item two, $\beta_2 = 0.300$, and the set of δs are: $\delta_1 = -1.600$; $\delta_2 = 0.224$; $\delta_3 = -0.184$; $\delta_4 = 1.560$. This set of four δs was estimated and fixed for all the items in the scale despite that item threshold parameters, β_i, are free to vary across items such as 0.300 estimated for our example item 2. Using $\beta_{ij} = \beta_i + \delta_j$, we yield threshold values for item two: $\beta_{21} = -1.300$; $\beta_{22} = 0.524$; $\beta_{23} = 0.116$; $\beta_{24} = 1.860$. One can see that the RS model is more restricted than the PC model because it assumes the set of category thresholds is equal across items in addition to equal discrimination assumption in the PC model.

As with the PC model, the conditional probability of endorsing a particular point on the rating scale can be obtained through direct operation specified in Equation 12. The RS and PC models share the same advantages and drawbacks because they are both 1PL models assuming item and category discrimination parameters to be "1." One additional limitation of the RS model is that it is not suitable for test items with different response formats. However, for the same reason, the RS model has the advantage of being more parsimonious because it entails even fewer parameter estimates than the PC model. Another advantage is that this set of identical categorical threshold estimates could provide some information on the psychological distances between the scale points for the underlying construct being measured.

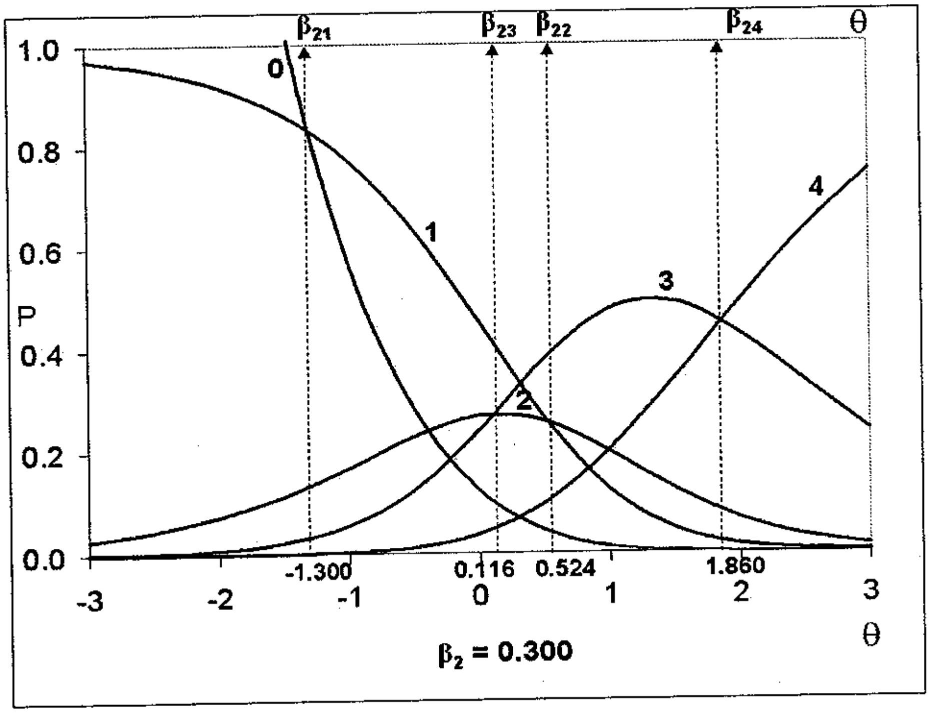

Figure 16.3. The Category Characteristic Curve for the RS model.

Graded Response Model (GR)

Samejima (1969, 1996) developed the 2PL GR models for ordered categorical response. In the GR model, the response categories are contrasted with J dichotomies such that responding in $u_i < j$ is contrasted with $u_i \geq j$. For instance, an item with 5 score points, 0, 1, 2, 3, and 4, would have four contrasting dichotomies as: (1) 0 versus 1, 2, 3, and 4; (2) 0 and 1 versus 2, 3, and 4; (3) 0, 1, and 2, versus 3 and 4; (4) 0, 1, 2, and 3, versus 4. Consequently, the GR model specifies the probability in terms of responding in u_i or higher ($u_i \geq j$) in relation to θ scores. In other words, for each of the J dichotomies, a probabilistic relationship will be modeled. Embretson and Reise (2000) referred to the J curves as Operating Characteristic Curves (OCCs) and can be written as,

$$P(u \geq j|\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{1 + \exp[\alpha_i(\theta - \beta_{ij})]}, \qquad j = 0, \ldots, J, \qquad (13)$$

where $P(u \geq j|\theta)$ is the probability of responding in a particular category score j or higher on item i. Hence, the probability of responding in the lowest category or higher, $P_{i0}(\theta)$, is equal to 1. Note that the contrasting and probability modeled are opposite to those of LogR we introduced earlier where $u \leq j$ is contrasted with $u > j$ and the cumulative probability of $P(u \leq j|X)$ is modeled. However, the logic for contrasting the $J + 1$ outcomes using J cumulative dichotomies remains the same. The β_{ij} parameter in the GR model is the threshold indicating the θ level needed to make a response that is equal to and greater than the threshold j with a 50% probability for item i (see Figure 16.4). In the GR model, the discrimination parameters α_i are always allowed to differ across items. However, the slopes may or may not vary across response categories within an item. When α_i is constant across the response categories, it is referred to as a *homogeneous* GR model and when α_i is not constant across response categories, it is referred to as a *heterogeneous* GR model. Homogeneous GR models are more commonly applied in practice and are conceptually equivalent to the cumulative logit LogR, where the slopes are assumed to be parallel. Figure 16.4 illustrates the OCCs for our example item 2 based on the homogeneous GR Model. The corresponding parameters are: $\alpha_2 = 1.42$ and $\beta_{21} = -2.07$; $\beta_{22} = -0.22$; $\beta_{23} = 0.93$; $\beta_{24} = 2.42$.

As in the cumulative logit LogR for ordinal outcomes, the GR models are viewed as *indirect* IRT models because the probability of an examinee's response to a particular category $P_{i(jth)}(\theta)$, hence the CCC, is obtained by,

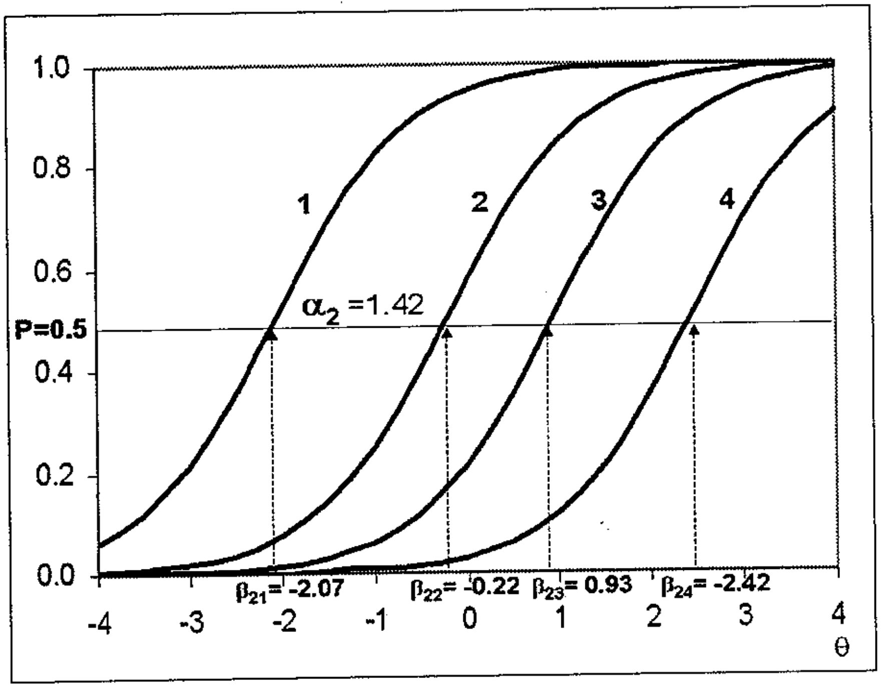

Figure 16.4. Operating Characteristic Curves for the graded response model.

$$P_{i(jth)}(\theta) = P_{ij}(\theta) - P_{i(j+1)}(\theta). \tag{14}$$

For example, the probability of endorsing category 2 would be: $P_{2\text{nd}}(\theta) = P_2(\theta) - P_3(\theta)$. In this sense, the GR model is also referred to as a "*difference*" model (Embretson & Reise, 2000; Thissen & Steinberg, 1986). Note that the homogeneous GR model is analogous to the equal slopes model in ordinal LogR in terms of (a) equal slopes assumption across response categories and (b) indirect specification of the probability. Figure 16.5 illustrates the CCCs for our example item 2. Note that the middle point of two adjacent threshold parameters β_{ij} and $\beta_{i(j+1)}$ in the CCCs depicts the point on the θ continuum where the probability of jth category peaks on the CCCs. One can see that the CCC for category1 peaks at the midpoint of β_{21} and β_{22}, and is equal to $\frac{(-2.07) + (-0.22)}{2} = -1.145$, which indicates the θ level needed to have the maximum probability of endorsing category 1 (i.e., disagree).

Readers may have noticed that, all the three ordinal IRT models we have introduced assume equal slopes across response categories. However, the GR model differs from the PC and RS models in three aspects: (1) PC and RS model fix the values of the category slopes to be 1, whereas

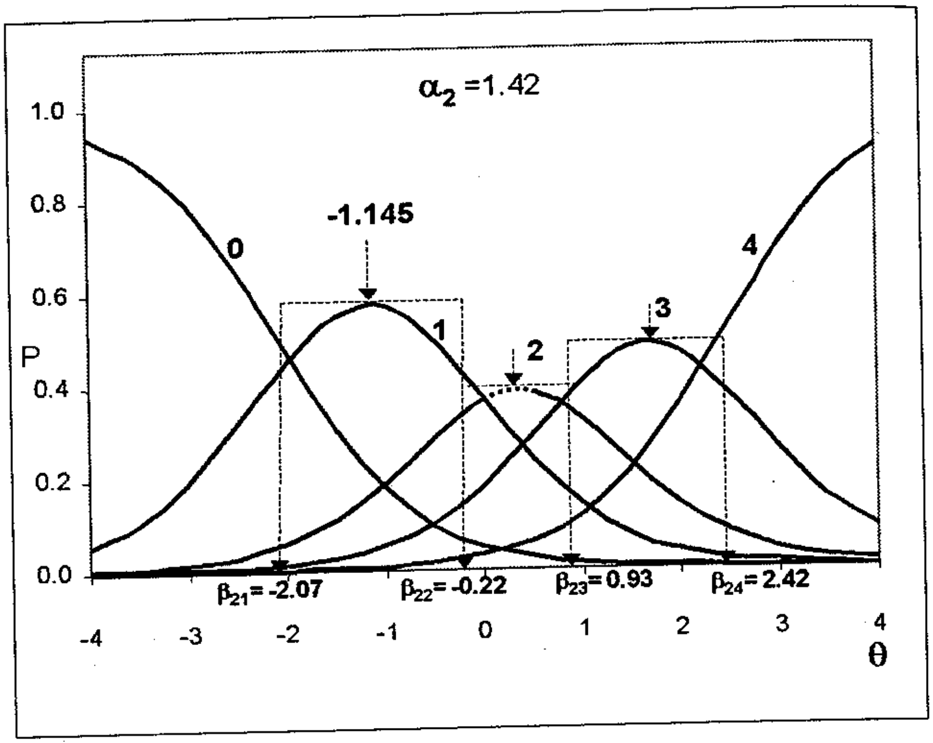

Figure 16.5. Category Characteristic Curve for the GR model.

the GR model allows the category slopes to be estimated or fixed at values other than 1 within each item, (2) the discrimination parameters are allowed to differ across items such as our example item 2 being estimated at 1.42, (3) items within a test need not have the same number of response categories like the RS model (i.e., J does not have to the same for all items), and (4) the β_{ij} in a GR model are always ordered such that $\beta_{(j+1)} > \beta_j$.

De Ayala (1993) showed that 2PL binary IRT models are simply a special case of GR models with two response categories. One of the advantages of the GR model is that the item discrimination parameters, unlike PC, GPC, and RS models, are allowed to vary across items. This advantage is welcomed by a set of test items that are likely to have differential discrimination power as in the attitude or personality measures. In addition, the GR model has the flexibility to accommodate test items with different response formats. Notice that a modified graded response model was developed by Muraki (1990, 1992) to model a Likert-type response format where the items are of the same number of response categories. One of the drawbacks of the GR model is the indirect calculation of the CCCs.

NOMINAL ITEM RESPONSE

Nominal Response Model (NR)

Bock (1972) proposed a 2PL polytomous IRT model characterizing item responses that were on a nominal scale. His initial intention was to model the alternatives in multiple-choice items. Conventionally, multiple-choice items are scored into a dichotomization of correct or incorrect and modeled accordingly. The NR model argued that the information provided by examinees' wrong responses by choosing a certain distracter is not all the same and should not be treated uniformly as "incorrect." Modeling an item's incorrect response to distracters may provide more information about an examinee's level of ability. The NR model is a direct probability model and can be written as:

$$P_{ij}(\theta) = \frac{\exp(c_{ij} + \alpha_{ij}\theta)}{\sum_{j=0}^{J} \exp(c_{ij} + \alpha_{ij}\theta)}, \qquad (15)$$

where $\sum \alpha_{ij} = \sum c_{ij} = 0$ or $\alpha_{i0} = c_{i0} = 0$.

The α_{ij} parameters are interpreted the same way as the discrimination parameters in the ordinal IRT, and c_{ij} are the intercept parameters of the nonlinear response function associated with *j*th category of item *i*. Specifically, c and α are the intercept and the slope, respectively, for the linear regression in the logit form. Readers should be cautious not to interpret the c parameter in the same manner as one would interpret the γ parameter in the 3PL binary model, which actually indicates the pseudo-chance parameter of an item. For each item, as usual, there are J threshold parameters (a.k.a., location parameter in the logit linear literature) that are often assumed to be unordered, although there are occasions where data indicates that they are, in fact, ordered. Because Equation 15 is invariant with respect to translation of the logit, the constraint on $\sum \alpha_{ij} = \sum c_{ij} = 0$ or $\alpha_{i0} = c_{i0} = 0$ is needed as an anchor to solve the identification problem. One can see that the expression and identification restriction in Equation 15 are almost identical to that in Equation 4 for multinomial LogR except that IRT is subscribed by "*i*" indicating that "*I*" items are simultaneously modeled. De Ayala (1992) showed that the threshold parameters could be obtained by,

$$\beta_{ij} = \frac{c_{(j-1)} - c_j}{\alpha_j - \alpha_{(j-1)}}. \tag{16}$$

The β_{ij} parameters are analogous to the step difficulty of the PC model and are located at the intersection of adjacent CCCs. In fact, all the divide-by-total or direct methods (e.g., PC model and RS model) can be shown to be special cases of the nominal response model (Embreston & Reise, 2000; Thissen & Steinberg, 1986). The NR model is the most general specification of polytomous IRT models. This means that it has the least assumptions made about the number of item and category parameters as well as the order of the category thresholds. Namely, both the threshold and discrimination parameters are free to vary across items and across categories except for the identification restrictions. Like other polytomous IRT models, the NR model can also be applied to binary item response. To illustrate, we borrowed the example of Tatsuoka (1983) in de Ayala (1993): a multiple-choice item with three options (i.e., two distracters). Using our notation system, J equals to 2, where the first option was coded as "0," second as "1," and third as "2." The J is also the maximum code indicating that there are $J + 1 = 3$ options. This item is a mathematics addition problem asking "$-6 - 10 = ?$" with three alternatives: (a) -16, (b) -4, and (c) 4. Figure 16.6 shows the values for α were $\alpha_0 = -0.75$, $\alpha_1 = -0.25$, and $\alpha_2 = 1.0$ and, although not shown, the values for c were $c_0 = -1.5$; $c_1 = -0.25$; $c_2 = 1.75$. Using Equation 16, we yielded $\beta_1 = -2.5$ and $\beta_2 = -1.6$. The advantage of the NR model is that it is the most flexible model for the different types of item responses. Its limitation, for the same reason, is that it is less parsimonious to specific types of item responses.

SUMMARY

A brief introduction to LogR was given in terms of the purpose, assumptions, functional forms, when to use which LogR models, and its nature as a generalized linear model. Building on this introduction, we showed that IRT is a special form of LogR with the explanatory variable being a continuous latent variable constructed by accounting for the joint distributions among the test items. In IRT, the probabilistic relationship between examinees' responses to an item and their latent ability estimates is described by a nonlinear logistic function characterized by the item parameters. In addition to binary IRT, two branches of polytomous IRT models were described: ordinal and nominal. For ordinal item responses, the Partial Credit model, the Rating Scale model, and the Graded Response model were described. These models differ in the manner of

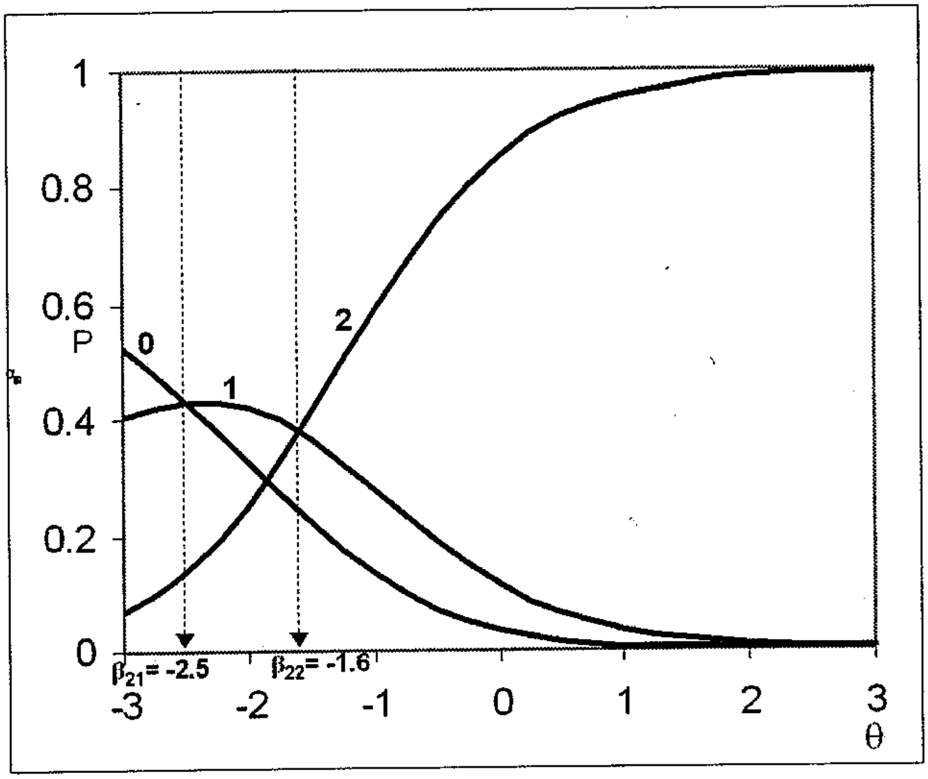

Figure 16.6. Category Characteristic Curves for the nominal response model.

how item and category parameters are constructed and how the probability is modeled. Table 16.3 summarizes the features of the four polytomous IRT models discussed in this chapter and Table 16.4 compares their advantages and disadvantages.

Closing Remarks and Filling in the Portrait of IRT From a LogR Perspective

Although we have focused on the connection between LogR and IRT, it is important to note that there is one important difference between these two methodologies. That is, because the predictor variable is a latent variable logistic IRT models, as compared to LogR, require a large sample size to achieve unbiased person and item parameter estimates. In addition, a large number of items are also needed to guarantee unbiased person parameter estimation and small sample-to-sample variation in theta estimates. An insufficient number of examinees or items will lead to an inappropriate specification of the statistical relationship. In the case of

Table 16.3. Comparisons of the Model Specifications for Four Polytomous IRT Models

Model	*PC (1PL)*	*RS (1PL)*	*GR (2PL)*	*NR (2PL)*
Across items				
α	Fixed at 1	Fixed at 1	Free	Free
β	Free	Free	Free	Free
Within an item				
α	Fixed at 1	Fixed at 1	Fixed / Free	Free
β	Free	Free, Equal across items	Free & Ordered	Free
Contrasting coding	j vs. all	j vs. all	$u \geq j$ vs. $u < j$	j vs. all
Probability modeling	Direct	Direct	Indirect	Direct

Table 16.4. Comparisons of the Advantages and Disadvantages for Four Polytomous IRT Models

Models	*PC*	*RS*	*GR*	*NR*
Advantages				
Total score is sufficient for ability estimates	✓	✓		
Requiring (relatively) smaller sample size than 2PL models	✓	✓		
Accommodating tests with different response formats	✓		✓	✓
Providing the psychological distance of the measured construct		✓		
Item discrimination is allowed to vary			✓	✓
Flexibility to all types of item responses				✓
Disadvantages				
Equal discrimination restricts use in practice	✓	✓		
Requiring (relatively) larger sample size than 1PL models			✓	✓
Not suitable for test items with different response formats		✓		
Indirect specification of probability			✓	
Less parsimonious to specific types of items responses				✓

small sample sizes and short tests or scales, practitioners should consider using nonparametric IRT models in which no prespecified functional form, such as LogR, would be imposed to describe the relationship between the item response and the ability score (see, e.g., Sijtsma & Molenaar, 2002).

The parametric logistic IRT framework described herein should be increasingly utilized in day-to-day measurement research and data analysis in the health, social, and behavioral sciences because of its versatility and practicality in solving many problems in measurement and testing such as measurement bias, item parameter drift, test score equating, and computer adaptive testing (Embreston & Reises, 2000; Hambleton et al.,

ZUMBO

1991). These advantages of IRT models, especially when compared to classical test theory, reside in the fundamental premise that IRT measurement models generate item-independent person parameters and person-independent item parameters. However, these advantages are not guaranteed by simply fitting an IRT model to the response data. Rather, they are subject to the empirical assessment of parameter invariance of the specified model (Hambleton et al., 1991; Rupp & Zumbo, 2004), a cornerstone principle, yet often misunderstood element of IRT.

At this point, it is appropriate to say a few words about invariance in IRT models and its implications for IRT practice in terms of item bias, item drift, computer adaptive testing, and equating/linking of test versions or forms. Invariance is a population property dictating that the values of the parameters of a statistical model are identical across the subpopulations or the test conditions for which the test items are designed. Parameter invariance is often construed in applications and the applied literature as a magical yet mythical property; however, in fact, it is a universal phenomenon of all model based regression-like analyses such as least square regression, logistic regression, structural equation modeling, and IRT models (Breithaupt & Zumbo, 2002; Zimmerman & Zumbo, 2001; Zumbo & Rupp, 2004). Simply put, if a model is correctly specified (i.e., the regression function is correct for the population), then the regression parameters are invariant across the subpopulations or test conditions. IRT parameter invariance, hence, cannot be explicitly tested because it is a theoretical property in the population. At best, it may be indirectly and empirically tested by the model-data fit and by examining whether the parameters remain invariant across different calibration samples after the parameters are put on the same metric. In other words, IRT person parameter invariance and item parameter invariance hold if the set of parameters calibrated on one data is the linear transformation of those calibrated on the other (Rupp & Zumbo, 2003, 2004, in press). This exercise of linear transformation is necessary because the metrics of the IRT person and item parameters are often set arbitrarily from calibration to calibration. Note that the latent predictor, θ, is constructed from the joint distribution of the items in a scale and hence has no inherent mean or variance (i.e., metric).

Therefore, the versatile IRT day-to-day applications will succeed only if the model fits the data well and parameter invariance hold true. Following the same premise, IRT based investigation of differential item functioning, in essence, can be regarded as a statistical method for detecting item bias through the examination of lack of invariance where item parameters are variant across subpopulations such as gender or ethnic groups. The same logic applies to the investigation of item parameter drift where the initial parameters of items in an item pool show drift in a later calibration after prolonged use. Also, because of the item-independent person parameter property, a result of IRT parameter invariance, computer adaptive testing is able to assign unbiased ability scores to examinees regardless of what items in the item pool are administered to the examinees. When used to equate test scores, IRT naturally overcomes the problems of incomparability in scores of examinees taking different tests. If an IRT model fits the data well, examinees' ability scores are made directly comparable by the item-independent invariance property.

The invariance property is of less importance in LogR, even though the same premise still holds. This is because most researchers utilize LogR primarily for the explanatory purpose of statistically testing whether a set of predictors contributes to the explanation of the variation in the outcome variable. Thus, one is more concerned about whether the proportion of deviation (i.e., –2 log likelihood) explained away by the chosen predictors is just a result of sampling capitalization, rather than whether the specified model is correct in the population. In other words, LogR modellers make fewer demands on the perfect model-data fit and do not expect the model to explain away all the variation in the outcome variable; a close enough approximation of the population model would suffice. In contrast, conventional IRT is utilized as a measurement model that adopts a "model fitting" perspective which dictates that a single latent ability variable, theta, is the sole drive for people's responses and should be sufficient to account for nearly all the variation in people's responses and hence the model is expected to fit the data nearly perfectly so that the beneficial applications of IRT parameter invariance will succeed. Currently, IRT modellers have moved ahead from the simple conventional IRT models to more expanded models. For example, whether it be binary or polytomous, multidimensional IRT models (i.e., more than one single latent ability variables) have been developed to more aptly describe examinees' item responses (see, e.g., Ackerman, 1992; Embreston & Reise, 2000, p. 82). Also latent class (i.e., discrete grouping variable) IRT models have been developed where examinees are assigned to latent classes that serve as the explanatory variable for the item responses.

Finally, we believe that the future of IRT, in many ways, will move from the traditional "response fitting" measurement approach to a more explanatory approach. For example, this could be done by framing the IRT models under the generalized linear and nonlinear mixed effects models, a model with random coefficients in which the fixed and/or random effects enter the model nonlinearly (e.g., Rijmen, Tuerlinckx, de Boeck, & Kuppens, 2003) where items are nested within the examinees and one or more latent ability variables are treated as random effects, and multiple examinee-level predictors can now be incorporated into the extended model. Also the beneficial property of IRT parameter invari-

ZUMBO

ance follows naturally under this generalized linear and nonlinear mixed effects framework: the specified IRT model, within transformation, is the same model (i.e., item parameters are fixed) at various levels of the random variable, theta, when the model fits the data. In moving toward a more "explanatory" modeling strategy, measurement and psychometrics are becoming more than just a descriptive or normative process but rather one that tells the researcher why and how item responses arise. This is a relatively new avenue in IRT (De Boeck & Wilson, 2004; Lu, Thomas, & Zumbo, 2005) and more generally to a new perspective on validity and the practice of validation in measurement (Zumbo, 2005).

ACKNOWLEDGEMENT

We would like to thank Zhen Li and Siok Leng Ng for helpful and detailed comments on an earlier draft of this manuscript.

NOTE

1. Please note that exp(x) is the same function as e^x, where e is about 2.718.

REFERENCES

Ackerman, T. A. (1992). A didaction explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Agretsi, A. (2002). *Categorical data analysis.* New York: Wiley.

Andrich, D. (1978a). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-94.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-73.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques.* New York: Marcel Decker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

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

Breithaupt, K., & Zumbo, B. D. (2002). Sample invariance of the structural equation model and the item response model: a case study. *Structural Equation Modeling, 9,* 390-412.

Costa, P. T., & McCrae, R. R. (1992). *The revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16,* 327-43.

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

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York: Springer.

Dodd, B. G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent trait models. (Doctoral dissertation, University of Texas at Austin, 1984). *Dissertation Abstract International, 45,* 2074A.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement, 11,* 371-84.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling, 12,* 263-277.

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

Menard, S. (2001). *Applied logistic regression analysis* (2nd ed.). Newbury Park, CA: Sage.

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

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

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological measurement, 17,* 351-363.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables.* Thousand Oaks, CA: Sage.

Ostini, R., & Nering, M. L. (2005). *Polytomous item response theory model.* Thousand Oaks, CA: Sage.

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96,* 3-14.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (G. Leunbach, Trans.). Copenhagen: The Danish Institute for Educational Research.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185-205.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49*, 264-276.

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement, 64*, 588-599. {*Errata, (2004) Educational and Psychological Measurement, 64*, 991}.

Rupp, A. A., & Zumbo, B. D. (in press). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement.*

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

Samejima, F. (1996). The graded response model. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory.* New York: Springer.

Sijtsma, K., & I. W., Molenaar. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-77.

van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory. *Applied Psychological Measurement, 25*, 273-282.

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory.* New York: Springer.

Zimmerman, D. W., & Zumbo, B. D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing, 1*, 283-303.

Zumbo, B. D. (2005). *Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing*. Samuel J. Messick Memorial Award Lecture, LTRC 27th Language Testing Research Colloquium, Ottawa, Canada.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.