Supplement to Chapter 16

Overview of Item Response Theory

Item response theory (IRT) is an alternative measurement model to classical test theory (CTT) for measuring unobservable constructs; in IRT, these constructs are called *latent traits*. IRT is best thought of as a *family* of statistical methods and theoretical models for analyzing item-level data from a composite, multi-item measure of a trait.

This Supplement provides a brief overview of IRT and a discussion of why IRT methods are advantageous. IRT is highly complex, and so interested readers are urged to consult other sources for more in-depth coverage. Fairly nontechnical overviews of IRT are offered by DeMars (2010) and Embretson and Reise (2000). Bond and Fox (2015) provide accessible descriptions of the related Rasch family of models.

BASICS OF ITEM RESPONSE THEORY

IRT analyses involve modeling the probability of people's response to an item as a function of the underlying trait and one or more **item parameters**. In IRT models, every respondent is assumed to have a true location on a continuous latent trait dimension, and the person's location on the continuum is assumed to underlie how he or she responds to an item. The latent trait being estimated in IRT is sometimes referred to as **theta (θ)**.

In IRT models, the amount of the latent trait is expressed on a continuum that is like a standard score, with a mean of zero and a standard deviation of 1. The standardized metric or "ruler" can be used to locate both people, in terms of the amount of the trait they possess, and items, in terms of how "difficult" it is to endorse them. Thus, model-based estimation is used to separate the measurement properties of the person's responses to items on the one hand and the person's underlying level of the trait being measured on the other.

Item Response Theory Models

IRT models provide a mathematical equation to characterize the relationship between the probability of a person's response to an item and the amount of his or her latent trait. One important way in which IRT models vary concerns how many item parameters are estimated. In brief, IRT includes models for one, two, three, and four item parameters. The one-parameter IRT model (1-PL) is similar to another model, the Rasch model. Three-parameter (3-PL) and four-parameter (4-PL) models are used infrequently in health research and will not be discussed.

Another key way in which IRT models vary concerns whether the items in the analysis are **dichotomous** or **polytomous** (i.e., three or more response options). The most basic IRT models are for dichotomous items, in which there are only two response options. In ability tests, the dichotomy is between *correct* and *incorrect*. In health scales, the more typical dichotomies are for responses of *yes* versus *no,* or *present* versus *absent* (e.g., for symptoms or conditions). Polytomous IRT models are used with items that have three or more ordered response categories, reflecting intensity or frequency of a symptom, feeling, or condition. Most self-report health scales rely on polytomous items.

Item Characteristic Curves

A basic feature of an IRT analysis is the **item characteristic curve** (ICC). An ICC is commonly defined as an S-shaped (logistic) function that models the relationship between people's responses to an item and their level of the latent trait. Figure 1 presents an example of an ICC for a single dichotomous item. In this figure, the x-axis represents the latent trait continuum, labeled theta (θ). The y-axis represents the probability of endorsing the item, ranging from 0.0 to 1.0.

As an example, let us assume that the latent trait for the item in Figure 1 is depression, with negative values on the latent trait corresponding to low levels of depression and higher positive values corresponding to increasingly more severe levels of depression. Suppose the item is: "Sometimes I feel unhappy," to which people answer *yes* or *no*. For this hypothetical item, the probability of endorsing the item is 0.5 (i.e., a 50–50 probability) for those whose value on the depression trait is exactly at the mean of zero. For those who are more depressed, the probability of endorsing the item increases. For example, the probability increases to about 0.75 for those with a depression trait value of 1.0. These visual representations of item properties are one of the many attractive features of IRT and provide useful information for selecting items that cover the desired range of the trait.

ITEM PARAMETERS IN ITEM RESPONSE THEORY

Item characteristic curves can vary along four dimensions: their location along the trait continuum theta, the steepness of their slopes, and where they flatten out at the bottom or at the top. These four dimensions correspond to four potential item

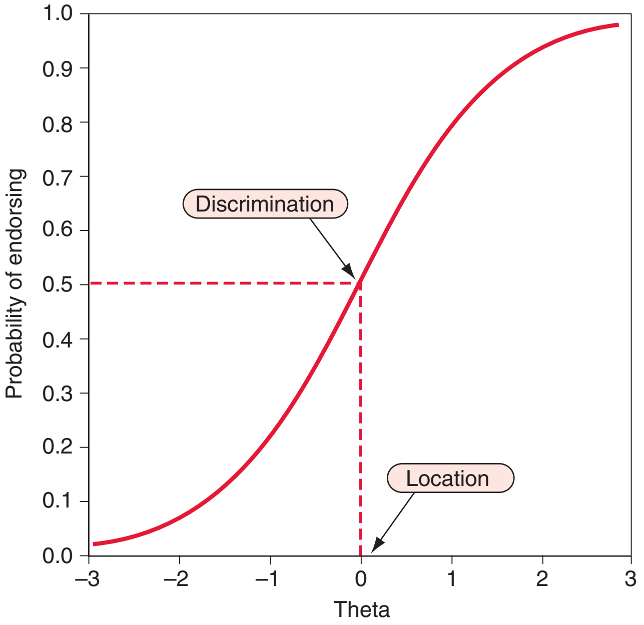

FIGURE 1 Item characteristic curve for a dichotomous item.

parameters, but only the first two are estimated in the development of health scales. In estimating item parameters, a set of items believed to comprise a unidimensional scale must be administered to a large sample of respondents.

Item Difficulty (Location)

The term **item difficulty** is used in educational testing to describe how difficult an item is along an ability trait to achieve a 50% probability of a correct response. The more difficult the test question, the higher a students' ability must be to have a 50-50 probability of answering the question correctly. An item's difficulty shows where along the trait continuum the item functions best. All IRT models estimate the difficulty of items under consideration.

In health fields, the term **item location** is sometimes used in lieu of "difficulty." However, one can also conceptualize an item for health-related constructs as being more "difficult" to endorse among people who do not have high levels of the trait. For example, it is more "difficult" for people who are only slightly depressed to agree with the item "Sometimes I think about committing suicide" than to agree with the item "Sometimes I feel unhappy." In Figure 1, the location of the ICC for the dichotomous "unhappy" item centers at the mean level on the depression continuum of 0.0. The ICC for the "suicide" item would be located far to the right on the theta continuum. By determining an item's difficulty level (location), researchers can establish how much of the trait is required for a person to have a specified probability of endorsing the item. In Figure 1, the location parameter (symbolized as *b*) for the "unhappy" item is equal to 0.0.

Item Discrimination (Slope)

The **item discrimination** parameter provides information about the degree to which an item can unambiguously differentiate between those whose trait level is below the item location and those whose trait level is above it. Item discrimination is also called the *slope parameter*, with steeper slopes at a particular theta level offering better discrimination than less steep slopes, as depicted on the ICC. In Figure 1, the ICC's steep slope directly above the trait level of 0.0 (at the 0.5 probability point) indicates good discrimination at that level. The discrimination parameter (symbolized as *a*) in ITT is similar to an item-total correlation in CTT.

Figure 2 shows ICCs for two dichotomous items that have identical location parameters ($b = 0.0$ for both), but that differ in slope. Item 1 is a better (more discriminating) item, with a steeper slope and $a = 1.5$. The flatter ICC of item 2 ($a = 0.5$) reflects an item with greater ambiguity. In other words, item 1 discriminates more effectively than item 2 between those who endorse or do not endorse the item. This illustrates how the ICCs of items can help scale developers to better understand the strengths and weaknesses of individual items.

ITEM RESPONSE THEORY AND RASCH MODELS

Item location and discrimination are the basic elements of IRT models. In this section, we briefly describe three IRT models: the one-parameter logistic model (1-PL), the Rasch model, and the two-parameter logistic model (2-PL).

The One-Parameter Logistic Model

The **one-parameter logistic (1-PL) model** is an IRT model that includes only the item location (difficulty) parameter. In a 1-PL model, it is assumed that only the underlying trait and the item's location influence a person's response to an item. In a 1-PL model, the slopes of the ICCs (the items' discrimination) are assumed to be the same. That is, it is assumed that the ICCs are parallel and do not cross each other. Figure 3 presents two ICCs that have the same discrimination but different locations on the trait continuum, one at −.5 and the other at +.5. This is an ideal situation for a 1-PL model.

The Rasch Model

In the literature on IRT, Rasch and 1-PL models are often discussed as though they were synonymous. The **Rasch model** is similar to the 1-PL IRT model in many respects, the most noteworthy being that both models estimate only one parameter—item difficulty/location. Although the Rasch and 1-PL models

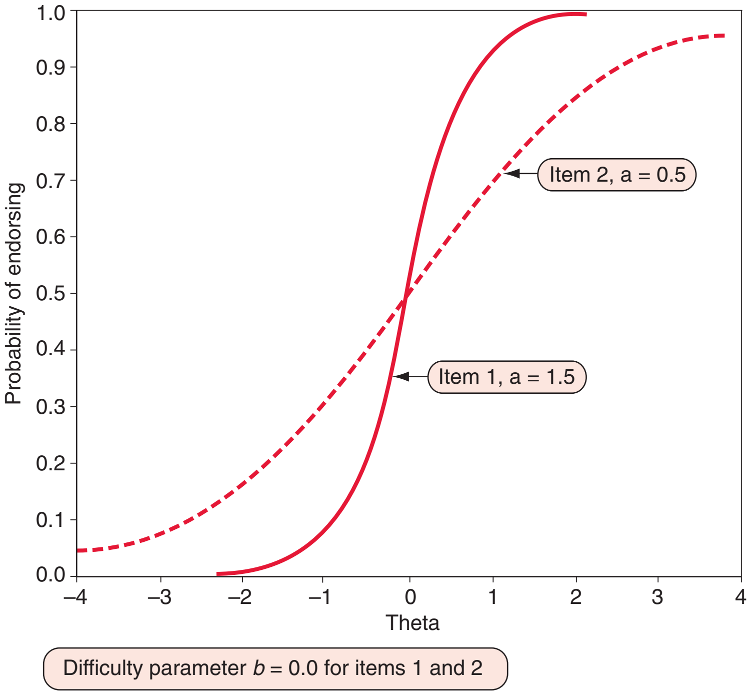

FIGURE 2 Item characteristic curves for two dichotomous items varying in item discrimination.

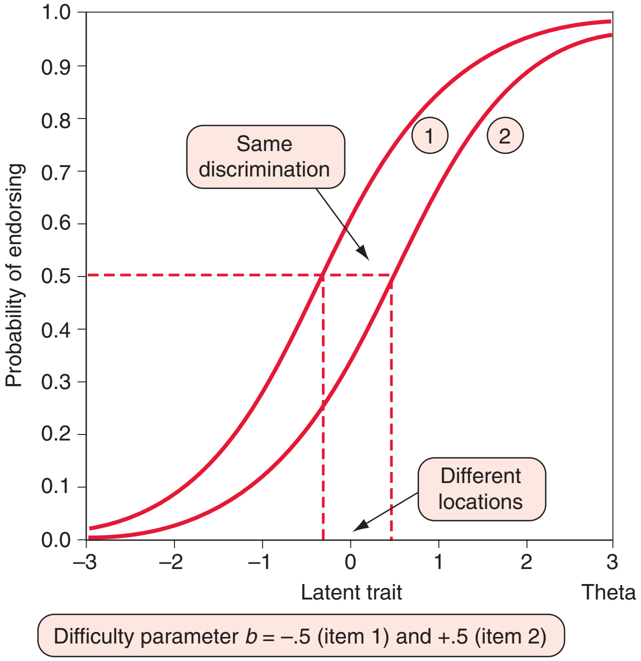

FIGURE 3 Item characteristic curves for two dichotomous items varying in item location.

are similar, there are some conceptual differences, as described more fully in Polit and Yang (2016). One major difference concerns the goal of the analysis. In IRT modeling, the objective is to identify a model that best fits a set of data and adequately describes item response patterns. In Rasch analysis, the model itself is paramount: the intent is to find a set of items that fit a Rasch model. IRT models predominate in the United States, whereas researchers in other parts of the world often prefer Rasch analysis.

The Two-Parameter Logistic Model

Two-parameter logistic (2-PL) models include both the discrimination and location parameters. In such models, item discriminations can be different from each other and thus the ICCs can cross.

A frequently used 2-PL model for polytomous items (e.g., Likert-type items) is called the **graded response model** (**GRM**). Within the GRM, a polytomous item is treated as a series of dichotomies, equal to the number of response options, minus one. In other words, with polytomous items there are *multiple* location parameters, which are sometimes called *category threshold parameters.* The GRM estimates the probability of a patient's response at or above a given category threshold on the latent trait continuum.

For example, on the CES-D depression scale (Radloff, 1977), there are four response options for a set of 20 questions: Rarely or none of the time (<1 day), Some or a little of the time (1-2 days), Occasionally or a moderate amount of the time (3-4 days), and Most or all of the time (5-7 days), with these rank-ordered options scored 0, 1, 2, or 3, respectively. Thus, for a CES-D item, there are three-category threshold parameters. For example, for the item "I felt depressed," the first location parameter (b_1) would correspond to the probability of moving from a response of "none of the time" to "some of the time." The second location parameter (b_2) would correspond to the probability of moving from "some of the time" to "a moderate amount of the time," and b_3 would correspond to the probability of moving from "a moderate amount of the time" to "most of the time."

With polytomous items, **category response curves** (CRCs) represent the probability of a person's response in each category, given his or her level on the latent trait. Figure 4 illustrates the category response curves for the "I felt depressed" item (Item 6), using the GRM model with data from a sample of 1,000 women. The category threshold parameters represent the point along the latent trait continuum at which the respondent has a 0.50 probability of responding above that threshold. For item 6 on the CES-D, a person with a depression trait level of −0.34 has a 0.50 probability of responding "none of the time" versus any response indicating greater frequency. A person with a trait level of +1.18 (b_3) has a 0.50 probability of answering "occasionally" versus "most of the time." This graphic depiction helps in the evaluation of items: the CRCs in Figure 4 suggest a good spread across the latent trait for the "I felt depressed" item and good differentiation across response categories.

Evaluation of Rasch and Item Response Theory Models and Items

In an IRT analysis, the items are evaluated, as is the overall model (i.e., the empirical "fit" of a set of items to form a unidimensional scale). Different criteria and statistics are used in Rasch and IRT models. It is beyond the scope of this brief summary to describe the various statistical methods used to test model fit, but we can make a few remarks.

Rasch analyses are more focused than IRT analyses on the local fit of each item to the model, and so item fit statistics are provided in Rasch software. *Mean square fit* is an index of item fit and is evaluated for two types of fit. The **infit statistic** is sensitive to response patterns across items; it measures unexpected responses to items with a difficulty level near a person's trait level. The **outfit statistic** captures unexpected responses to items that are at the extremes of the trait continuum. For items with a good fit, infit and outfit values are expected to range between 0.7 and 1.4 (Bond & Fox, 2015). Items with high or low infit and outfit statistics are candidates for deletion.

Another evaluative step concerns whether the items, taken together, adequately cover the trait range sufficiently. For example, it is usually more

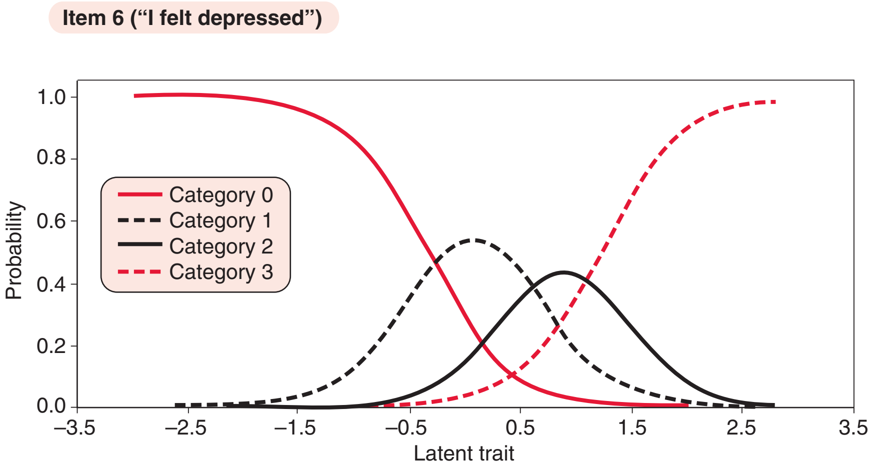

FIGURE 4 Category response curves for CES-D item 6 ("I felt depressed").

desirable to have 10 items that span the trait continuum (or a specific part of the continuum) than to have 10 items with an item difficulty of 0.0. Rasch software (but not IRT software) can be used to create a **person-item map**, which shows the distribution of people along the latent trait (often on the left side of a vertical "ruler" for the trait) and the distribution of items (on the right side of the same trait ruler). With such a map, it is easy to identify where additional items might be needed, or where there are undesirable redundancies in item difficulty.

Many statistics can be used to evaluate the overall fit of IRT models to data from a sample of respondents. Two common fit statistics in IRT analyses are the **root mean square error of approximation** (**RMSEA**) and the **comparative fit index** (**CFI**). The closer the RMSEA value is to zero, the better the fit of the model to the data. Hu and Bentler (1998) suggest a cut-off value of less than or equal to 0.06 as indicative of good model fit. The CFI ranges between 0 and 1; values greater than or equal to 0.95 generally indicate adequate fit (Bentler, 1990).

At the item level, scale developers using IRT models inspect the parameters for each item. When a 2-PL model is used, both location and discrimination parameters play a role in item selection. In terms of location, it is desirable to have items that span the desired trait range—and, for polytomous items, to have response options that have a good range across the trait, as the CES-D item in Figure 4. When examining the ICCs and CRCs, the steeper the slope, the better the item is at discriminating between high and low levels of the latent trait.

In IRT, an important concept concerns how much *information* an item provides. An item response function can be transformed into an **item information function** (IIF). Item information is maximized near the item's location, and the amount of information is a function of item discrimination. Item information is closely associated with measurement error in IRT and is another important tool for evaluating and selecting items for a scale.

Scoring

In IRT and Rasch scaling, the item responses from a set of items with known item response functions are used to estimate a person's position on the latent trait continuum. Specific scoring methods are complex and are not described here. Suffice it to say, though, that IRT scores are not merely the sum of item responses. A distinct advantage, however, is that the trait scores are immediately interpretable. For example, Polit and Yang (2016) illustrated an IRT analysis with seven items from the CES-D in a sample of about 1,000 women. The CTT-based summated scores ranged from 0 to 21. The IRT scores, by contrast, ranged from −1.59 (for the least depressed women who answered all seven items with "rarely") to +2.41 (for the most depressed women who answered all seven items with "most of the time"). These IRT-based scores are immediately interpretable because they can be understood relative to a "ruler" for the depression trait.

ADVANTAGES OF ITEM RESPONSE THEORY

Item response theory offers many advantages for creating multi-item scales, including the fact that the methods have numerous interesting applications. For instance, IRT has proved to be invaluable for creating short-form scales based on existing high-performing scales. In CTT, there is an inevitable tension between the desire to create an internally consistent scale on the one hand and minimizing respondent burden on the other. To improve internal consistency, one needs to only add items, but lengthening the scale makes it more burdensome. Brief scales also have the advantage of being amenable to adoption in busy clinical settings. IRT methods allow researchers to create short forms from existing scales and, at the same time, achieve a comparable (or improved) level of precision. IRT is also frequently used to refine items of an existing scale or to further examine the psychometric properties of a scale developed using CTT methods.

Computerized adaptive testing (CAT) is a popular application of IRT. With CAT, a computer algorithm is used to select a subset of discriminating items from a carefully calibrated *item bank* (a set of items with known item parameters) for the target trait. Items are selected to optimize measurement precision for each respondent. A person usually begins by answering an item of moderate difficulty—i.e., near the middle of the latent trait continuum. The response to that item provides a preliminary estimate of the latent trait level, and the computer then selects another item from the item bank that would improve the estimate. For example, those not endorsing an item would be given an "easier" item, whereas those endorsing it would be given a more "difficult" item. This iterative process continues until a good trait estimate is obtained—usually when a prespecified and low amount of measurement error is achieved. Through this process, it is typically possible to get a good trait estimate with relatively few items.

TIP PROMIS®, an NIH-funded initiative, is a particularly important example of computerized adaptive testing. PROMIS® offers carefully developed and tested CATs for numerous important health outcomes such as fatigue, depression, physical function, and pain intensity. Measures are available for both adult and pediatric populations. PROMIS® measures can easily be administered online, with instantaneous scoring—and with information about normed values for age and gender. PROMIS® item banks have been translated into several languages.

Another important application of IRT is the analysis of **differential item functioning (DIF),** which allows analysts to detect whether items function differently for different subgroups—and therefore may introduce biases. For example, there is an item on the CES-D that asks about crying, and this item has consistently shown DIF for men and women. DIF is an important tool for exploring *equivalence* in cross-cultural validation studies.

A particularly attractive feature of IRT-based scales is that, unlike CTT scales, the measurement of a latent trait is not *test-dependent*. In a CTT scale, adding or omitting an item results in a different scale and different scores, but IRT-based scores are not dependent on a particular set of items. In other words, with IRT methods, a person's position on a latent trait continuum does not depend on the specific items that are administered. Such "item-free" scaling of individual differences is possible because IRT includes both item and person parameters into the same model. This seems intuitively desirable—a person's level of a trait (say, depression) is, at any point in time, a given amount, and it is useful to estimate that value regardless of which items are completed. It is this "item-free" scaling that makes computerized adaptive testing feasible. Also, it means that missing data can be tolerated. A persons' trait level can be estimated with a subset of items that have been calibrated in an IRT analysis, even if some questions are left unanswered.

Another important difference between CTT and IRT concerns measurement error. In CTT scales, a single index of measurement error (the *standard error of measurement* or SEM) is computed for a sample, and the SEM is the same value for everyone in that particular sample (and the SEM is sample-dependent). In IRT scales, measurement error is different at different points along the latent trait continuum, and so the degree of precision is person-specific.

Many of the other advantages of IRT are a bit technical but suffice it to say that IRT scales have desirable features that may lead to them dominating the scale-construction landscape in the not-too-distance future. A major barrier to developing IRT-based scales is that the models themselves are complex, and both statistical and measurement sophistication are required to estimate them. Moreover, the software for IRT analyses is not particularly user-friendly. Another impediment is that in clinical settings, unit-weighted summated scoring is simple to calculate, whereas IRT-based scoring is not. The widespread availability of computer tablets and other handheld technology, however, will likely overcome this problem as IRT-based scales become available for instantaneous online scoring.

Example of a Rasch Analysis

Ma and colleagues (2017) used a Rasch model in their psychometric assessment of an existing scale to measure workplace bullying among nurses. The scale, which had been translated into Chinese, was the 22-item Negative Acts Questionnaire-Revised (NAQ-R). The NAQ-R asks respondents how often they have been subjected to 22 negative acts in the workplace during the last 6 months (e.g., "Spreading of gossip and rumors about you" and "Having your opinions ignored"), on a five-point scale from *never* to *daily*). A major goal was to develop an online computer adaptive testing (CAT) version of the NAQ-R using the Rasch model. The researchers used an experimental design in which their sample of 963 nurses were randomized to the standard NAQ-R or the CAT version. Measurement properties were then compared. The researchers concluded that the scale measured a unidimensional construct and that measurement precision in both forms was comparable. Of note, however, the CAT version achieved the same level of precision with an average of 8.9 items per respondent, compared to 22 items for the standard scale. Their paper, which was published as an open-access article, included a person-item map.

REFERENCES CITED IN THE CHAPTER 16 SUPPLEMENT

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

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

Hu, L., & Bentler, P. (1998). Fit indices in covariance structure analysis: Sensitivity to underparameterized model misspecifications. *Psychological Methods, 4*, 424–453.

*Ma, S. C., Wang, H., & Chien, T. (2017). A new technique to measure online bullying: Online computerized adaptive testing. *Annals of General Psychiatry, 16*, 26.

Polit, D. F., & Yang, F. M. (2016). *Measurement and the measurement of change: A primer for health professionals*. Philadelphia: Lippincott.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.

***A link to this open-access journal article is provided in the Toolkit for this chapter in the accompanying* Resource Manual.**