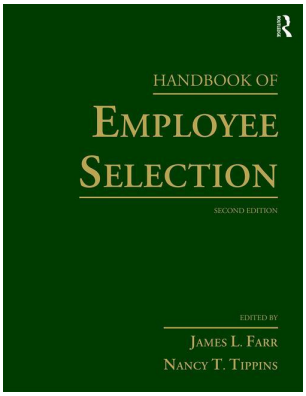

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

Modern Psychometric Theory to Support Personnel Assessment and Selection

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-42
Stephen Stark, Oleksandr S. Chernyshenko, Fritz Drasgow
Published online on: 22 Mar 2017

42

Modern Psychometric Theory to Support Personnel Assessment and Selection

STEPHEN STARK, OLEKSANDR S. CHERNYSHENKO, AND FRITZ DRASGOW

Advances in computing technology have rapidly expanded the options available for psychological assessment in work contexts. Faster computers, mobile devices with multimedia capabilities, and Internet access have virtually eliminated the need for paper-and-pencil tests. The types of items that can be presented to examinees have also expanded beyond traditional multiple-choice and Likert-type formats to include more complex stimuli, such as videos and immersive task simulations (Drasgow & Olson-Buchanan, 1999; Mills, Potenza, Fremer, & Ward, 2002; Stark, Martin, & Chernyshenko, 2015). Today, virtually all assessments can be offered "on-demand," meaning that they can be accessed by test takers at any time, and results can be provided readily to stakeholders to expedite the personnel screening process.

One of the important implications of computerization is that more sophisticated measurement technologies have gradually been implemented to support assessment needs. Item response theory (IRT) methods are particularly well-suited for designing and evaluating selection tests, because the parameters that describe items are invariant across examinee subpopulations, and neither the properties of individual items nor test scores depend fundamentally on the subset of items composing a test. IRT methods can thus be used to construct parallel or tailored test forms, to match test difficulty to individual examinee capabilities as in computerized adaptive testing (CAT), to link item properties and test scores across different measurement occasions, to identify aberrant examinee response patterns, and to test for measurement invariance across different examinee groups (e.g., Embretson & Reise, 2000; Hulin, Drasgow, & Parsons, 1983; Maydeu-Olivares & McArdle, 2005). As more flexible IRT methods are developed, IRT is likely to become the methodology of choice for supporting structured assessments.

The aim of this chapter is to introduce readers to IRT models and methods now being used in personnel assessment and selection. We hope to help readers who lack in-depth IRT training to better understand some models and applications. First, we describe four IRT models commonly used in cognitive ability and "noncognitive" (e.g., personality, attitudes, and interests) testing. Second, we discuss how examinee response data are scored and show how adding or removing items from a test affects measurement precision. Third, we discuss item response and item information functions and how they can be used to increase measurement efficiency with adaptive item selection. Finally, we describe the concept of "person fit" and how person-fit methods can be used to verify the results of unproctored tests and identify unmotivated or careless examinees. We also discuss some examples involving actual workplace tests and IRT techniques used in practice.

IRT MODELS

Three-Parameter Logistic Model (3PLM)

Item response theory involves probabilistic mathematical models that describe how an examinee's trait level and an item's properties jointly influence item responses. For example, one of the most parsimonious and well-recognized IRT models is the one-parameter logistic or Rasch model (Rasch, 1960) for dichotomous data (correct-incorrect; agree-disagree). The Rasch model uses examinee trait level (θ) and just one item property, *item difficulty* (the location of an item on the trait continuum), to predict the probability of answering an item correctly. In most psychological domains, however, more item properties (parameters) are used to model the probability of correct responses. Specifically, for multiple-choice items, like those often used in cognitive ability tests, *item discrimination* (how well an item differentiates examinees of different ability levels) and *guessing* are also taken into account.

In the three-parameter logistic model (3PLM; Birnbaum, 1968), each item is characterized by an item difficulty (a.k.a. extremity or location) parameter *(b)*, an item discrimination parameter *(a)*, and a lower asymptote or "guessing" parameter *(c)*. The probability of a correct or positive response to item *i* for the 3PLM is given by:

$$P(u_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}, \quad (1)$$

where u_i denotes an examinee's response to item *i* ($u_i = 1$ if correct; 0 if incorrect), $P(u_i = 1 \mid \theta)$ is the probability of a correct response for a randomly chosen examinee having trait level θ, and 1.7 is a scaling factor that is included for historical reasons. Note that the two-parameter logistic model (2PLM) and the one-parameter logistic model (1PLM or Rasch model) can be obtained from Equation 1 by placing "constraints" on the *a*- and *c*-parameters. If one assumes no guessing and sets $c = 0$ for all items, then the 2PLM results. If one also assumes that all items are equally discriminating (e.g., set all $a = 1$), then the 1PLM results.

The structure of the 3PLM is most easily understood by plotting the probability of a correct response as a function of θ on the latent trait range [−3.0, +3.0] and examining how the curve changes as a function of the item parameters. Figure 42.1 shows 3PLM *item response functions* (IRFs) for three hypothetical items having the same discrimination and guessing parameters ($a = 1.5$ and $c = 0.2$, respectively) but different difficulty parameters ($b = -1.0$, 0.0, 1.0, respectively). It can be seen that the item difficulty parameter affects the lateral position of the IRFs along the trait continuum. As the difficulty parameter increases from −1.0 to 1.0, the probability of a correct response, at a particular θ, decreases. For example, only examinees having $\theta > 1.5$ have a high probability of answering the item with $b = 1$ correctly.

Figure 42.2 illustrates how the item discrimination parameter affects the shape of IRFs. Shown are items that exhibit rather low discrimination ($a = 0.5$), medium discrimination ($a = 1.0$), and high discrimination ($a = 2.0$). It is evident that as *a* increases, the IRFs become steeper near $\theta = b$. Also note that the difference in response probabilities at trait levels of $\theta = -0.5$ and $\theta = 0.5$ increases as item discrimination increases. When $a = 0.5$, the response probability difference is only 0.2, but when $a = 2.0$, the response probability difference is nearly 0.8. Thus, items with large *a*-parameters better differentiate examinees at trait levels near the item difficulty parameter.

Finally, Figure 42.3 illustrates the effect of the *c*-parameter of the 3PLM. With a typical multiple-choice item, even low-ability examinees can sometimes guess the correct answer. As shown, the *c*-parameter affects the lower asymptote of the IRF so that the probability of a correct response remains above zero for any trait level. Values of the *c*-parameter typically range from 0.1 to 0.3 for items having four or five response options.

The 3PLM is often applied to cognitive ability test data, because in most cases, it is reasonable to assume that items are not equally discriminating and guessing is a realistic possibility (Hulin et al., 1983). The process of estimating item parameters is called *item calibration*. Because the

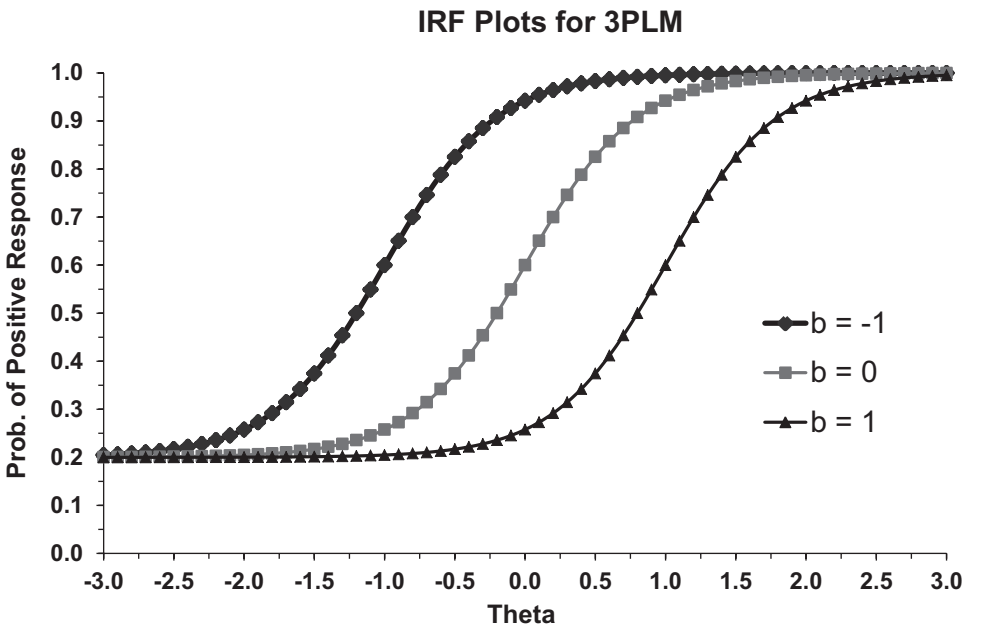

FIGURE 42.1 3PLM IRFs for Three Items Having Different Difficulty Parameters

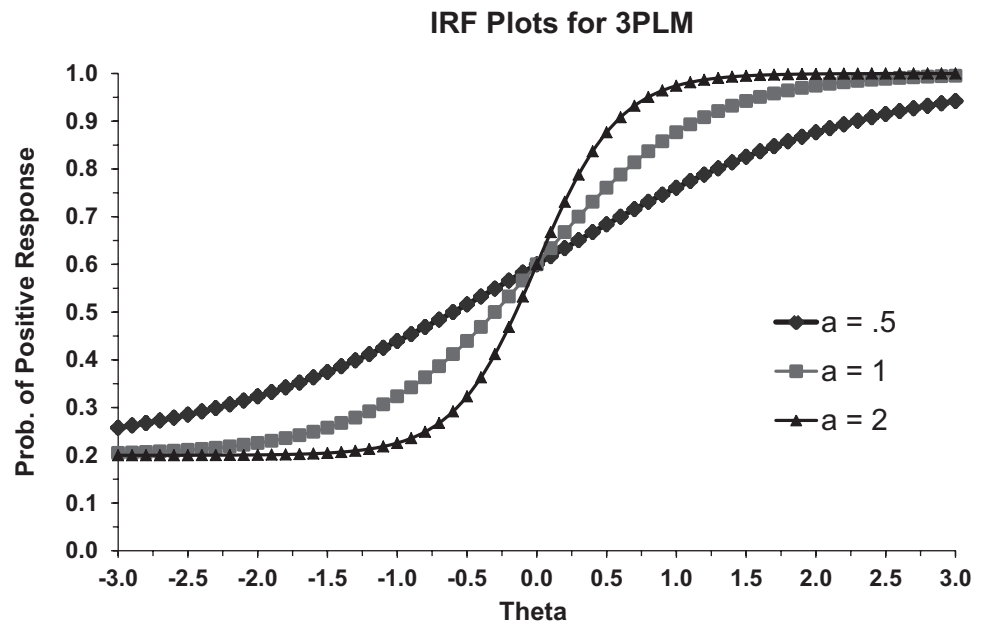

FIGURE 42.2 3PLM IRFs for Three Items Having Different Discrimination Parameters

3PLM equation has only one θ, the test data must be *essentially unidimensional*, meaning that just one dominant factor underlies the item responses. Drasgow and Parsons (1983), Hattie (1985), and Stout (1987) discuss several methods for testing the unidimensionality assumption with dichotomous data, but one of the simplest approaches, which works reasonably well in practice,

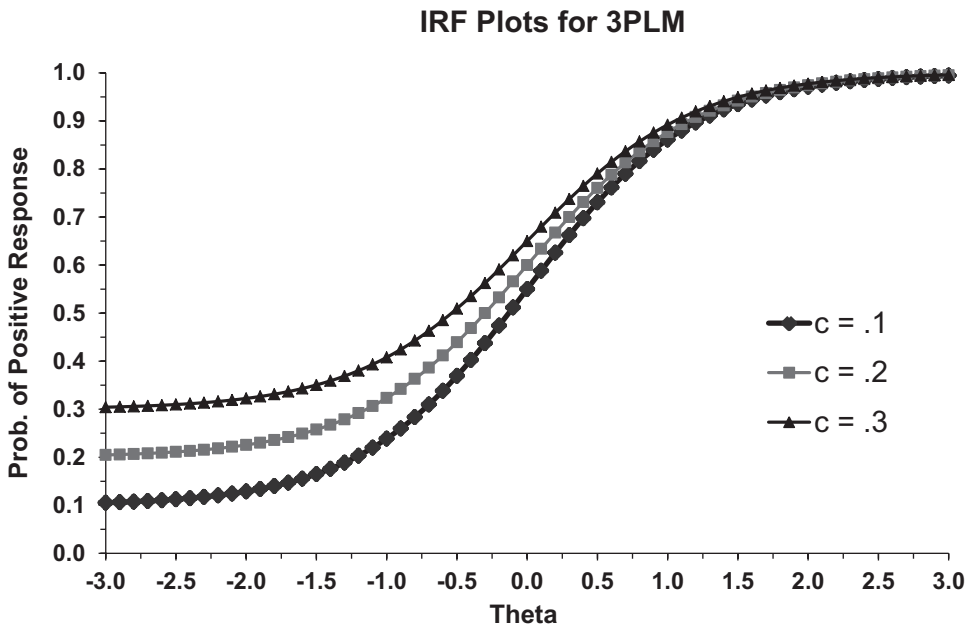

FIGURE 42.3 3PLM IRFs for Three Items Having Different "Guessing" Parameters

is exploratory factor analysis of item tetrachoric correlations. If the ratio of the first to second eigenvalue is greater than 3, then the response data may be viewed as sufficiently unidimensional (see Lord, 1980; also see Chapter 8 of Hulin et al. (1983) concerning violations of tetrachoric correlation assumptions, and the alternative *modified parallel analysis* procedure.).

Several specialized software packages are available for estimating 1PLM, 2PLM, and 3PLM IRT parameters. BILOG-MG (Zimowski et al., 2003) is still widely used, but many newer applications have been developed, and statistical packages such as Mplus (Muthén & Muthén, 2015), SAS, and R now contain functions for fitting these models. For best results, sample sizes of 250, 500, and 1000 are often recommended for the 1PLM, 2PLM, and 3PLM, respectively. Readers seeking guidance on parameter estimation are referred to Hulin et al. (1983) and Hambleton and Swaminathan (1985). For readers interested in model derivations, Baker and Kim (2004) is recommended.

Samejima's Graded Response Model (SGRM)

Models for polytomous data are more complex than models for dichotomous data, because they must account for multiple response categories. In polytomous IRT terminology, the function that relates trait level to the probability of endorsement, or choosing, a particular response category is called an *option response function* (ORF). Among the most widely used polytomous IRT models are Samejima's graded response model (SGRM; Samejima,1969), Bock's nominal model (Bock, 1972), and Muraki's generalized partial credit model (Muraki, 1992). Here, we focus on the SGRM, because it is the most commonly used with questionnaire data involving Likert-type response formats (e.g., strongly disagree, disagree, neutral, agree, strongly agree).

The mathematical form of the SGRM is

$$P(u = k|\theta) = \frac{1}{1+\exp[-1.7a(\theta - b_k)]} - \frac{1}{1+\exp[-1.7a(\theta - b_{k+1})]}, \quad (2)$$

where *u* denotes the response to a polytomously scored item, *k* is the particular option selected by the respondent, *a* is the item discrimination parameter, and b_k is referred to as a location or extremity parameter. Note that an item with *k* options will have one discrimination parameter, and *k−1* extremity parameters. These parameters are used to calculate what are known as *boundary response functions*, and the differences between successive boundary response functions give the respective *option response functions*, which relate trait level to the probability of endorsing a particular response category, as shown in Equation 2. Example SGRM ORFs for a five-option Likert-type item with a = 2.0, $b_1 = -1.5$, $b_2 = -0.5$, $b_3 = 0.7$, and $b_4 = 1.2$ are shown in Figure 42.4.

As shown in Figure 42.4, the *b*-parameters correspond to the points on the trait continuum where adjacent ORFs intersect. For example, the ORF for option 1, which is the left-most monotonically decreasing curve, intersects with the ORF for option 2 at $b_1 = -1.5$. Similarly, the ORF for option 2 intersects with the ORF for option 3 at $b_2 = -0.5$, and so on. The five ORFs are distinct (steep slopes and narrow peaks), because the discrimination parameter is large ($a = 2.0$). As a result, there are clearly identifiable regions of the trait continuum where the endorsement of a particular response option is most likely. For example, examinees located between −0.2 and 0.6 on the trait continuum have a 60% or higher chance of endorsing option 3, a 10–20% chance of endorsing options 2 or 4, and virtually zero chance of endorsing options 1 or 5. Thus, endorsing option 3 tells us that an examinee is most likely located in the −0.2 to 0.6 range.

In contrast, if an item has a low discrimination parameter, the endorsement of a particular response option provides less information with regard to an examinee's location. For example, Figure 42.5 shows the ORFs for a five-option item having the same *b*-parameters as in the previous figure but an *a*-parameter of 0.4. As can be seen in the figure, the ORFs for options 2, 3, and 4 are very flat across the trait continuum and overlap to a large extent. Thus, selecting any of these options tells us relatively little about the examinee's trait level.

Historically, the most widely used software for estimating SGRM parameters was the MULTILOG computer program (Thissen, 1991). However, as with the previously mentioned dichotomous models, SGRM parameters can now be estimated with Mplus (Muthén & Muthén, 2015), SAS, and R, as well as newer specialized applications. Samples of 1,000 have been recommended for SGRM parameter estimation, but much smaller samples may be acceptable in many settings. One concern, however, with small samples is that at least 20 to 50 persons *per category* are recommended for parameter estimation. Otherwise, infrequently

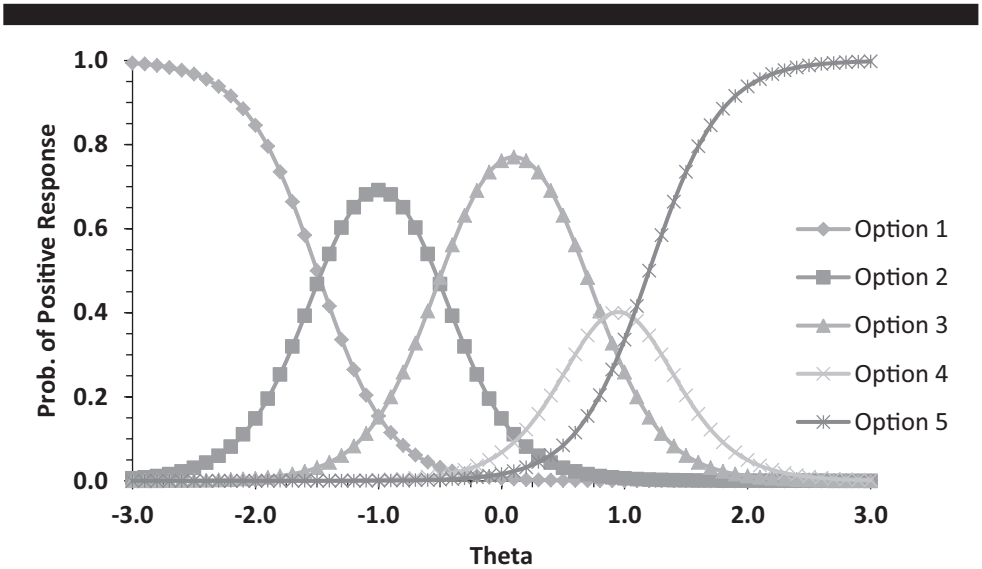

FIGURE 42.4 SGRM ORFs for a Five-Option Item Having a High Discrimination Parameter

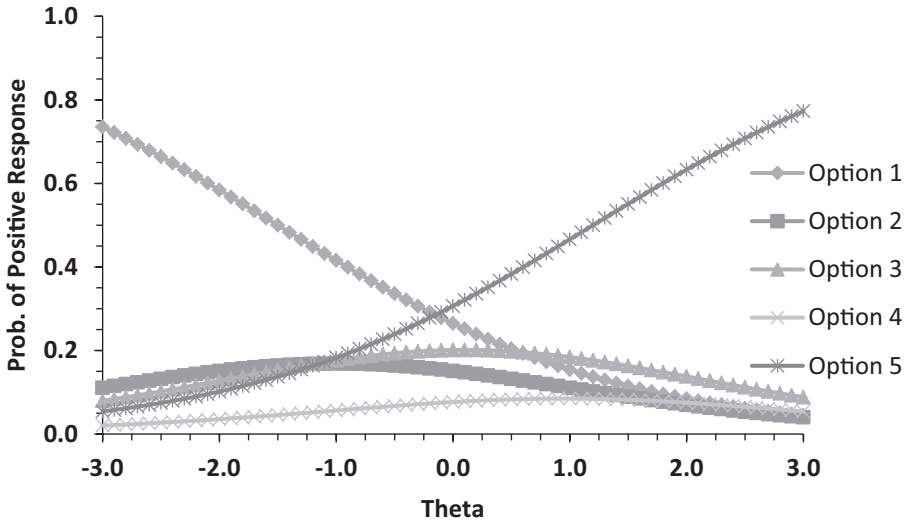

FIGURE 42.5 SGRM ORFs for a Five-Option Item Having a Low Discrimination Parameter

endorsed categories should be collapsed. Readers interested in a more detailed description of the SGRM and applications in noncognitive testing may refer to Embretson and Reise (2000) and Chernyshenko et al. (2001).

Generalized Graded Unfolding Model (GGUM)

Both the 3PLM and SGRM described above belong to a class of IRT models known as dominance models. *Dominance models* assume that the probability of a correct response (or endorsement of the highest response category) increases as an examinee's trait level increases. Thus, respondents with very high trait scores (e.g., +3.0) are those who are most likely to answer an item correctly or endorse a "strongly agree" response option (assuming negatively worded items have been reverse scored). An alternative family of models, known as *ideal point models*, makes a different assumption about item responding—namely, the probability of endorsement increases as a function of the similarity between a person and an item. For example, a person is most likely to endorse an attitude item expressing an opinion similar to his or her own. Ideal point models have been found to work well with items measuring job attitudes (Carter & Dalal, 2010), personality (Stark et al., 2006), and vocational interests (Tay et al., 2009). In particular, these models have been shown to accommodate neutral (moderate) items, which are typically discarded when dominance models are applied (Chernyshenko et al., 2007).

James Roberts and colleagues have developed a number of item response models that implement an ideal-point-response process. The most general and widely used is the Generalized Graded Unfolding Model (GGUM; Roberts, Donoghue, & Laughlin, 2000). Under the GGUM, the probability of obtaining an observed response $U_i = u$ is defined as:

$$P\left[U_i = u \mid \theta_j\right] = \frac{\exp\left(\alpha_i\left[u\left(\theta_j - \delta_i\right) - \sum_{k=0}^{u}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M-u)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{u}\tau_{ik}\right]\right)}{\sum_{w=0}^{C}\left[\exp\left(\alpha_i\left[w\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w}\tau_{ik}\right]\right) + \exp\left(\alpha_i\left[(M-w)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w}\tau_{ik}\right]\right)\right]} \quad (3)$$

where:
θ_j = the location of respondent *j* on the continuum underlying responses,
α_i = the discrimination parameter for item *i*,
δ_i = the location of item *i* on the continuum underlying responses,
u = 0, 1, 2, . . ., *C*; *u* = 0 corresponds to the response option with the strongest level of disagreement and *u* = *C* corresponds to the response option with strongest level of agreement,
C = the number of observable response options minus 1,
w = an index for summing over observable response options 0 to *C*,
M = 2 * *C* + 1 is the number of subjective response categories, indexed *k* = 0, 1, 2. . .., *M*,
τ_{ik} = the location of the k^{th} subjective response category threshold on the latent continuum relative to the location of item *i* where $\tau_i 0 = 0$ and $\sum_{k=0}^{M} \tau_{ik} = 0$.

The GGUM equation defines the option response function for each observable response. According to the model, two subjective responses underlie each observable response; so, a respondent may agree or disagree with an item from a position that is above or below the item on the trait continuum (the formula for subjective response functions can be found in Roberts et al., 2000). Consequently, each ORF is obtained by summing the two corresponding subjective response functions, as shown in Equation 3. Figure 42.6 displays ORFs for a hypothetical item (*i*) with four response options, where 1= strongly disagree, 2= disagree, 3= agree, 4= strongly agree, $\alpha i = 2$, $\delta_i = 0$, $C = 3$, $M = 7$, $\tau_{i1} = -1$, $\tau_{i2} = -.7$, $\tau_{i3} = -.4$, $\tau_{i4} = 0$, $\tau_{i5} = .4$, $\tau_{i6} = .7$ and $\tau_{i7} = 1$. The values of τ_{ik} indicate where successive subjective response functions intersect.

GGUM item parameters may be estimated using the GGUM2004 computer program (Roberts & Fang, 2006) or Markov chain Monte Carlo (MCMC) algorithms developed in various statistical programming languages (e.g., de la Torre, Stark, & Chernyshenko, 2006; Wang et al., 2014). Although samples of 700 or more have been recommended for GGUM calibration, we have found that 400–500 may be satisfactory when the primary emphasis is on scoring.

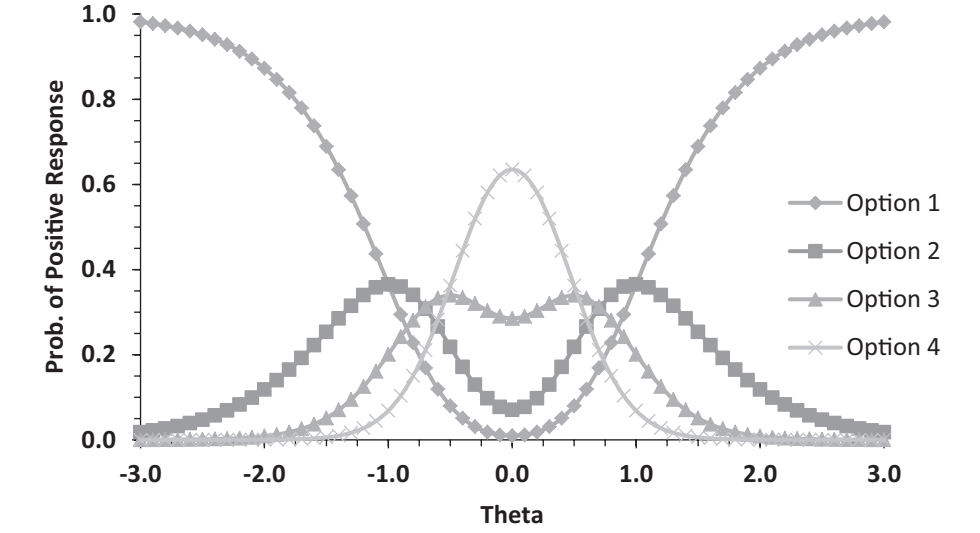

FIGURE 42.6 GGUM ORFs for an Item Having Four Observable Response Categories

Multi-Unidimensional Pairwise Preference (MUPP) Model

Historically, most noncognitive measures used for personnel screening have required respondents to indicate their level of agreement with individual statements using a Likert-type response format. More recently, however, there has been a great deal of interest in forced-choice formats that require respondents to rank or choose between two or more statements. These statements typically measure different dimensions and are matched in terms of social desirability in an effort to reduce faking and other response biases (Stark et al., 2014). Some recent examples of forced-choice measures in the personality domain are OPQ-32 (https://online.shl.com/gb/en-gb/products/opq32r), TAPAS (Stark et al., 2014), and ETS WorkFORCE Assessment (Naemi et al., 2014).

Traditional approaches to scoring forced-choice measures suffer from the problem of ipsativity (Cattell, 1944; Hicks, 1970; Salgado, Anderson, & Tauriz, 2014). A set of scales is said to be *ipsative* when the total score, obtained by summing the scale scores, is a constant. In this situation, scores can be compared meaningfully within persons, but between-person comparisons are problematic. However, IRT methods have since been developed to overcome these ipsativity problems (Böckenholt, 2004; Brown & Maydeu-Olivares, 2011; de la Torre et al., 2012; Stark, 2002; Stark, Chernyshenko, & Drasgow, 2005), thus opening new possibilities for the use of forced-choice measures in job selection.

An example of a forced-choice IRT model for pairwise preference data is the Multi-Unidimensional Pairwise Preference model (MUPP; Stark, 2002; Stark et al., 2005). The model assumes that when a respondent is presented with a pair of statements, denoted *s* and *t*, and is asked to choose the statement that is "more like you," he or she evaluates each statement independently until a preference is reached. The probability of preferring statement *s* to statement *t* in item *i*, given trait scores $(\theta_{d_s}, \theta_{d_t})$ on the dimensions, d_s and d_t represented by those statements, can be written as

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t})}{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t}) + P_s(0 \mid \theta_{d_s}) P_t(1 \mid \theta_{d_t})}, \tag{4}$$

where

$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t})$ = probability of a respondent preferring statement *s* to statement *t* in pairwise preference item *i*;

d = index for dimensions, where $d = 1, \ldots, D$, d_s represents the dimension assessed by statement *s*, and d_t represents the dimension assessed by statement *t*;

s, *t* = indices for first and second statements, respectively, in an item;

$(\theta_{d_s}, \theta_{d_t})$ = latent trait scores for the respondent on dimensions d_s and d_t, respectively;

$P_s(1 \mid \theta_{d_s})$ = probability of endorsing statement *s* given trait score θ_{d_s};

$P_s(0 \mid \theta_{d_s})$ = probability of not endorsing statement *s* given trait score θ_{d_s};

$P_t(1 \mid \theta_{d_t})$ = probability of endorsing statement *t* given trait score θ_{d_t};

and

$P_t(0 \mid \theta_{d_t})$ = probability of not endorsing statement *t* given trait score θ_{d_t}.

The probability of preferring a statement in a pairwise preference item thus depends on a respondent's trait scores, θ_{d_s} and θ_{d_t}, and the unidimensional model chosen to compute endorsement probabilities for the individual statements composing a pair. To date, the majority of measures based on the MUPP model have used the dichotomous version of the GGUM (Roberts et al., 2000), but other IRT models have been explored (e.g., Seybert, 2013). In most applications, parameters for the statements representing each dimension have been estimated by administering statements individually to large samples of examinees (e.g., 400–500) using an ordinal response format (Stark, 2002). The ordinal responses are then dichotomized and calibrated using software for the selected unidimensional IRT model. Alternatively, statement parameters may be calibrated directly from pairwise preference responses by using MCMC methods (e.g., Lee, 2016; Seybert, 2013).

With pairwise preference items that involve statements representing different dimensions, the relationship between trait levels and endorsement probabilities is represented by a three-dimensional surface, which has many peaks and valleys. An example *item response surface* for personality statements reflecting Dominance and Responsibility is shown in Figure 42.7. In the figure, values along the vertical axis indicate the probability of preferring statement *s* to statement *t* given a respondent's standing on the respective dimensions and each statement's GGUM parameters; these values were computed using Equation 4.

Note that when forced-choice items involve more than two statements (e.g., triples or tetrads), more complex IRT models are needed (e.g., de la Torre et al., 2012; Joo, Lee, & Stark, 2016; Lee, 2016). When such items involve more than two dimensions, there is no point in creating item response surfaces because they would be difficult to display and interpret.

IRT SCORING

The logic of scoring examinees in IRT is different from classical test theory (CTT). In CTT, item responses are scored dichotomously and summed over items to obtain a total (number correct) score, which may be standardized and transformed to another metric for score reporting (e.g., the IQ metric with mean 100 and standard deviation 15 or the SAT metric with mean 500 and standard deviation 100). In IRT, estimating trait levels is analogous to clinical diagnosis, where a clinician tries to estimate the most likely "disease" given a set of presenting "symptoms" (see Embretson & Reise, 2000). In IRT, the symptoms are item responses and the disease is an examinee's trait level. Note that both IRT and clinical diagnosis assume that other outcomes are also possible (an examinee may have a different trait level or a patient can have a different disease), but the diagnosed outcome (trait level) is *the most likely one*. Therefore, in IRT, scoring is a search

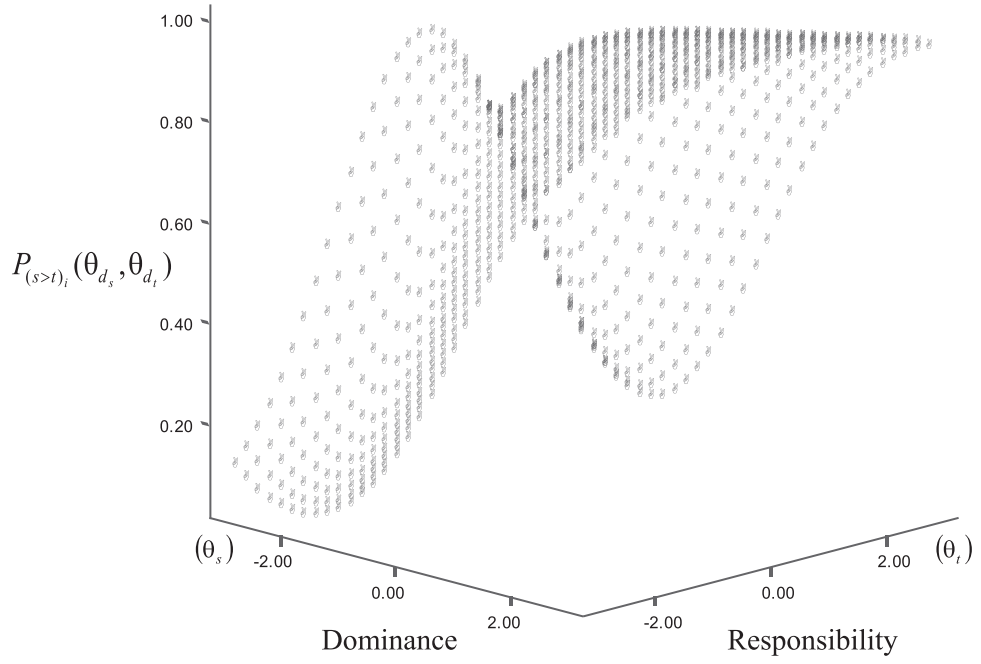

FIGURE 42.7 MUPP Item Response Surface for a Pairwise Preference Item Measuring Dominance and Responsibility

process in which the presenting behaviors (item responses and their parameters) are used to determine what trait level is most likely.

Consider, for example, an examinee with a correct and an incorrect response to two 3PLM items. Suppose the correctly answered item has an item discrimination parameter of $a = 1.0$, an item difficulty parameter of $b = -1.0$, and a guessing parameter of $c = 0.1$. Suppose the incorrectly answered item has item parameters of 1.0, 2.0, and 0.2, respectively. The conditional probability or "likelihood" of observing the response pattern (correct, incorrect), given a value of θ and the item parameters specified above, is simply the *product of the individual item response probabilities* given by Equation 1. The product rule comes from the fact that, in IRT, item responses are assumed to be independent after conditioning on θ; this is known as the *local independence* assumption. In nontechnical terms, local independence implies that the response probability for a given item is a function only of an examinee's trait level and that item's parameters; thus, the response for one item does not depend on how the examinee answers other items.

Formally, the likelihood of a response pattern, $\mathbf{u} = <u_1, u_2, \ldots u_n>$, for examinee j, given a value of θ and a vector of 3PLM parameters for item i, $\beta_i = <a_i, b_i, c_i>$, is given by

$$L(\mathbf{u}|\theta_j, \beta_1, \ldots, \beta_n) = \prod_i P_i(\theta_j)^{u_i} Q_i(\theta_j)^{1-u_i}, \quad (5)$$

where P_i is the probability of an correct response to the ith item and $Q_i = 1 - P_i$.

To illustrate the product rule written in Equation 5 above, we have multiplied the probability of the correct response to Item 1 and the probability of the incorrect response to Item 2, computed at trait levels on the interval [−3, −2.9, . . ., +3.0], and plotted the results in Figure 42.8. The resulting curve, which is called the "likelihood function," is single-peaked with a maximum at $\theta = 0.7$. The value of theta corresponding to the peak of the curve is called the *maximum likelihood estimate* (MLE) of theta and represents the value of the latent trait that makes the observed response pattern most likely. Note that this procedure for finding the MLE of theta is known as a grid search. It can be used to estimate trait scores in most situations, but more computationally efficient methods, such as Newton-Raphson iterations, are available and typically used. Readers interested in the details of these procedures should refer, for example, to Hambleton and Swaminathan (1985).

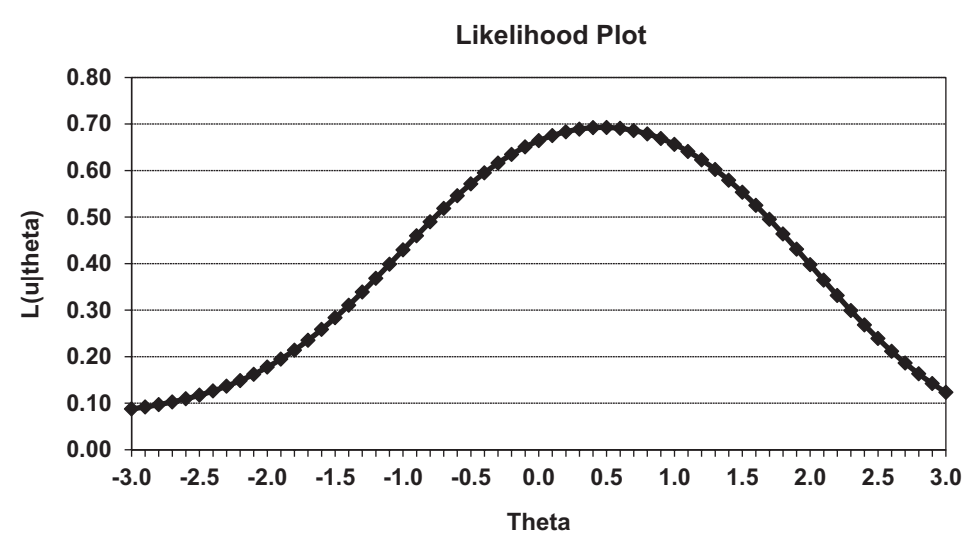

FIGURE 42.8 Likelihood of Response Pattern for a Hypothetical Examinee

With dominance models, a problem with maximum likelihood estimation is that trait scores cannot be estimated for examinees with all "correct" or all "incorrect" response patterns: the maximum of the likelihood function would occur at plus or minus infinity. Therefore, in practice, the Bayes model or expected a posteriori (EAP) estimation is used instead (e.g., Thissen & Wainer, 2001). The same is generally true for ideal point and forced-choice models. In either case, scoring is accomplished by computing the posterior likelihood of an observed response pattern, using previously estimated item parameters, and then the mode or mean of this posterior likelihood is found.

An important feature of IRT is that trait scores do not depend inherently on the specific subset of items that are administered. Specifically, examinee trait scores can be compared even if the examinees answered different subsets of items, provided that the item parameters are all on the same metric (that can be accomplished via *concurrent calibration* or *linking*; see Kolen & Brennan, 2014). This invariance property is critical to applications, such as *computerized adaptive testing* (CAT), where examinees receive individually tailored item sets to optimize measurement precision.

Information and the Precision of IRT Trait Scores

An important question for IRT applications is: How do the items administered to an examinee affect the precision of the θ estimate? The MLE procedure, described above, yields the single most likely trait score for a given response pattern and set of item parameters. A different pattern of responses or a different set of item parameters would change the shape of the likelihood function (its height and width) and possibly the precision of the trait estimate because different response probabilities would be used in the computations. In general, it is best to present items having IRFs that are steep in the ability range where an examinee is located because the resulting likelihood function will be higher and narrower, thus reducing trait score estimation error. The steepness of an IRF over a particular range of theta influences how much *information* an item provides in that part of the trait continuum. Information is an IRT concept that is commonly used to judge item quality and suitability for a particular examinee(s). The more information an item provides, the more it reduces trait score estimation error.

To examine the quality of a test form (i.e., a set of items), one can sum the item information functions to obtain the *test information function*, which shows where the test provides the best measurement. An example of a test information function for a three-item 3PLM test is presented in Figure 42.9a. As can be seen in the figure, the test information function for this short measure peaks at $\theta = 0.0$, and it is relatively high between $\theta = -1.0$ and $\theta = 1.5$. At the same time, the test provides almost no information at the extremes of the trait continuum. This translates into greater measurement precision in the central region of the trait continuum and high imprecision at the extremes, as illustrated by the standard error plot in Figure 42.9b. In IRT, *standard error* is computed as the inverse square root of test information. The error at the extremes of the trait continuum could be reduced by adding some items that provide information at high or low trait levels.

COMPUTERIZED ADAPTIVE TESTING (CAT)

Unlike traditional testing environments where one or more test forms are constructed in advance and items are administered to examinees in a prescribed order, CATs can be constructed on the fly so that each examinee receives a unique set of items that provides near-maximum information at his or her estimated trait score at every stage of an exam. At the start of a test, it is often assumed that an examinee has an average trait score on the construct being assessed. The first item is chosen to provide near-maximum information at that trait score. After answering the item, the examinee's trait score is updated, and the next item is selected to provide near-maximum information at the new trait score, subject perhaps to content and item

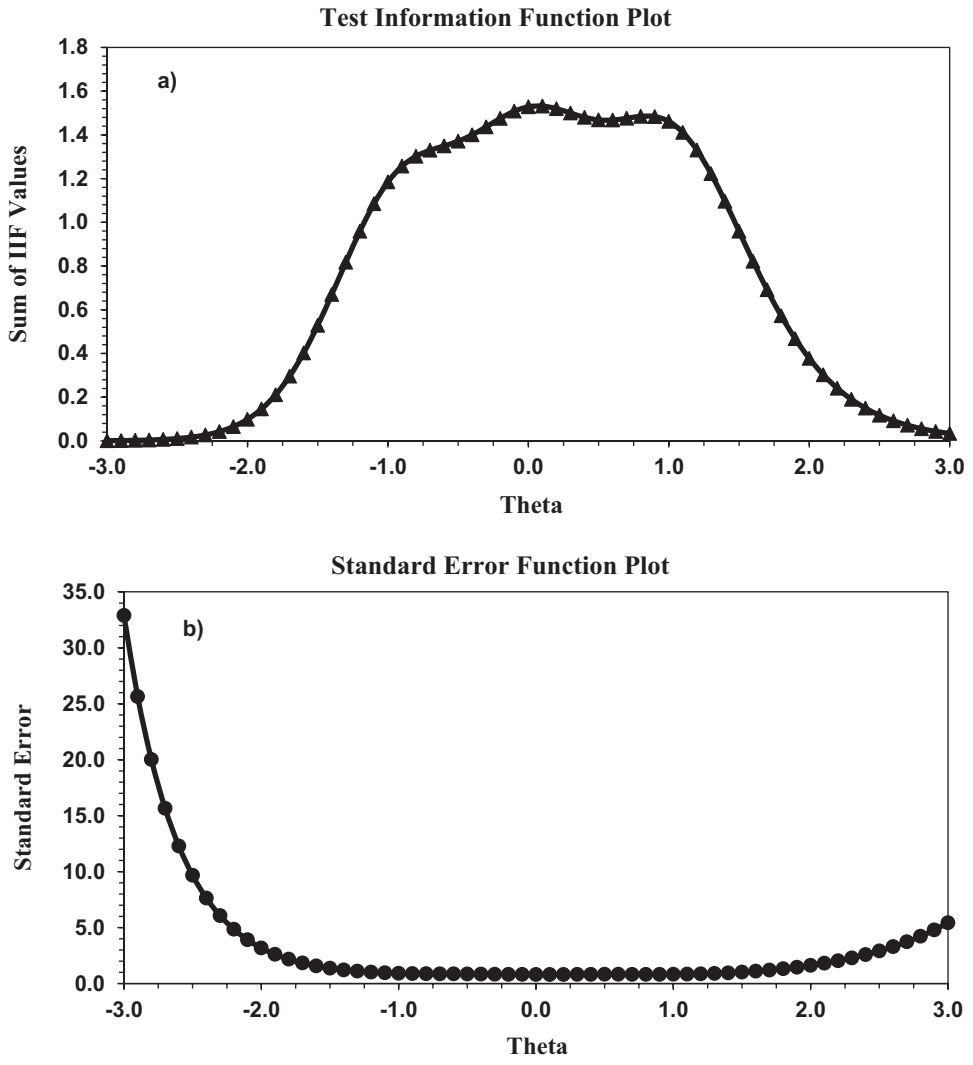

FIGURE 42.9 Test Information and Standard Error Functions for a Hypothetical Three-Item Test

exposure constraints. This process continues until a predetermined number of items has been administered or until the standard error of the trait score falls below a preset level of acceptability. (These test termination criteria are known as fixed-length and variable-length *stopping rules*, respectively.) Adaptive testing in this fashion is psychometrically efficient, often yielding precision similar to nonadaptive tests having nearly twice as many items. In addition, CATs tend to provide better accuracy and precision at extreme trait levels than do nonadaptive tests, which improves the utility of CATs for decision making and diagnostic feedback.

To illustrate CAT in more detail, consider, for example, the item information equation for the 3PLM shown below:

$$I_i(\theta) = \left[a_i^2 \frac{1 - P_i(\theta)}{P_i(\theta)} \right] * \left[\frac{\left(P_i(\theta) - c_i\right)^2}{\left(1 - c_i\right)^2} \right]. \tag{6}$$

Before an item is selected, Equation 6 can be used to compute the information provided by each available item at the examinee's estimated trait score. Selecting the item that provides the most

information is optimal in a psychometric sense but leads to items with the largest *a*-parameters being overused across testing sessions, particularly in the early stages of exams when examinees have very similar scores. Methods to control item exposure vary in complexity. The Sympson-Hetter procedure (Hetter & Sympson, 1997) is a sophisticated *item exposure control* method that uses exposure parameters, derived from simulation research, to prevent overuse. A much simpler method is to identify, for example, the top five most informative items and select one randomly from that group. Importantly, any procedure that results in less discriminating items being administered reduces the efficiency of CAT somewhat, but the added security provided by controlling item exposure may justify the cost.

Figure 42.10 presents illustrative test information and standard error functions for simulated 20-item adaptive and nonadaptive 3PLM tests. The nonadaptive tests were constructed by random selection from a diverse pool of items, while the fixed-length adaptive tests were created

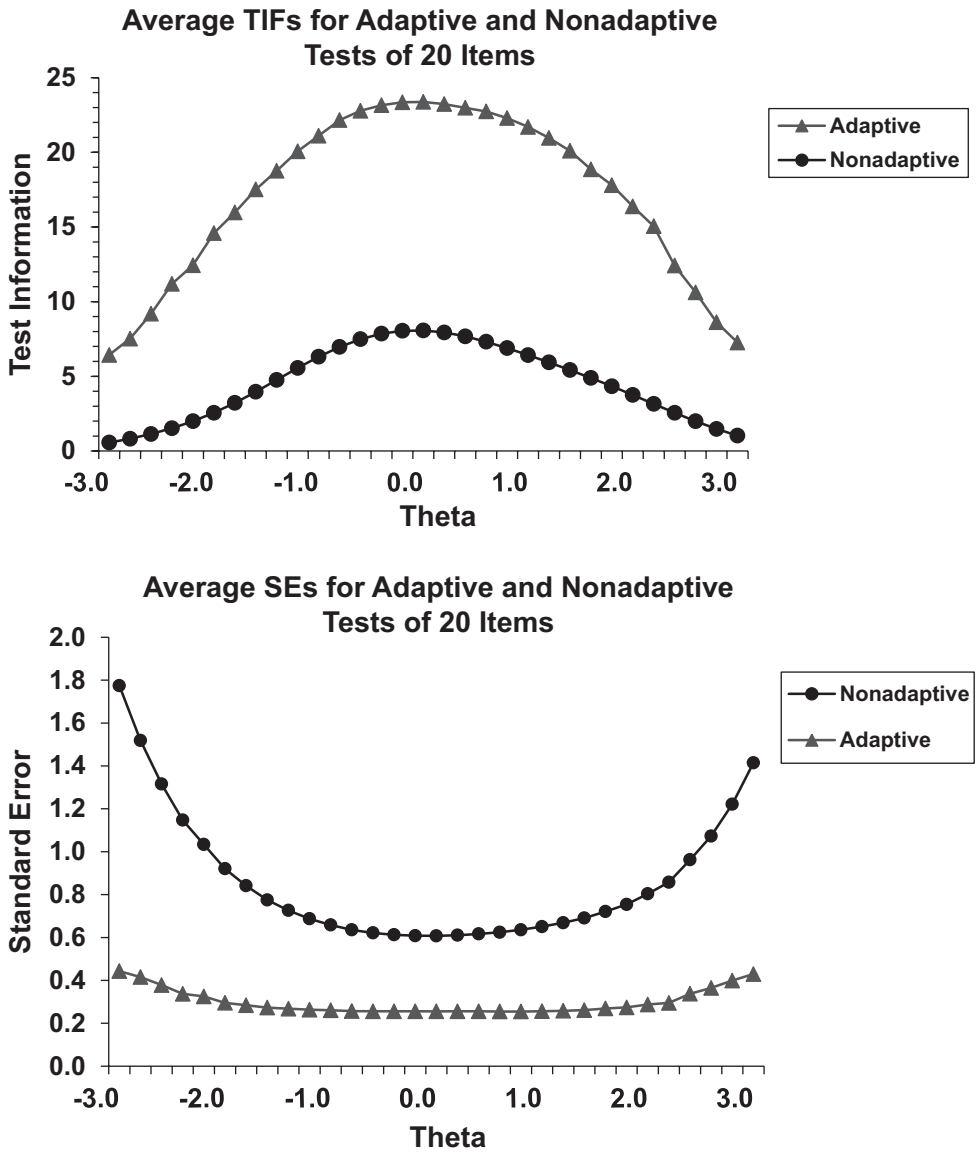

FIGURE 42.10 Comparison of Test Information Functions (TIFs) and Standard Error Functions (SEs) for Simulated 20-Item Adaptive and Nonadaptive Tests

by sequentially selecting items that provided near-maximum information at an examinee's estimated trait score at every point during a test. As can be seen in the figure, test information is highest in the central regions of the trait continuum, with adaptive tests yielding two to four times as much information as the nonadaptive tests. In addition, it can be seen that the corresponding standard error is markedly lower for adaptive tests, especially at extreme trait scores.

Examples of well-known 3PLM CATs used for personnel screening, selection, or classification include CAT-ASVAB (see Sands, Waters, & McBride, 1997) and the Proctor and Gamble Reasoning Screen (see McCloy & Gibby, 2011). CAT-ASVAB is a cognitive ability test battery used for screening and classifying U.S. military applicants. CAT-ASVAB was developed as an alternative to the paper-and-pencil ASVAB to shorten the number of items administered and time needed to evaluate applicants. Today, CAT-ASVAB consists of ten 11-item to 16-item unidimensional subtests, which all together take about 2.5 hours to complete. The Proctor and Gamble (P&G) Reasoning Screen is a 15-item *unproctored* Internet-based CAT used to screen P&G job applicants worldwide. The Reasoning Screen is available in 20 languages and contains items measuring figural, numeric, and logical reasoning. Applicants passing the Reasoning Screen must subsequently complete the *proctored* nonadaptive Reasoning Test, which serves as an independent hurdle as well as a score verification tool in the hiring process.

Although the adaptive testing principles and examples above focused on the 3PLM, it is important to note that CATs are being developed based on a variety of models to measure both cognitive and noncognitive constructs. For example, the National Institutes of Health PROMIS assessment (Reeve et al., 2007) uses SGRM-based CATs to efficiently measure a wide variety of psychological and physical health indicators. The Tailored Adaptive Personality Assessment System (TAPAS) uses MUPP-based CATs to measure a collection of narrow personality factors for screening military job applicants (Stark et al., 2014), and the ETS WorkFORCE Assessment uses MUPP-based CATs to predict applicants' job fit (Naemi et al., 2014). In noncognitive testing applications, CAT is especially useful because organizations typically want to measure many constructs in a short time. The trait scores for the various constructs may be used to form profiles or composites for evaluating the suitability of applicants for multiple job roles.

DETECTING ABERRANT RESPONSE PATTERNS

When validating and using tests for selection and licensure, it is important to screen examinee data for potential aberrant responding. This is especially true for unproctored and noncognitive tests, for which cheating and faking good are major concerns (National Research Council, 2015; Tippins et al., 2006). Unmotivated and random responding are also key concerns when pretesting new items, especially in research contexts where there are no clear incentives to answer carefully. For these reasons, many CTT and IRT methods for detecting aberrance have been developed (Drasgow, 1982; Karabatsos, 2003; Meade & Craig, 2012; Meijer & Sijtsma, 1995). One IRT index that has consistently been found to perform well is *l*z (Drasgow, Levine, & McLaughlin, 1987), which represents the standardized log likelihood of a response pattern.

For unidimensional dichotomous models (e.g., 3PLM), the log likelihood of an *n*-item response pattern can be written

$$l_0 = \sum_{i=1}^{n} u_i \log P_i(u_i = 1 \mid \hat{\theta}) + (1 - u_i) \log(1^{-u_i}(u_i = 1 \mid \hat{\theta})),$$

where $\hat{\theta}$ is an estimate of θ. The approximate expectation of this log likelihood is

$$E(l_0) \approx \sum_{i=1}^{n} P_i(u_i = 1 \mid \hat{\theta}) \log P_i(u_i = 1 \mid \hat{\theta}) + [1 - P_i(u_i = 1 \mid \hat{\theta})] \log [1 - P_i(u_i = 1 \mid \hat{\theta})]$$

.

The approximate variance is

$$Var(l_0) \approx \sum_{i=1}^{n} P_i(u_i = 1 \mid \hat{\theta})[1 - P_i(u_i = 1 \mid \hat{\theta}]\left\{\log \frac{P_i(u_i = 1 \mid \hat{\theta})}{[1 - P_i(u_i = 1 \mid \hat{\theta})]}\right\}^2.$$

Finally, the approximately standardized index is

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}} \tag{7}$$

Since then, l_z has been extended for use with multiple subtests, polytomous responses, and most recently forced-choice models (Drasgow, Levine, & McLaughlin, 1987, 1991; Lee, Stark, & Chernyshenko, 2014; Stark, Chernyshenko, & Drasgow, 2012). The l_z indices developed by Drasgow et al. focus on identifying persons who respond inconsistently with model predictions. Responding in a way that is incongruent with one's true trait scores over the course of a long test leads to large negative l_z values. Thus, based on early research showing that the distribution of l_z is approximately standard normal for long tests (e.g., 80 items), critical values for a one-tailed z-test can be used to classify response patterns as normal or aberrant. For example, if one wants to screen response patterns with a 5% false-positive rate (i.e., 5% of normal response patterns will be misclassified as aberrant), the critical l_z for a lower one-tailed z-test would be −1.65. If a respondent's observed l_z were less than the critical value, then the response pattern would be flagged as aberrant; otherwise, the pattern would be considered normal.

In addition to indices such as l_z that are generally sensitive to inconsistencies with model predictions, methods have been developed to detect specific forms of aberrance, such as patterned responding (AAA, BBB) and rapid responding (Chernyshenko, Stark, & Drasgow, 2012). With modern computer-based testing, response time has become particularly easy to track, and, in noncognitive testing, the number of items answered in less than two seconds can serve an effective flag for careless responding. Indices have also been developed to detect item pool and test compromise as well as to verify the integrity of scores on tests administered in unproctored environments (e.g., Segall, 2001, 2002; Wang, Zheng, & Chang, 2014).

Regardless of which methods are used to identify aberrant responders and vet test scores, it is incumbent upon organizations to have policies describing how flagged examinees will be treated. Nonzero false-positive rates guarantee that some percentage of examinees will be inappropriately flagged. Therefore, to promote fairness and guard against potential litigation, retesting, rather than disqualification of flagged examinees, may be the more prudent course of action.

SUMMARY AND CONCLUSION

IRT is a continuously expanding and improving technology for constructing, administering, scoring, and evaluating a variety of assessment tools. Unlike classical test theory statistics, IRT item parameters are invariant across subpopulations; person parameters do not depend on the specific set of items administered; and measurement precision can be readily evaluated as a function of trait level. These properties make IRT methods useful for computerized adaptive testing, for detecting measurement bias and assessing growth (or decline) in trait levels over time, and for model-based detection of aberrant responding. One limitation, of course, is that large samples are needed for some applications (e.g., 250 or more per group for measurement bias analyses), and minimum sample size recommendations tend to increase with model complexity.

In this chapter, we discussed just a few models and two IRT applications. However, as described in a 2015 National Research Council report entitled *Measuring Human Capabilities*, there are many more IRT models and methods for improving the quality of structured assessments used for organizational decision making. There remains, however, a pressing need for new psychometric technology to support emerging forms of assessment, such as simulations,

serious games, collaborative exercises, and constructed response tasks, which involve stochastic elements and interdependencies that most current psychometric models cannot account for (Chernyshenko & Stark, 2015; Stark, Martin, & Chernyshenko, 2015). We anticipate that future IRT research will attempt to address these challenges, and it will be interesting to see whether today's prevailing models will play a central role in future assessment programs.

REFERENCES

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

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. 397–472). Reading, MA: Addison-Wesley.

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

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9*, 453–465.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*, 460–502.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI work satisfaction scale. *Personality and Individual Differences, 49*, 743–748.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*, 292–303.

Chernyshenko, O. S., & Stark, S. (2015). Mobile psychological assessment. In F. Drasgow (Ed.), Volume 2 of the NCME Book Series. *Technology in testing: Measurement Issues* (pp. 206 – 216). NJ: Wiley-Blackwell.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. A. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523–562.

Chernyshenko, O. S., Stark, S., & Drasgow, F. (July 2012). *Investigating effects of unmotivated responding on validities of multidimensional forced choice personality tests*. Invited presentation at the 8th conference of the International Test Commission. Amsterdam, NE.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*, 88–106.

de la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2012). *Some extensions of the multiunidimensional pairwise preference model*. Paper presented at the 26th annual meeting of the Society for Industrial and Organizational Psychology. Chicago, IL.

de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement, 30*, 1–17.

Drasgow, F. (1982). Choice of test models for appropriateness measurement. *Applied Psychological Measurement, 6*, 297–308.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59–79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for multidimensional test batteries. *Applied Psychological Measurement, 15*, 171–191.

Drasgow, F., & Olson-Buchanan, J. B. (Eds.) (1999). *Innovations in computerized assessment*. Mahwah, NJ: Erlbaum.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189–199.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones Irwin.

Joo, S-H., Lee, P., & Stark, S. (April 2016). *Information functions of multidimensional forced-choice IRT models*. Paper presented at the annual conference of the National Council on Measurement in Education. Washington, DC

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking. Methods and practices*. New York, NY: Springer.

Lee, P. (2016). *Investigating parameter recovery and item information for triplet multidimensional forced choice measures: An application of the GGUM-RANK model*. Doctoral dissertation. University of South Florida. Tampa, FL.

Lee, P., Stark, S., & Chernyshenko, O. S. (2014). Detecting aberrant responding on unidimensional pairwise preference tests: An application of Lz based on the Zinnes Griggs ideal point IRT model. *Applied Psychological Measurement, 38*, 391–403.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.

Maydeu-Olivares, A., & McArdle, J. (Eds.) (2005). *Contemporary psychometrics. A Festschrift to Roderick P. McDonald*. Mahwah, NJ: Lawrence Erlbaum.

McCloy, R., & Gibby, R. (2011). Computer adaptive testing. In N. T. Tippins & S. Adler (Eds.), *Technology-enhanced assessment of talent* (pp. 153–190). San Francisco, CA: Jossey-Bass.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8*, 261–272.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.) (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Naemi, B., Seyert, J. M., Robbins, S., & Kyllonen, P. (2014). *Examining the WorkFORCE assessment for job fit and core capabilities of the FACETS engine (Research report ETS-RR-14–32)*. Princeton, NJ: ETS.

National Research Council. (2015). *Measuring human capabilities: An agenda for basic research on the assessment of individual and group performance potential for military accession*. Committee on Measuring Human Capabilities: Performance Potential of Individuals and Collectives. Board on Behavioral, Cognitive, and Sensory Sciences; Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reeve, B., Hays, R. D., Bjorner, J., Cook, K., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., Cella, D., & on behalf of the PROMIS cooperative group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcome Measurement Information System (PROMIS). *Medical Care, 45*, 22–31.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3–32.

Roberts, J. S., & Fang, H-R. (2006). GGUM2004: A Windows-based program to estimate parameters in the Generalized Graded Unfolding Model. *Applied Psychological Measurement, 30*, 64–65.

Salgado, J. F., Anderson, N., & Tauriz, G. (2014). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 88*, 797–834.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*, 1–100.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Segall, D. O. (2001). *Detecting test compromise in high-stakes computerized adaptive testing: A verification testing approach*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics*, *27*, 163–179.

Seybert, J. (2013). *A new item response theory model for estimating person ability and item parameters for multidimensional rank order responses*. Doctoral dissertation. University of South Florida. Tampa, FL.

SHL. (2015). *Occupational Personality Questionnaire (OPQ32)*. Retrieved from https://online.shl.com/gb/en-gb/products/opq32r

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment: The generalized graded unfolding model for multi-dimensional paired comparison responses*. Doctoral dissertation. University of Illinois at Urbana-Champaign. Urbana-Champaign, IL.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. *Applied Psychological Measurement, 29*, 184–201.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (July 2012). *Development of a person-fit index for multidimensional pairwise preference tests*. Invited presentation at the 8th conference of the International Test Commission. Amsterdam, NE.

Stark, S., Chernyshenko, O. S., Drasgow, F., White, L. A., Heffner, T., Nye, C. D., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26*, 153–164.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring. *Journal of Applied Psychology, 91*, 25–39.

Stark, S., Martin, J., & Chernyshenko, O. S. (2015). Technology and testing: Developments in education, work, and healthcare. In F. Leong, F. Cheung, K. Geisinger, D. Bartram, & D. Iliescu (Eds.), *ITC international handbook of testing and assessment*. NY: Oxford University Press.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Tay, L., Drasgow, F., & Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology, 94*, 1287–1304.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59*, 189–225.

Wang, W., de la Torre, J. D., Drasgow, F., Meade, T., & Louden, R. (2014). *MCMC GGUM v1.2 user's guide*. Orlando, FL: University of Central Florida.

Wang, C., Zheng, Y., & Chang, H. H. (2014). Does standard deviation matter? Using "standard deviation" to quantify security of multistage testing. *Psychometrika, 79*, 154–174.

Zimowski, M., Muraki, E., & Bock, D. (2003). *BILOG-MG*. Lincolnwood, IL: Scientific Software International.