CHAPTER 22

ITEM RESPONSE THEORY AND MUSIC TESTING

BRIAN C. WESOLOWSKI

This chapter presents an introductory overview of concepts that underscore the general framework of item response theory (IRT).

Latent Measurement

Traits, abilities, and attitudes (i.e., constructs) in music teaching and learning are not directly measurable. Examples of these constructs include but are not limited to musical aptitude, music performance achievement, musical preference, self-efficacy, affective response to music, musical expectancy, and so forth. These constructs can be easily explained by descriptive criteria and qualitative attributes; however, they cannot be *directly measured*. The measurement of these constructs can only be inferred *indirectly* through the measurement of secondary behaviors that are considered to be theoretically representative of the construct. Any construct that cannot be directly measured but rather inferred through the measurement of secondary behaviors is considered to be *latent*.

In order to infer a latent construct from a secondary behavior, an apparatus must be constructed with the intent to collect data that specifically gathers empirical evidence of these secondary behaviors. Items must be carefully constructed with the explicit intent to provoke persons' responses that directly reflect the latent construct. The data collected from the interaction between each person's observed response and each item provide observable, empirical evidence that can serve as a starting point to establish inferences. A mere ordering of observed responses by persons who answered the least correct of items to persons who answered the most correct of items is not measurement, however. Similarly, an ordering of observed responses of items that were answered the least correct to items that were answered the most correct is not measurement. These are

simply examples of ordinal rankings based on proportion-correct observed responses. This data answers the question of "how many" but not "how much" In order to answer the question of "how much," the implementation of a measurement model is necessary. Measurement models provide a mechanism for transforming observed responses into estimated measures of person ability and item difficulty. It is only with the implementation of a measurement model that persons and items can be validly, reliably, and fairly compared. Furthermore, it is only with the implementation of a measurement model that inferences can be drawn as to how well the items empirically define the intended construct being measured. The implementation of measurement models into steps of the scientific method, therefore, is necessary for meaningfully connecting the substantive theory of a latent construct with the measurement of persons and items.

"Item response theory," or latent trait theory, is broad umbrella term used to describe a family of mathematical measurement models that considers observed test scores to be a function of latent, unobservable traits (Birnbaum, 1957, 1958a, 1958b, 1967; Lazarsfeld, 1950; Lord & Novick, 1968). Item response theory uses probabilistic distributions of responses as a logistic function of person and item parameters in order to define a latent construct. In other words, IRT models provide methods of data analysis that use the latent characterizations of objects of measurement (i.e., persons) and latent characterizations of agents of measurement (i.e., items) as predictors of observed responses in order to empirically define a latent construct. Figure 22.1 represents a conceptual operationalization of a unidimensional latent construct.

The construct represented in Figure 22.1 has several notable characteristics:

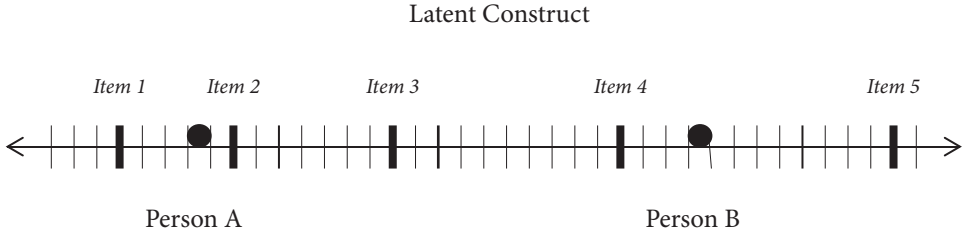

FIGURE 22.1 Operational definition of a unidimensional latent construct with calibrations of persons and items.

1. The latent construct is represented by a unidimensional, continuous line;
2. The line, acting as a scale of measurement (i.e. a "ruler"), is marked off in equal, interval-level units;
3. The items are calibrated to the line with a relative positioning that reflects each item's difficulty level; and
4. The persons are calibrated to the line with a relative positioning that reflects how much or how little of the construct each person possesses.

Items calibrated to the line from left to right represent less difficult items to more difficult items. Another way to conceptualize an item calibration is the item's rank ordering based on its discriminatory ability to best distinguish between persons at various locations on the continuum. This is discussed later under the section titled "Item Information Function." As demonstrated in Figure 22.1, item 1 is the least difficult item and item 5 the most difficult item. Similarly, persons calibrated on the line from left to right represent persons with less possession of the latent construct (e.g., lower ability or lower achievement) to persons with more possession of the latent construct (e.g., higher ability or higher achievement). In the example presented in Figure 22.1, Person B has a higher ability than Person A.

One important premise of IRT is to ascertain a *conceptual measurement* of a person's ability using a *conceptual ruler* the same way one would ascertain a *physical measurement* of one's height using a *physical ruler*. As a substantive example, assume a measurement apparatus was constructed in order to measure musical aptitude. A person's musical aptitude is not directly observable in the same way that the measurement of a person's height is. Therefore, the amount of musical aptitude one possesses can only be inferred based on a person's responses to test items that theoretically represent the construct of musical aptitude. In this case, the continuous line would represent the unidimensional construct of musical aptitude. The line acts as a conceptual "ruler" that is marked off in equal interval-level log odds (i.e., logit) units. Logits are discussed later in the section titled "Constructing Interval Units on the Continuum: Log Odds Units." Items are developed to represent the construct based on the test constructor's theory of specific observable tasks that represent musical aptitude. Each person would then interact with (i.e., respond to) each item, resulting in a collection of observed responses. Assuming the items called for dichotomous (i.e., correct/incorrect) scoring, each correct response would be marked with a score of "1" and each incorrect response would be marked with a score of "0." Once all of the observed responses are collected and dichotomously scored, both items and persons could be rank ordered based on their observed proportion-correct responses. Then, the implementation of an appropriate IRT measurement model would transform the observed responses into estimated linear measures. These estimated measures supply information that indicates: (1) which items appropriately define the latent construct of musical aptitude; (2) how well the items define the latent construct of musical aptitude; (3) how well the items discriminate between persons at various ability levels; (4) which persons were appropriately measured; and (5) how much musical aptitude those persons possess.

It is important to note that a measurement apparatus is a conceptual representation, or operational definition, of the developer's definition of the latent construct. Although the content of the apparatus is driven by theoretical, research-based principles and understandings, the unique collection of items is only one possible operationalization of the construct. Item response theory is the mechanism that provides an empirical rationale for the developer's definition of the operationalization.

Item Response Functions and Item Characteristic Curves

The example of musical aptitude provides one instance of a music assessment context where multiple persons respond to multiple items. This can be more specifically characterized by a single person's response (s) to a single item (i) resulting in an individual interaction (X_{is}). For a dichotomous item, only two possible *observed* outcomes for each interaction can be achieved: (1) a correct response ($X_{is}=1$); and (b2) an incorrect response ($X_{is}=0$). In order to model the *probability* of these responses as a distribution of the persons on a latent continuum, the ability of the person ($\theta \in (\infty,+\infty)$) must be parameterized as a logistic function of the item's difficulty ($b_i \in (-\infty,+\infty)$). An *item response function* (IRF) is the function of a person's ability (θ_s) to an item's difficulty (b_i). In other words, the IRF is a mathematical function that relates the latent construct to the probability of a single person answering a single item correctly. The probability of a correct response is denoted as $P(x_{is}=1)$, and the probability of an incorrect response is denoted as $P(x_{is}=0)$. The IRF is a logistic function, meaning that the probability of a correct response $P(x_{is}=1)$ increases with respect to the increasing position of a person's ability (θ_s) on the unidimensional continuum. Conversely, the probability of an incorrect response $(P(x_{is}=0))$ increases with respect to the decreasing position of a person's ability (θ_s) on the unidimensional continuum.

Each IRF can be characterized by a monotonically increasing function called an item characteristic curve (ICC). Figure 22.2 is a graphical depiction of an ICC.

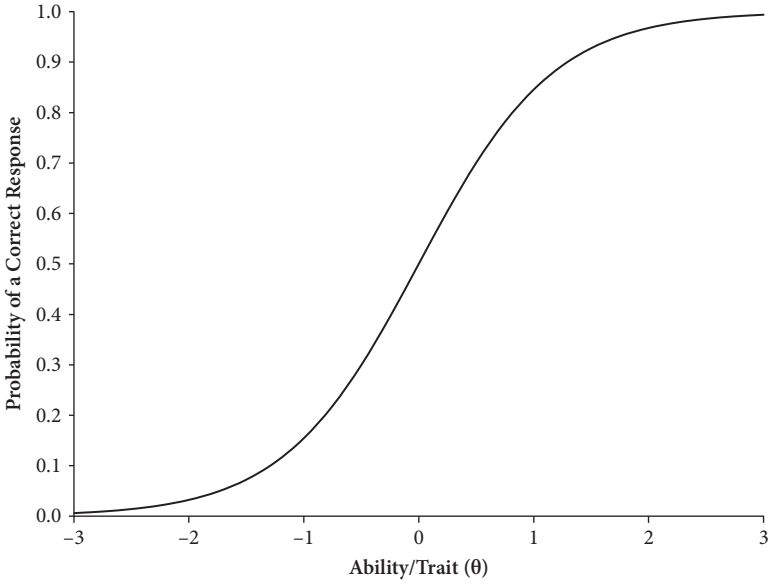

FIGURE 22.2 Graphical depiction of an item characteristic curve (ICC).

The abscissa (i.e., horizontal axis) represents person ability on the latent construct ($\theta \in (-\infty, +\infty)$), and the ordinate (i.e., vertical axis) represents the probability of a correct response in the range between 0 (0% chance of a correct response) and 1 (100% chance of a correct response). The S-shaped curve (i.e., ogive) is a graphical representation that provides a full illustration for items at all ability levels on the latent continuum.

The ICC is defined by the following mathematical expression that connects a person's probability of success on an item to their ability on the latent continuum:

$$P(x_{is} = 1) = \frac{\exp(b_i - \theta_s)}{1 + \exp(b_i - \theta_s)},$$

where:

$P(x_{is} = 1)$ = the probability that person *s* answers item *i* correctly;
b_i = a threshold parameter for the difficulty of item *i*;
θ_s = a threshold parameter for the ability of person *s*.

An important characteristic of the ICC is its inflection point. For most IRT models, the inflection point represents the intersecting point at which the probability of answering an item correct is .50. This can also be described as the threshold for where a person's odds change from a 50% chance of answering an item incorrectly to a 50% chance of answering an item correctly. The item parameters control the location of the IRFs. Therefore, this function indicates that a person possessing more of the latent trait should have a higher chance of correctly answering a more difficult item than a person possessing less of the latent trait. As discussed later in the chapter, this intersection point is of importance because this is the point at which there is the most item information. More specifically, the inflection point is the point at which the item maximally discriminates persons.

Constructing Interval Units on the Continuum: Log Odds Units

In the same way that inches or centimeters are equal divisions of a physical ruler used to measure height, log odds units (i.e., logits) are equal divisions of a conceptual ruler used to measure items and persons on a latent construct. If we are interested in measuring musical aptitude, we may ask each person to respond to four items (i_1, i_2, i_3, i_4). If person 1 (θ_1) answers i_1 and i_2 correctly and i_3 and i_4 incorrectly, they would receive an observed sum score of 2. If θ_2 answers i_1 and i_2 incorrectly and i_3 and i_4 correctly, they would also receive an observed sum score of 2. However, let us suppose that items i_3 and i_4 were more difficult than items i_1 and i_2. Do θ_1 and θ_3 deserve to have the same observed sum score? If we assume for this example that the items and persons are functioning according to the expectations of the model, the answer is no. The next logical questions would then be "How much do the items differ?" and "How much do the persons differ?" In order to answer these two questions, there must be a mechanism in place to compare how much

the questions differ in difficulty and how much the persons differ in ability. Therefore, a nonlinear transformation of proportion scores for both items and persons must be made to interval-level units. A logistic transformation of the nonlinear proportion correct scores of θ_1 and θ_2 and the proportion correct scores of i_1, i_2, i_3, and i_4 provides the mechanism needed to answer the "How much?" question. The transformation assigns both the persons and items an estimated interval-level measure (i.e., calibration of items and persons on the latent continuum) in logits.

This logit scale is developed independently of both the particular items included in the test as well as the particular persons being measured due to the assumption of parameter invariance. Wright (1993) notes:

> when any pair of logit measurements have been made with respect to the same origin on the same scale, the difference between them is obtained merely by subtraction… the logit scale is unaffected in variations in the distribution of measures that have been previously made, or which items…may have been used to construct and calibrate the scale. The logit scale can be made entirely independent of the particular group of items that happen to be included in a test…or the particular samplings of persons that happen to have been used to calibrate the items. (p. 288)

The transformation produces values that theoretically fall between $-\infty$ and ∞. The example figures provided in this chapter limit those values to a more practical range of −3.0 to 3.0. A logistic transformation is defined as:

$$\Psi[x] = \ln\left[\frac{x}{1-x}\right],$$

where:

$\Psi[x]$ = logit transformation for x;
ln = natural logarithm;
$[x / x{-}1]$ = proportion correct responses.

Engelhard (2013) clearly delineates between the logistic transformations of persons versus the logistic transformations of items. Person logits are defined by:

$$\Psi[p] = \ln\left[\frac{p}{1-p}\right],$$

where:

p = number of correct items/total number of items for x.

Item logits are defined by:

$$\Psi[(p-value)] = \ln\left[\frac{(p-value)}{-(p-value)}\right],$$

where:

(*p-value*) = number of correct responses/total number of persons responding to the item.

The logit units allow for the comparison of items and persons in a meaningful way that answers the question of "how much."

Assumptions of Item Response Theory

The following section describes the assumptions of item response theory, including parameter invariance, unidimensionality, and local independence.

Parameter Invariance

Parameter invariance indicates an equality of item and person parameters from different person populations or measurement conditions. In other words, person and item parameters are sample independent. Item parameters are independent of (i.e., invariant across) the ability levels of persons responding to them. Likewise, person parameters are independent of (i.e., invariant across) the items measuring the ability of the persons.

The importance of parameter invariance comes in the form of providing inferences of generalizability between person ability and item difficulty. In order for generalizations to be valid, measurement models must be used that provide measurement conditions where parameters are invariant. Measurement models that do not maintain properties of parameter invariance succumb to the variability attributed to the sample. In other words, estimations of item difficulty and person ability are based on the observed interactions of the sample within each individual assessment context. Parameter invariance is estimated for unidimensional IRT models when θ is normally distributed with $M = 0$ and $SD = 1$. As a result, any variation in item or person estimates across different samples from the same population is considered to be a result only of measurement error. Perfect parameter invariance, however, is considered to be a measurement ideal that can never be achieved (Engelhard, 2013). Therefore, perfect model data fit is never expected.

Unidimensionality

Unidimensionality implies that persons and items can be described by a single latent construct. From a psychological perspective, unidimensionality refers to the specific construct that influences a person's performance (McNamara, 1996). Unidimensionality is possible when the items collectively measure the same weighted composite of ability. More specifically, the psychometric assumption of unidimensionality is met when: (1) all of the items used in the measurement apparatus measure the same construct; and (2) persons only use their ability on the construct to respond to the test items.

Wright and Linacre (1989) indicate that the empirical analysis of dimensionality can be addressed in three steps:

1. Analyze the relevant data according to a unidimensional measurement model;
2. Find out how well and in what parts these data do conform to our intentions to measure; and
3. Study carefully those parts of the data, which do not conform, and hence cannot be used for measuring, to see if we can learn from them how to improve our observations and so better achieve our intentions.

When the assumption of unidimensionality is true, local independence may be obtained (Lord, 1980; Lord and Novick, 1968).

Local Independence

The graphical representation of the ICC in Figure 22.2 represents one ICC. In other words, it represents only one probable outcome $P(x_{is}=1)$ resulting from the interaction between an item difficulty (b_i) and a person ability (θ_s). Multiple items, however, are necessary to operationalize latent constructs. Each item maintains its own representative ICC, and the likelihood of a person's success on an item is a represented by a function of only their ability related to the latent trait and the characteristic of that item. Parameters are therefore considered to be conditionally independent when each item response is independent given each examinee's position on the latent continuum. This means that after controlling for θ_s the item responses should be uncorrelated. Local independence provides statistically independent probabilities of item responses and can be characterized by the following function:

$$P(X_{is1}=x_1, X_{is2}=x_2 \,|\, \theta_s) = P(X_{is1}=1 \,|\, \theta) P(X_{is2}=1 \,|\, \theta).$$

The assumption of local independence posits that only the characteristics of the test items and person's ability relate to the construct being measured. As an example, if a test item somehow aids the test taker in correctly responding to another item, the assumption of local independence is violated.

Although Lord (1980) considered local independence to be met if the underlying assumption of unidimensionality was met, Hambleton, Swaminathan, and Jane Rogers (1991) argue that local independence and unidimensionality are distinct qualities: "local independence will be obtained whenever the complete latent space has been taken into account" (p. 11). It is argued that the results of item dependency present potential implications of biased parameter estimation, inflated reliability estimates, false estimations of measure precision, and artificially small estimates of standard error. Each of these implications has potential to affect the overall dimensionality of the measure, thereby making unidimensionality and local independence separate issues of concern.

Common Item Response Theory Models

There are over 100 IRT models that can be classified into six basic categories: (1) models for items with binary-scored and/or polytomously-scored response categories, (2) non-parametric models, (3) models for response time or multiple attempts, (4) models for multiple abilities or multiple cognitive components, (5) models for nonmonotone items, and (6) models with special assumptions about response models (van der Linden & Hambleton, 1997). In this chapter I examine three of the more common models for binary response (dichotomous) items: one-, two-, and three-parameter logistic models and two of the more common models for ordered response (polytomous) response items: the partial credit (PC) model and the rating scale (RS) model.

One-Parameter Logistic Model

The one-parameter logistic (1-PL) model (Rasch, 1960) predicts the probability of a correct response from an interaction between a person's ability and one parameter: the item difficulty parameter. In other words, a person's chance of answering an item correctly is based on the relationship between the person's ability and how difficult the item is. The 1-PL model is mathematically specified as follows:

$$P(X_{is}=1\,|\,\theta_s, b_i) = \frac{\exp[a(\theta_s - b_i)]}{1+\exp[a(\theta_s - b_i)]},$$

where:

θ_s = ability of person *s*;
b_i = difficulty of item *i*;
a = common discrimination parameter.

If persons *s*'s ability is greater than the difficulty of item *i*, than θ_s will be greater than .00 logits and the probability of answering the question correctly will be greater than 50%. Conversely, if a person *s*'s ability is less than the less than the difficulty of item *i*, than θ_s will be less than .00 logits and the probability of answering the question correctly will be less than 50%. The item difficulty parameter (b_i) represents an index of item difficulty that corresponds to the value of θ_s at the ICC's inflection points of the curve.

A unique characteristic of the 1-PL model is that there is a common discrimination parameter (*a*). The discrimination parameter provides information on how related the item is to the latent trait. More specifically, it is the slope of the curve as it crosses the probability of .50. A result of holding this parameter constant is that the items do not cross, meaning they are all equally discriminating. This indicates that the measures produced by the 1-PL model are sample-free for the agents of measurement (i.e., items) and test-free for the objects (i.e., persons). This sample/item-free characteristic is a property of measurement referred to as *specified objectivity*, and only occurs in the 1-PL

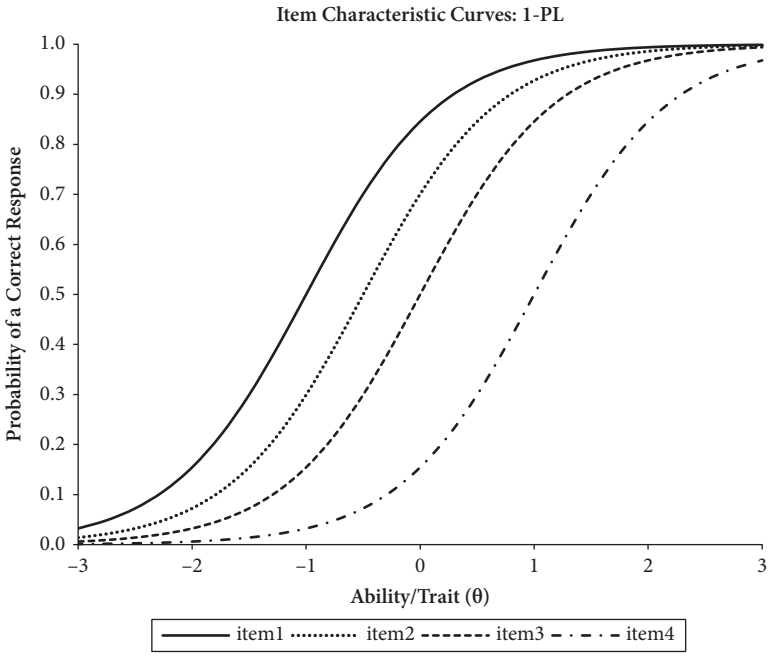

FIGURE 22.3 Graphical depiction of four item characteristic curves (ICCs) in the context of a one-parameter logistic model (1-PL): i_1 ($a_1 = 1.00$; $b_1 = -1.00$; $c_1 = 0.00$), i_2 ($a_2 = 1.00$; $b_2 = -0.50$; $c_1 = 0.00$), i_3 ($a_3 = 1.00$; $b_3 = 0.00$; $c_1 = 0.00$), i_4 ($a_4 = 1.00$; $b_4 = 1.00$; $c_1 = 0.00$).

model. The implication of employing only the item difficulty parameter is the displacement of the ICC from left to right. In all other aspects, the ICCs are identical. Figure 22.3 demonstrates four items with varying difficulty levels.

Each of the item ICCs run parallel to each other and do not cross because they have a common discrimination parameter that results in the same slope. As a result, each item is equally discriminating. The differences in the items are their difficulty, or the location (left to right) on the latent trait. The difficulty level is indicated by where the point of inflection occurs across the horizontal axis. Item 1 (i_1) has a difficulty level of −1.00 logits, i_2 has a difficulty level of −.50 logits, i_3 has a difficulty level of .00 logits, and i_4 has a difficulty level of 1.00 logits. Therefore, person s (θ_s) with a value of −1.00 has a 50% chance of correctly answering i_1 correctly, person s (θ_s) with a value of −.50 has a 50% chance of correctly answering i_2 correctly, and so forth.

Two-Parameter Logistic Model

The two-parameter logistic (2-PL) model (Birnbaum, 1957, 1958a, 1968) predicts the probability of a correct response to a test item based on two parameters: item difficulty (b_i) and item discrimination (a_i). The 2-PL model is mathematically specified as follows:

$$P(X_{is}=1\,|\,\theta_s, b_i, a_i)=\frac{\exp[a_i(\theta_s-b_i)]}{1+\exp[a_i(\theta_s-b_i)]},$$

where:

θ_s = ability of person s;
b_i = difficulty of item i;
a_i = item i's discrimination parameter.

In the equation represented by the 1-PL model, the common discrimination parameter was represented by a. In the equation representing the 2-PL model, a_i represents an item discrimination parameter that is freed to vary by item. The item discrimination parameter in the 2-PL model, therefore, describes the unique relationship of each item to the latent trait.

Figure 22.4 graphically depicts four items. Items i_1 and i_2 have the same difficulty level (b_1 = −1.00; b_2 = −1.00). Note the crossing of the ICCs at the intersection of .50 probability of a correct response and −1.00 logits. However, they differ in their discrimination slope (a_1= .50; a_2 = 1.00). Note the flatter slope of i_1 compared to the steeper slope of i_2. The same is true for items i_3 and i_4. Items i_3 and i_4 have the same difficulty level (b_3 = .00; b_4 = .00). Note the crossing of the ICCs at the intersection of .50 probability of a correct

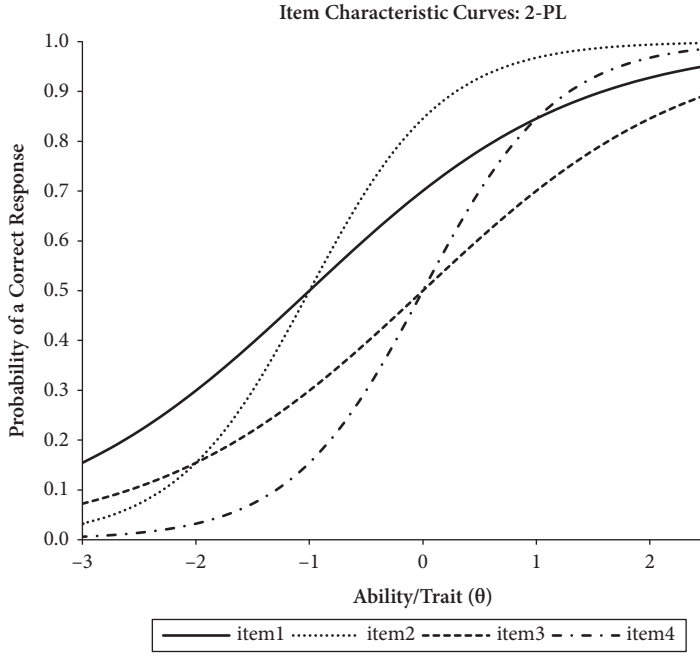

FIGURE 22.4 Graphical depiction of four item characteristic curves (ICCs) in the context of a two-parameter logistic model (2-PL): i_1 (a_1 = 0.50; b_1 = −1.00; c_1 = 0.00), i_2 (a_2 = 1.00; b_2 = −1.00; c_1 = 0.00), i_3 (a_3 = 0.50; b_3 = 0.00; c_1 = 0.00), i_4 (a_4 = 1.00; b_4 = 0.00; c_1 = 0.00).

response and .00 logits. However, they differ in their discrimination slope (a_3 = .50; a_4 = 1.00). Note the flatter slope of item i_3 compared to the steeper slope of item i_4. The discrimination is an index represented by the steepness of the ICC at its inflection point. Parameters with larger a_i values will demonstrate steeper ICCs, and parameters with smaller a_i values demonstrate flatter ICCs. Substantively, this means that items i_2 and i_4 are stronger items, as they have stronger discriminating power. However, from a visual perspective, both pairs of ICCs cross at their inflection point, changing the ordering of persons and items. For a person where (θ_s) = −1.00, items i_3 and i_4 are more difficult. However, for a person where (θ_s) = 2.00, item i_1 is more difficult than item i_4. As opposed to the 1-PL model, the 2-PL model violates the measurement characteristic of specified objectivity, where the ordering of the persons and the ordering of the items does not remain constant across the continuum of theta values.

Three-Parameter Logistic Model

The three-parameter logistic (3-PL) model (Birnbaum, 1957, 1958a, 1968) describes the relationship between person ability and the probability of a correct response using three parameters: difficulty, discrimination, and guessing. The 3-PL model is mathematically specified as follows:

$$P(X_{is}=1\,|\,\theta_s, b_i, a_i, c_i) = c_i + (1-c_i)\frac{\exp[a_i(\theta_s - b_i)]}{1+\exp[a_i(\theta_s - b_i)]},$$

where:

θ_s = ability of person s;
b_i = difficulty of item i;
a_i = discrimination parameter for item I;
c_i = lower asymptote for item i.

The difference between the 3-PL model and the 2-PL and 1-PL models is found in the following portion of the equation: $c_i + (1–c_i)$. The parameter c_i represents the item lower asymptote, or the lower bound of probability that is independent of theta. This is often referred to as the "guessing" parameter. Because the probability is independent of the item difficulty and item discrimination parameters, the probability of answering a question correctly starts with the estimation of c_i ($c_i > 0$) then becomes dependent on θ_s, a_i, and b_i. Figure 22.5 demonstrates four items.

The items represented in Figure 22.5 have the same difficulty values and discrimination values as demonstrated in the 2-PL model example. In this example, however, the lower asymptote varies by item. Item i_1 has a lower asymptote value of .20, item i_2 has a lower asymptote value of .40, item i_3 has a lower asymptote value of .10, item i_4 has a lower asymptote value of .15. These values can be interpreted as the probability of a correct response as a result of guessing. In comparing Figure 22.5 to Figure 22.4, the compression of the ICCs as a result of adding the new parameter causes the slope to

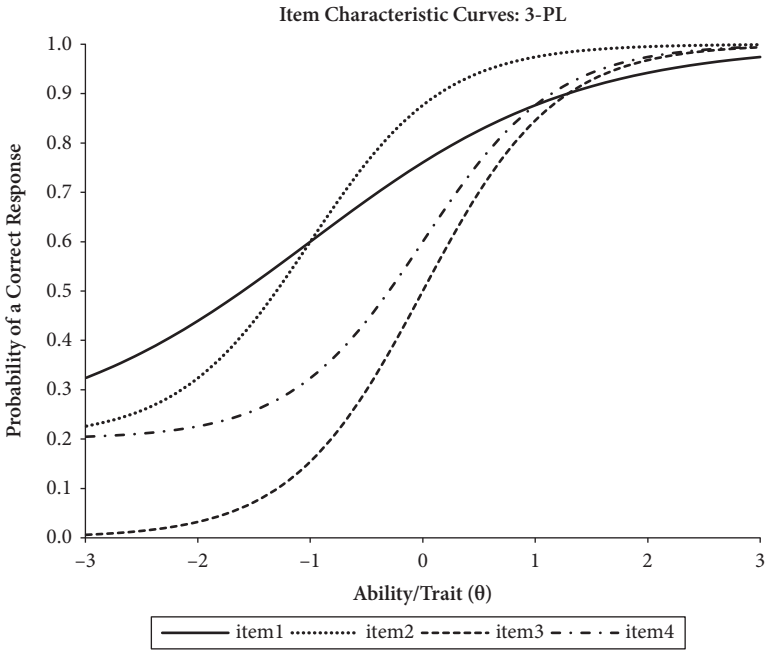

FIGURE 22.5 Graphical depiction of four item characteristic curves (ICCs) in the context of a three-parameter logistic model (3-PL): i_1 ($a_1 = 0.50$; $b_1 = -1.00$; $c_1 = .20$), i_2 ($a_2 = 1.00$; $b_2 = -1.00$; $c_1 = .30$), i_3 ($a_3 = 0.50$; $b_3 = 0.00$; $c_1 = .10$), i_4 ($a_4 = 1.00$; $b_4 = 0.00$; $c_1 = .15$).

decrease, or become more flat. Because the values have been compressed between c_i and 1.00, the discriminatory power of the item is reduced. Therefore, the larger the c value for an item, the less it can discriminate between persons, and the less item information it provides.

The 3-PL model is most popular in high-stakes educational settings and is most appropriately used when persons possessing low ability level answer difficult items correctly. The drawback, however, is that c_i is often difficult to estimate because very often there are few individuals with low θ values that provide helpful item responses to estimate c_i. The implication is that large sample sizes are needed to estimate the values with adequate precision.

The Rasch Model

Historical debates over "which model is better" inundate the psychometric and educational assessment literature. Most notable is the debate between Benjamin Wright and Ronald Hambelton (Wright, 1992) in the context of the model selection for the measurement of academic achievement. These debates are plentiful and too great to cover in this short chapter. However, there is one important distinction worth noting: the relationship between the Rasch model and the 1-PL model.

The Rasch model (Rasch, 1960) is often considered a pseudonym for the 1-PL model. However, the Rasch model is a specialized model with philosophical and developmental underpinnings that contrast with the 1-PL model or any other IRT models. The most important philosophical difference is the notion of "model-data fit" versus "data-model fit." The IRT paradigm argues that the rationale for the choice of one model over another is that the chosen model accounts better for the observed data. In this paradigm, most texts compare and contrast the models as 1-PL versus 2-PL versus 3-PL, and so forth. The Rasch perspective, however, focuses on the compatibility of measurement with properties of invariance and the quantitative laws of fundamental measurement. Andrich (2004) notes:

> the main challenge in the traditional paradigm is to those with expertise in statistics or data analysis to identify a model that accounts better for the given data, notwithstanding that they may find other problems in the data; the main challenge in the Rasch paradigm is for those with expertise in the substantive field of the construct to understand the statistical misfits as substantive anomalies and, if possible or necessary, to generate new data that better conform to the model while enhancing substantive validity of the variable. (p. I–15)

Item Information Function

In IRT, item information refers to the value of the ability parameter. More specifically, it is an index that represents the item's ability to discriminate between persons. Fisher (1925) defined information as the reciprocal of the precision with which a parameter could be estimated. In IRT, precision is the standard error of measurement, or more broadly, the variance of the latent trait. Information, then, is the reciprocal of variance, and can therefore be connected to reliability. Under the CTT paradigm, reliability is equal to true variance divided by the added sum of true variance and error variance. Therefore, reliability is equal to information divided by the sum of information plus 1. The more information, the more precise the estimate of a person's ability. The less information, the less precise the estimate of a person's ability.

As an example, Figure 22.6 provides a visual representation of four item characteristic curves (ICCs) and item information functions (IIFs) for the 1-PL, 2-PL, and 3-PL models.

For the 2-PL model, items i_2 and i_4 have steeper slopes, indicating more discriminating power. However, more discriminating power does not equate to more information. The nonlinearity of the slopes at the extremes indicates less information. For persons with theta values between −1.00 and .00, items i_2 and i_4 are informative. However, for persons with theta values between 1.00 and 2.00, the items have less information. Additionally, because the discrimination varies freely by item, information is different for each item. If persons have a theta value between .00 and 1.00, item i_3 has more information than items i_1 and i_2. In this example then, the question of "how much information" is affected by both the theta value and the item. Information is maximized

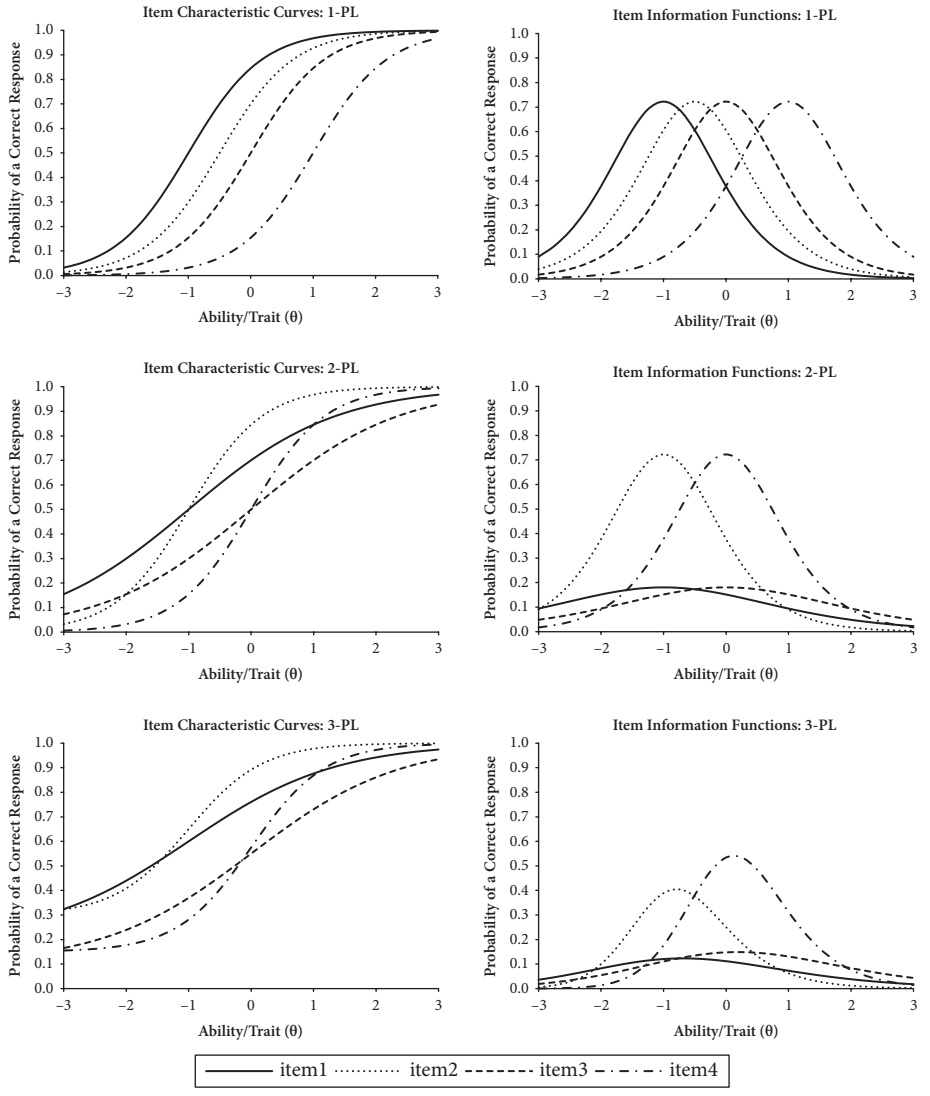

FIGURE 22.6 Item characteristic curves (ICCs) and item information functions (IFFs) for 1-PL, 2-PL, and 3-PL models.

at each ICC's inflection point and decreases as the ability level approaches the extremes of the IIF. Evaluation of the IIF for item i_2 in the 2-PL model demonstrates that the maximum amount of information is at the theta value of −1.00. In other words, the item best discriminates persons with an ability level of −1.00. In models other than the 1-PL where each item discriminates similarly, items can differ in both discrimination and difficulty, each affecting the item information. Therefore, both parameters play a role in selecting *how* good an item is, *where* the item is good, and *for whom* the item is good.

Test Information Function and Standard Error of Measurement

Figure 22.7 provides a graphical depiction of test information functions (TIFs) and their reciprocal standard errors of measurement (SEMs) for the 1-Pl, 2-PL, and 3-PL models.

Item information, because it is on an interval-level scale, has additive properties. Therefore, test information can be computed through the sum of each item response functions over all the items on the test. Test information provides the information, or reliability, of a test at any given ability level. The test information function is defined as follows:

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta),$$

where:

I = test information at a given ability level (θ);

I_i = the amount of information for item i at ability level θ;

N = total number of items.

Test information is valuable, as it provides detailed levels of precision at all ability levels across the latent continuum. This information then provides insight into how the test functions in relation to the latent trait and provides diagnostic information in terms of particular areas to be targeted in adding or removing items. The range of a test information function is 0 to the number of items. The TIF approaches 0 as ability approaches ∞ and approaches the most information as ability approaches $+\infty$. Therefore, a TIF is an increasingly monotonic function of ability.

Standard error of measurement is the reciprocal of test information, and is useful in building confidence intervals around an ability estimate. The lower the SEM, the higher the information. Conversely, the higher the SEM, the lower the information. The SEM function is defined as follows:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}} = \frac{1}{\sqrt{test\ information}}.$$

Procedures for Estimating Ability

Unlike the classical test theory paradigm, where persons are scored by summing the correct responses to items and converting the observed sum score to a standardized score, person measures in IRT are estimated based on persons' corresponding probabilistic

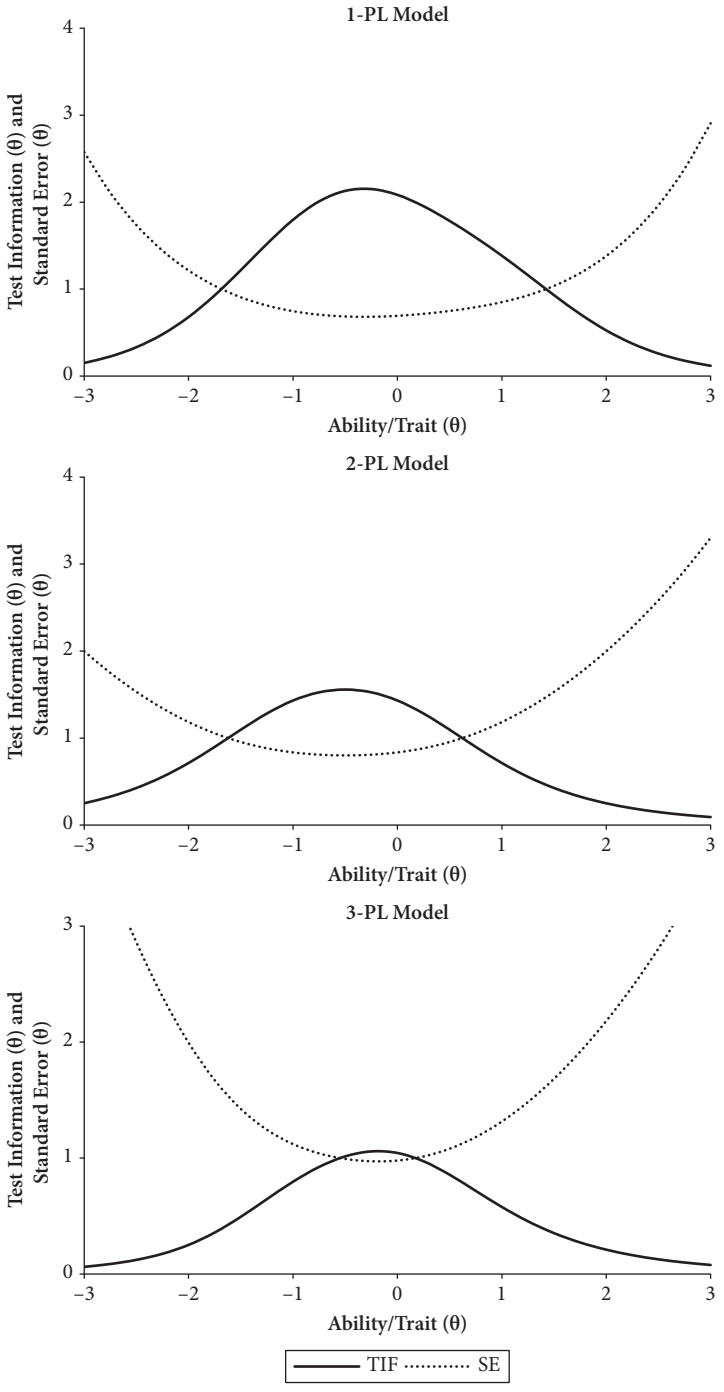

FIGURE 22.7 Test information functions (TIFs) and standard errors of measurement (SEMs) for the 1-Pl, 2-PL, and 3-PL models.

response pattern to a set of items. As an example, assume a person with high ability on the latent construct answers difficult items correctly. This is a likely scenario, as a person with a high ability has a higher probability of answering more difficult items correctly. Similarly, assume a person with low ability on the latent trait answers difficult items incorrectly. This, too, is a likely scenario, as a person with a low ability has a higher probability of answering more difficult items incorrectly. If a person with a low ability on the latent construct answers a more difficult question correctly, a less probable scenario is created. In this case, the IIF to this question would violate the probabilistic patterning.

In estimating the ability of a person, then, a value of θ is sought to maximize the highest likelihood for the predicted item response pattern based on the proposed measurement model. As discussed earlier, the assumption of local independence posits that each item information function is an autonomous reference to the underlying trait (see Figure 22.7). If item information functions serve as a referent, than the converse (the probability of an incorrect response) can be expressed as:

$$Q_i = 1 - P(X_i \mid \theta),$$

IRT models only express the possibility of a correct response. Therefore, in order to express joint probability of both answering items correctly and incorrectly, a likelihood function must be employed.

Under the umbrella of IRT, there are many procedures for estimation and many rationales for choosing among them, each with its own strengths and weaknesses. Due to the space limitation of this chapter, one of the more conventional approaches to ability estimation, *maximum likelihood estimation* (MLE), is discussed.

Likelihood Function

In an assessment context where items are restricted to binary response options, the response pattern (i.e., the number of possible outcomes) is represented by 2^k. If an assessment has four items, the total amount of distinct response patterns is 16. Joint probability of item responses is possible due to the assumption of local independence. Persons' responses to individual items are conditionally independent functions of their theta value. Therefore, they can be multiplied to obtain the probability of the pattern. A likelihood function is the expression of joint probability: the probability of both answering items correctly and incorrectly. As an example, the raw likelihood function of one item is expressed as follows:

$$L(u_{s1}, u_{s2} \ldots, u_{sI} \mid \theta_s) = \prod_{i=1}^{I} P_i(\theta_s)^{u_{si}} Q_i(\theta_s)^{1-u_{si}}.$$

This equation expresses the likelihood of a person's correct or incorrect response to an item and each successive item (in difficulty) through the last item. As an example, suppose a person takes a four-item exam. If each of the four items is ranked by difficulty

and the person answers questions 1 and 2 correctly and 3 and 4 incorrectly, it could be expressed as $x_1 = 1$, $x_2 = 1$, $x_3 = 0$, and $x_4 = 0$. Within the context the equation, this can be expressed as:

$$L(x_{s1}, x_{s2} \ldots, x_{sI} \mid \theta_s) = (P_1^1 Q_1^0)(P_2^1 Q_2^0)(P_3^0 Q_3^1)(P_4^0 Q_4^1).$$

With consideration to last two equations, the scores lie in the range of 0 to 1. Therefore, when joint probabilities are calculated, their quotient becomes increasingly small as test items increase. Therefore, the raw likelihood function is transformed to a log-likelihood function by calculating the natural logarithm of each IRC. The log-likelihood function is expressed as follows:

$$-\log L(u_{s1}, u_{s2} \ldots, u_{sI} \mid \theta_s) = \sum_{i=1}^{I} u_{si} \log[P_i(\theta)] + (1 - u_{si}) \log[Q_i(\theta)].$$

Although the likelihood functions provide a brief overview of the conceptual idea behind estimation, this process is not necessarily convenient for large datasets consisting of many persons and many items.

The MLE function provides one of the more conventional and efficient methods for locating the exact maximum of the log likelihood in a pattern of person responses. The MLE function is an iterative method of estimation for obtaining item and person locations. More specifically, the Newton-Raphson method is a popular procedure for converging on an MLE in a manner that successively improves the estimation. The MLE procedure is a complex mathematical procedure that includes many steps to converge on an estimate. Full details and examples can be found in Embretson and Reise (2000).

Common Models for Polytomous Response Items

Due to the responsive nature of music assessment contexts, student performances cannot always be appropriately measured in binary (right/wrong) terms. Music assessment and music psychology contexts often require more compelling evaluation experiences in order to more clearly define the latent construct. Therefore, many music assessment contexts necessitate the need for items requiring ordered responses, such as Likert-type items, semantic differential items, or other ordered-category items. Polytomous IRT models are extensions of binary models that examine the interaction between person response and ordered response categories. The principles of IRT in the context of binary responses can be extended to contexts where polytomous, ordered responses are more appropriate. These contexts, however, are much more complex. Therefore, I provide a brief overview of two common models that may be most relevant to the field of music from an introductory perspective: the partial credit (PC) model and the rating scale (RS) model.

For all polytomous IRT models, item information can be represented at both the item level and the category level. Category information can be represented as the log of the category response probability:

$$I_{ik}(\theta) = -\frac{\partial^2}{\partial \theta^2} \log P_{ik}(\theta),$$

where:

$I_{ik}(\theta)$ = information for category k of item i across the ability range of θ;
$P_{ik}(\theta)$ = probability of responding to in category k of item i.

The category information can then be summed to produce item information:

$$I_i(\theta) = \sum_k^m I_{ik} P_{ik}.$$

The Partial Credit Model

The partial credit (PC) model (Masters, 1982) is an extension of the 1-PL model that allows for the assignment of "partial credit" to a series of steps within a technical problem. As an example, if an item prompts person s to visually identify a chord in a written example of a sonata, this may include four distinct tasks: (1) identification of the root; (2) identification of the chord type; (3) identification of the inversion; and (4) proper figured bass labeling of the chord. For this particular item, four categories exist, each with a distinct probability for answering correctly and each with an independent difficulty threshold. This model is ideal for testing the assumption of the particular ordering. The PC model is mathematically specified as follows:

$$P_{ik}(\theta_s) = \frac{\exp \sum_{j=0}^{k} (\theta_s - \delta_{ik})}{\sum_{i=0}^{m-1} \exp \sum_{j=0}^{k} (\theta_s - \delta_{ik})},$$

where:

$P_{ik}(\theta)$ = probability of person s responding in category k for item i;
δ_{ik} = difficulty (i.e., location) of the category threshold parameter for item i.

Item information for the PC model is specified as follows:

$$I_i(\theta) = \sum_k k_i^2 P_{ik} - \left(\sum_k k_i P_{ik} \right)^2,$$

where:

$I_i(\theta)$ = information evaluated across the range of θ across item i summed across k categories ($k = 0, 1, \ldots, m$);
$P_{ik}(\theta_s)$ = probability of person s responding in category k of item i.

The δ_{ik} term is often referred to as a "step difficulty" of moving from category k to the adjacent category (k-1). For an item that has four distinct categories, three difficulty thresholds exist. If we assume that a monotonic relationship between categories exists where category b is more difficult than category a, category c is more difficult than category b, and category d is more difficult than category c, than three category thresholds exists: (1) the difficulty of moving from category a to category b (δ_{1i}); (2) the difficulty of moving from category b to category c (δ_{2i}); and (3) the difficulty of moving from category c to category d (δ_{3i}). Figure 22.8 illustrates an example of category response curves for a polytomous item with four response categories.

Each of the category thresholds is defined by where each of the category response curves intersect. For the exemplar item in Figure 22.8, the threshold parameters are $\delta_{1i} = -2.00,\ \delta_{2i} = -0.50,\ \delta_{3i} = 0.50$. At these intersection points, each step to the next category becomes probabilistically more likely to move to the next step as θ increases. The important conceptual idea to consider in applying the PC model is that categories are most often not equal in difficulty. Therefore, summed category responses in a polytomous item are equally as problematic as summed binary responses in a dichotomous item and warrant a similarly important empirical investigation.

The Rating Scale Model

The rating scale (RS) model (Andrich, 1978) allows for the analysis of polytomous items where the ordered response format is the same across all items. An example may include a Likert-type scale where the response categories are identical for each item (strongly

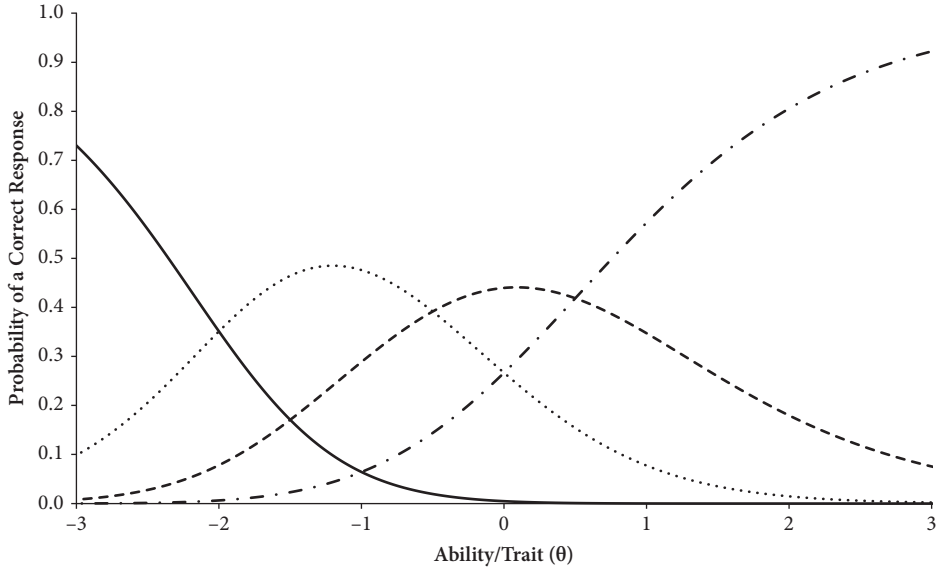

FIGURE 22.8 Category response curves for a four-category polytomous item.

agree, agree, disagree, strongly disagree). Each item is described by having a single scale location parameter (λ_i) that reflects the difficulty of the item. In the RS model, a category threshold measure (δ_j) is similarly described for all items in the measure. In other words, the RS model specifies that all of the items within the measure share the same rating scale structure. The RS model is mathematically specified as follows:

$$P_{ik}(\theta_s) = \frac{\exp \sum_{j=0}^{k} (\theta_s - (\delta_i + \tau_k))}{\sum_{i=0}^{m-1} \exp \sum_{j=0}^{i} (\theta_s - (\delta_i + \tau_k))},$$

where:

$P_{ik}(\theta_s)$ = the probability of person s responding in category k ($k = 0, 1, \ldots, m$) of item I;
δ_i = the location (i.e., difficulty) of the item parameter;
τ_k = the common category threshold (i.e., boundary) for all items.

Item information for the PC model is specified as follows:

$$I_i(\theta) = \sum_k k_i^2 P_{ik} - \left(\sum_k k_i P_{ik} \right)^2,$$

where:

$I_i(\theta)$ = information evaluated across the range of θ across item i summed across k categories ($k = 0, 1, \ldots, m$);
$P_{ik}(\theta)$ = probability of responding to in category k of item i.

Note that the model specification of item information for both the RS and PC models is the same. Both the PC and RS models as presented here adhere to the requirement of specified objectivity as in the 1-PL model. As a result, the item discriminations are both held constant and the model is in the same metric. However, $P_{ik}(\theta)$ will yield different results across the same dataset, thereby providing an overall difference in the analysis between the two models.

Linacre (2000, p. 768) provides nine important detailed considerations in deciding between a PC model and a RS model: (1) design of the items, (2) communication with the audience, (3) size of the dataset, (4) construct and predictive validity, (5) fit considerations, (6) constructing new items, (7) unobserved categories, (8) statistical information, and (9) antifragility (the simpler the better).

Multidimensionality

In contrast to unidimensional models, where only one parameter represents a person, multidimensional item response theory (MIRT) models include two or more parameters to represent persons (Reckase, 2009). Although unidimensional models are more sensible in educational contexts where most tests purport to assess one specific construct,

arguments have been made that unidimensional models do not sufficiently model complex domains often associated with many psychological areas where multiple latent ability dimensions are manifested simultaneously or nested within a more broad construct as either compensatory (i.e., latent ability scores for multiple dimensions are assumed to be independent and combine additively to influence the probability of a correct response) or noncompensatory (latent ability scores for multiple dimensions are assumed to be confounded and combine multiplicatively to influence the probability of responding to an item correctly) factors. Where the probability of a correct response is dependent on a single estimate of ability (θ) for unidimensional IRT models, MIRT models posit that the probability of a correct response is dependent on a vector θ on K-dimensional, continuous latent ability dimension (θ_k).

There are a wide variety of MIRT models for a multitude of data types that are too exhaustive for this chapter (see Reckase, 2009). Therefore, only the basic multidimensional extensions of the 1-PL, 2-PL, and 3-PL models are outlined.

The multidimensional extension of the 1-PL model can be expressed as:

$$P(X_{is}=1\,|\,\theta_s,d_j)=\frac{\exp(\sum_m \theta_{sm}+d_i)}{1+\exp(\sum_m \theta_{sm}+d_i)},$$

where:

x_{is}= response of person *s* to item *i*;
θ_{sm}= ability for person *s* on dimension *m*;
d_i = easiness intercept for item *i*.

The multidimensional extension of the 2-PL model can be expressed as:

$$P(X_{is}=1\,|\,\theta_s,d_i,a_i)=\frac{\exp(\sum_m a_{im}\theta_{sm}+d_i)}{1+\exp(\sum_m a_{im}\theta_{sm}+d_i)},$$

where:

x_{is}= response of person *s* to item *i*;
θ_{sm} = ability for person *s* on dimension *m*;
d_i = easiness intercept for item *i*;
a_{im} = discrimination for item *i* on dimension *m*.

The multidimensional extension of the 3-PL model can be expressed as:

$$P(X_{is}=1\,|\,\theta_s,d_i,a_i,c_i)=c_i+(1-c_i)\frac{\exp(\sum_m a_{im}\theta_{sm}+d_i)}{1+\exp(\sum_m a_{im}\theta_{sm}+d_i)},$$

where:

x_{is}= response of person *s* to item *i*;
θ_{sm} = ability for person *s* on dimension *m*;
d_i = easiness intercept for item *i*;
a_{im} = discrimination for item *i* on dimension *m*;
c_i = lower asymptote for item *i*.

Substantive Implications for Measurement and Evaluation in Music

In an educational climate that is becoming increasingly data-driven, the field of music is at a clear pivot point where the "subjectivity" of music making is becoming progressively interconnected with the "objectivity" of measuring student achievement and program accountability. The field can no longer ignore the demands of having to provide empirical data that validly, reliably, and fairly reflect both students' growth in the classroom and program effectiveness. Furthermore, the field can no longer withstand minimal approaches to collecting, analyzing, interpreting, and disseminating such data, as there have never been higher consequences for the implications of its interpretation and use. It is the field's ethical responsibility to provide sound empirical data and robust assessment processes that reflect true student learning in meaningful ways. In order to do so in a valid, reliable, and fair manner, the field must (1) recognize the complex nature of measuring and evaluating musical constructs, (2) spend considerable time and energy gaining a more grounded and fundamental understanding of psychometric theories, and (3) provide better mechanisms and opportunities for training researchers and practitioners in appropriate selection and implementation of psychometric theories.

Item response theory is a viable option for music researchers who want to enhance the data analysis and measurement instrument construction processes. As the field of music becomes more familiar with the theories and applications of IRT, important music assessment issues may continue to be addressed and reexamined from new, interesting, and informative perspectives.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Applications of Rasch Analysis in Health Care*, *42*(1), I7–I16.

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems* (Series Rep. No. 58–16, Project No. 7755- 23). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1958a). *On the estimation of mental ability* (Series Rep. No. 15, Project No. 7755-23). Randolph Air Force Base, TX USAF School of Aviation Medicine.

Birnbaum, A. (1958b). *Further considerations of efficiency in tests of a mental ability* (Tech. Rep. No. 17, Project No. 7755-23). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Birnbaum, A. (1967). *Statistical theory for logistic mental test models with a prior distribution of ability* (Research Bulletin No. 67-12). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novik (Eds.), *Statistical theories of mental test scores* (Chapters 17–20). Reading, MA: Addison-Wesley.

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

Engelhard, Jr., G. E. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Psychology Press.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725. doi: 10.1017/S0305004100009580

Hambleton, R. K., Swaminathan, H., & Jane Rogers, H. (1991). Fundamentals of Item Response Theory. Newberry Park, CA: SAGE.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Princeton, NJ: Princeton University Press.

Linacre, J. M. (2000). Model selection: Rating scale model (RSM) or partial credit model (PCM)? *Rasch Measurement Transactions*, *12*, 641–642.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Masters, G. (1982). A Rasch model for partial credit scoring. *Applied Measurement in Education*, *1*, 279–298.

McNamara, T. F. (1996). Measuring second language performance. New York, NY: Longman.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Wright, B. (1992). *IRT in the 1990s: Which Models Work Best? 3PL or Rasch?* Ben Wright's opening remarks in his invited debate with Ron Hambleton, Session 11.05, AERA Annual Meeting 1992.

Wright, B. D. (1993). Logits? *Rasch Measurement Transactions*, *7*, 288.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*, 857–860.