AJSLP

Tutorial

An Introduction to Item Response Theory and Rasch Models for Speech-Language Pathologists

Carolyn Baylor,[a] William Hula,[b] Neila J. Donovan,[c] Patrick J. Doyle,[b] Diane Kendall,[a,d] and Kathryn Yorkston[a]

Purpose: To present a primarily conceptual introduction to item response theory (IRT) and Rasch models for speech-language pathologists (SLPs).
Method: This tutorial introduces SLPs to basic concepts and terminology related to IRT as well as the most common IRT models. The article then continues with an overview of how instruments are developed using IRT and some basic principles of adaptive testing.
Conclusion: IRT is a set of statistical methods that are increasingly used for developing instruments in speech-language pathology. While IRT is not new, its application in speech-language pathology to date has been relatively limited in scope. Several new IRT-based instruments are currently emerging. IRT differs from traditional methods for test development, typically referred to as classical test theory (CTT), in several theoretical and practical ways. Administration, scoring, and interpretation of IRT instruments are different from methods used for most traditional CTT instruments. SLPs will need to understand the basic concepts of IRT instruments to use these tools in their clinical and research work. This article provides an introduction to IRT concepts drawing on examples from speech-language pathology.

Key Words: item response theory, outcomes measurement, Rasch model

Speech-language pathologists (SLPs) rely on a wide range of assessment instruments in their work—for example, articulation tests, language batteries, and questionnaires asking clients to rate their communication experiences. These instruments are important for quantifying the performance or perspective of clients, tracking changes in clients over time, and reporting this information succinctly to key stakeholders including the clients, referral sources, and payers. Applied researchers conducting treatment outcomes studies depend on these instruments to document treatment efficacy. SLPs are aware of many questions they should ask about instruments before choosing one to use with a client: Does the instrument address the trait or skill that I need to assess with this client? Does the instrument address skill levels appropriate to my client? Was the instrument developed using population samples reasonably similar to my client? Is the instrument appropriate in terms of age, language, and cultural concerns? Have I done enough background study with the instrument so that I can administer it correctly? One issue that SLPs may not be familiar with is the different theoretical approaches to measurement and how these different approaches affect instrument development, administration, scoring, and interpretation.

Many of the instruments SLPs currently use were developed using the psychometric methods of classical test theory (CTT) and are based on a definition of measurement as "the assignment of numerals to objects or events according to a rule" (Stevens, 1946, p. 677). The four scales that are associated with this definition are referred to as nominal, ordinal, interval, and ratio scales, descriptions of which are in most introductory statistics textbooks (e.g., Gravatter & Wallnau, 2000). Others have proposed that measurement requires more stringent criteria than the Stevens (1946) definition—in particular, that measurement requires a linear, additive scale with equal units (like inches on a ruler) that allow the numbers to be manipulated mathematically (i.e., addition, subtraction, or averaging; Bond & Fox, 2001; Michell, 1997, 2003; Wright, 1999). Item response theory (IRT; Lord & Novick, 1968) and the related set of measurement models based on the work of Georg Rasch (Rasch, 1960,

[a]University of Washington, Seattle
[b]VA Pittsburgh Healthcare System Geriatric Research Education and Clinical Center, and University of Pittsburgh, PA
[c]Louisiana State University, Baton Rouge
[d]VA Puget Sound Medical Center, Seattle, WA
Correspondence to Carolyn Baylor: cbaylor@u.washington.edu

Editor and Associate Editor: Laura Justice
Received September 10, 2010
Revision received February 10, 2011
Accepted April 25, 2011
DOI: 10.1044/1058-0360(2011/10-0079)

1980) offer a different set of statistical methods that attempt to address these more stringent definitions of measurement.

There are many differences between CTT and IRT in how instruments are constructed, administered, and interpreted. Even if SLPs are not familiar with the terminology of CTT, they are probably aware of many CTT guidelines. For example, all items in an instrument (or subscale) need to be administered to obtain a valid and reliable score. Longer instruments are generally more reliable than shorter instruments. Scores across clients or test situations cannot be compared unless the same items were administered or the instruments are equated psychometrically. Test scores are interpreted relative to a normative or reference sample, meaning that a client's ability is understood by comparing that client to the normative sample (e.g., percentile rankings). Many features of IRT differ from those of CTT (Embretson & Reise, 2000). For example, options for adaptive testing allow for only a subset of items from a larger item bank to be administered. This tailors assessment for each client and lowers response burden. These shorter item subsets can be as reliable as longer CTT instruments resulting in greater measurement efficiency (Cook, O'Malley, & Roddey, 2005). Adaptive testing leads to different item subsets being administered to different clients, or to the same client at different testing points, yet these scores can be directly compared facilitating comparison across time and clients. Finally, test scores are usually interpreted on the logit scale (discussed below), with the scale representing the difficulty of item content. Comparing a client's performance or response directly to the difficulty level of the content or skill is a more direct method of understanding the client's ability than knowing only where the client ranks in a normative sample.

IRT and Rasch models are not new. They date back to the 1960s and have been in use since then particularly in educational and aptitude testing. Over the past 10–15 years, IRT has been increasingly applied in psychology and the health sciences, and recently more IRT work has gained momentum in speech-language pathology. Early examples of the application of IRT to instruments relevant for speech-language pathology include the Token Test (Willmes, 1981), the Test of Adolescent/Adult Word Finding (German, 1990), the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997), the Aachen Aphasia Test (Willmes, 2003), and the Rehabilitation Institute of Chicago Evaluation of Right Hemisphere Dysfunction—Revised (Cherney, Halper, Heinemann, & Semik, 1996). However, this work has either tended to be technical in orientation with limited impact on clinical practice or has not capitalized on some of the particular advantages of IRT. Several research groups are now working on a variety of new IRT-based instruments intended for clinical use (Baylor, Yorkston, Eadie, Miller, & Amtmann, 2009; Donovan, Kendall, Young, & Rosenbek, 2008; Donovan, Velozo, & Rosenbek, 2007; Hula, Austermann Hula, & Doyle, 2009; Hula, Doyle, & Austermann Hula, 2010; Justice, Bowles, & Skibbe, 2006; Kendall et al., 2010; Milman et al., 2008), and more are likely to emerge in coming years.

Because IRT and Rasch instruments will become more prevalent in the SLP's toolkit, SLPs need to be familiar with the basic terminology and concepts of IRT. The purpose of this article is to provide an introduction to IRT and Rasch models for clinical SLPs and applied researchers who may be interested in using IRT instruments but are unfamiliar with these measurement models. This tutorial provides an overview of the concepts and terminology, a brief description of how IRT instruments are constructed, and a description of adaptive administration, which is a particular advantage of IRT instruments.

Before proceeding, an explanation of how terminology is used in this article is warranted. In the prior paragraphs, IRT and Rasch have been mentioned together. IRT models were developed in the 1950s and 1960s by Fred Lord and Alan Birnbaum (Birnbaum, 1968; Bock, 1997; Lord & Novick, 1968). For a review, see Bock (1997). Around the same time, the Danish mathematician Georg Rasch independently developed a model that is mathematically identical to the one-parameter IRT model (Rasch, 1960). However, Rasch's work had a different theoretical orientation than that of Lord and colleagues (Wright, 1997). Whereas the latter were more oriented to the practical needs of large-scale educational testing, Rasch was more focused on the principles underlying scientific measurement and on developing psychometric models that would meet the most rigorous standards of scientific measurement.[1] There are philosophical and methodological differences between IRT and Rasch models. However, most SLPs will not need to dwell on these differences for their clinical purposes. Because of this, the distinctions between IRT and Rasch will be mentioned only briefly, and the focus will be on the concepts and applications that are common to both IRT and Rasch. For ease of reading, the term *IRT* is used in this article to encompass both IRT and Rasch-based instruments where they have commonalities.

Key IRT Concepts

Much of the information presented in this tutorial can be found in introductory texts on IRT or Rasch models as well as in peer-reviewed publications (Bond & Fox, 2001; Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991; Hays, Morales, & Reise, 2000; Jette & Haley, 2005; Reeve et al., 2007; Smith & Smith, 2004). For ease of reading, this list will not be referenced repeatedly throughout this text for basic concepts, but additional references will be used for specific points.

Latent Traits

Our discussion of IRT begins with the definition of a *latent trait*. A latent trait is a characteristic or ability of an individual that is not directly observable but instead must be inferred based on some aspect of a person's performance or presentation. Many areas of communication are latent traits such as language and cognitive processes. For example, we cannot directly observe a client's auditory comprehension.

[1] The work of Lord and colleagues may be considered data-driven, in that they developed increasingly complex models to better describe empirical data, while the work of Rasch and his students can be characterized as model-driven, in that they have sought to develop more restrictive models that would satisfy the requirements of scientific measurement.

We can only infer or estimate a client's accuracy of auditory comprehension based on that client's performance on comprehension tasks such as following commands or answering yes/no questions. Examples of other latent traits are related to psychosocial issues such as coping or adjustment. We cannot directly observe a client's feelings about his or her experiences with a communication disorder, but we can make inferences about that latent trait based on the client's answers to questions on that topic.

One of the challenges in estimating latent traits is to understand how well the instruments we are using represent the intended latent traits.[2] The relationships between latent traits and instruments can be explained by measurement models. In CTT, the focus is on a client's total score on an instrument that is assumed to represent the person's actual trait level (i.e., the true score) and any measurement error (Crocker & Algina, 1986). In contrast, IRT models focus on responses to individual items instead of the total score. IRT models describe the probability of a client's response to each item as a function of the person's latent trait level and the properties of the items. For these reasons, IRT is often referred to as both model-based and item-based measurement. It is model-based because there is a mathematical equation, or model, that explains the relationship between the latent trait and a client's responses to the items, and it is item-based because the model explains the relationship between the instrument and the latent trait on an item-by-item basis according to the specific characteristics or parameters of each item. There are many different IRT models, and some of the more common ones that SLPs will encounter are introduced below. Before we introduce the specific models, however, the following sections provide additional information about some terminology and concepts that are incorporated into IRT models. The next section discusses a unit of measurement commonly used in IRT, the logit scale, and this is followed by an introduction to item parameters.

Scale Properties (Logits)

In most CTT instruments, SLPs report a client's latent trait level in terms of the number of items a client answers correctly on a test or where the client ranks relative to normative data using a percentile score. While these scoring methods can be used in IRT as well, they have some drawbacks. When reporting the number of items correctly answered, the interpretation of that score depends on the items that were administered. For example, the SLP might assume that scoring 9 out of 10 correct represents a high level of proficiency. However, knowing that a client scored 9 out of 10 correct does not tell the SLP much about the client's underlying skill level if the skill level represented by the items composing the test is unknown. Scoring 9 of 10 correct on a naming task on which all of the items are relatively easy to name (i.e., common, familiar objects with high-frequency names) represents a different skill level than 9 of 10 correct on a naming task on which all of the items are relatively difficult to name (i.e., abstract concepts or uncommon objects with low-frequency names). It is possible that a person scoring 6 of 10 on the harder test is more able than a person scoring 9 of 10 on the easier one. Interpretation of a score reported in number of items correct, therefore, is dependent on knowing the items that were administered.

A similar problem of dependence exists when interpreting scores with reference to a group such as normative data. Scores such as percentile rankings are dependent on knowing who was in the reference group. For example, the SLP might assume that a ranking at the 95th percentile represents a relatively high level of proficiency. However, knowing that ranking really does not tell the SLP much about the client's underlying skill level if the skill level represented in the reference sample is unknown. If the reference group was in general a very highly skilled set of individuals, ranking 95% may reflect a higher level of proficiency than achieving a ranking of 95% compared to a reference group of relatively low-ability individuals. The ranking can only be interpreted in the context of the reference group.

IRT addresses the two problems of dependence described above by typically interpreting scores through the use of the logit scale.[3] With the logit scale, a person's ability or trait level can be understood without direct reference to the items that were administered or any normative or comparison group. It may help to think of the logit scale as a type of "ruler" for latent traits, with the units on the ruler being logits instead of inches. Logit is short for log odds unit. The units on the logit scale represent increasing odds of answering an item correctly (for correct/incorrect items) as the logit value increases. An increase of one unit (i.e., from 1.0 to 2.0 logits) corresponds to an increase in the odds of a correct response by a factor equal to the base of the natural logarithm, approximately 2.718. Said more simply, higher logit values are associated with a higher probability of a correct answer to an item and therefore represent higher levels of the latent trait. The logit range most commonly used in IRT instruments is from –3.0 logits to 3.0 logits, with 0 logits representing the mean of the ability levels represented in the test sample (or the mean item difficulty level). However, logit values beyond –3.0/3.0 can be used, and SLPs may see these values in instrument administration. The logit score is either calculated automatically in computer-aided administrations or by using scoring guides for manual administration.

One of the key advantages often attributed to the logit scale is that of approximating an equal-interval scale (i.e., just as on a ruler where an inch always represents a specific unit of length, a logit always represents a specific change in the odds of answering an item correctly). Interval scaling is important because it supports the kinds of inferences we often wish to draw from clients' test scores, such as

[2] IRT typically refers to "estimating" the latent trait, which serves as a reminder that we are observing the trait indirectly in our instruments and cannot measure it directly.

[3] The logit scale is presented here because it is probably the most common score-reporting format that SLPs will see. However, other metrics, such as probits, have also been used in IRT. IRT can also utilize norm referencing as well. When learning to administer a new instrument, SLPs need to check to be sure they understand the score-reporting format the scale developers have chosen.

determining the extent to which a client's ability has changed or whether a client has responded more strongly to one intervention as opposed to another. An equal-interval scale can be used to conduct mathematical calculations such as differences between scores across clients or across time, or averages across a group (i.e., 4 logits – 2 logits = 2 logits, just as 4 in. – 2 in. = 2 in.). These mathematical operations are valid only with equal-interval scales. This point is highly relevant because of the frequent use by SLPs of ordinal categorical scales that do not have equal intervals. For example, suppose a researcher asks clients to rate the difficulty level of communication tasks using the following categories: 1 = *not difficult,* 2 = *somewhat difficult,* 3 = *very difficult,* 4 = *so difficult I cannot do*. A common practice is to calculate the average of the ratings across items or across individuals to obtain one data point to represent that individual or a group. This is not a valid mathematical operation, however, because in this case the 1, 2, 3, and 4 values do not represent numbers on an interval scale. The numbers are only labels for categories. They do reflect that there is an order to the categories that represent increasing levels of difficulty. But there is no evidence, for example, that the difference in difficulty between Categories 1 and 2 is the same as between Categories 3 and 4. On an ordinal scale, 3 – 2 does not equal 1. The averaging of ordinal scores has the potential to lead to mistaken inferences, for example, the conclusion that two groups differ on some measure when in fact they do not (Maxwell & Delaney, 1985). Through IRT, the categorical responses can be converted to the logit scale, allowing valid mathematical operations and giving a truer depiction of distances between persons and across time.

Although the prior paragraph describes an advantage of using the logit scale as an equal-interval scale, two notes of caution are warranted on this topic. First, using the logit scale does not necessarily mean that the underlying latent trait inherently exists in equal-interval units. The logit scale does provide a method, however, of transforming the ordinal nature of traits (i.e., more to less) into an interval scale for the mathematical reasons described above. Second, the strength of claims to interval scale measurement is one area of difference between Rasch and IRT models that should be noted. While both IRT and Rasch models use the logit scale, not all IRT models achieve the ideal of interval scale measurement of the latent trait to the same degree. Claims to interval scale measurement are strongest for the Rasch and one-parameter IRT models (discussed below), while two- and three-parameter IRT models can only claim "limited" interval scale properties due to the complexities of the models (Hambleton et al., 1991).

The logit scale is used to report the latent trait level of people. The logit scale is also used when describing item parameters, particularly item difficulty. The following section provides more details about item parameters that are important for understanding IRT as item-based measurement.

Item Parameters

IRT models use information about individual items, the item parameters, to explain the relationship between how clients answer each item and the clients' underlying trait levels. The two item parameters that are most important for SLPs to be aware of are *item difficulty* and *item discrimination.*[4]

Each item has an item difficulty value that represents the location of the item on the trait or ability range. For dichotomous items (items that have only two response choices, i.e., correct/incorrect or agree/disagree), item difficulty is defined as the point in the trait range where clients with that same ability level have a 50% probability of a correct response. Clients with latent trait levels higher than an item's difficulty value will have a higher likelihood of answering the item correctly and vice versa for clients with lower trait levels. Items with higher logit values are said to be more difficult because clients need to possess a higher level of the latent trait to correctly respond to or endorse the item. Item difficulty is that point on the trait range where the item is most effective in discriminating between trait levels (i.e., clients with different trait levels are most likely to answer the item differently).

Item discrimination refers to how sharply an item differentiates among people who have different trait levels. When items have high discrimination values, clients with different trait levels are likely to respond differently to the item. In contrast, clients with different trait levels may respond similarly on items with low discrimination. Items with higher discrimination are generally preferred and are given more weight in estimating a person's ability levels because they are assumed to represent a stronger relationship between the item and the underlying trait. To understand how the item parameters of item difficulty and item discrimination are used to estimate an individual's latent trait, the item parameters need to be "plugged in" to an IRT model that will explain the relationships among these variables. IRT models are explained in the following section.

IRT Models

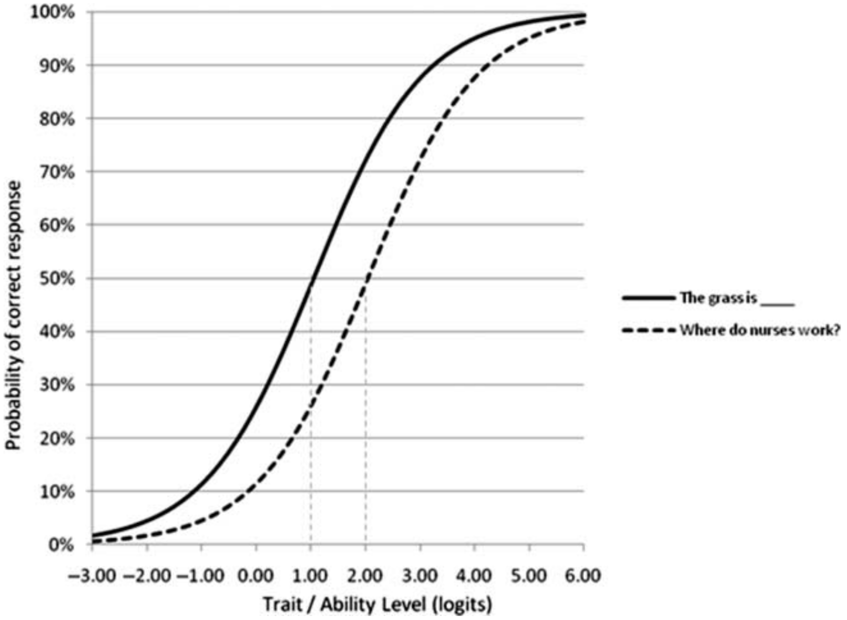
Now that some of the key terminology and elements used in IRT models have been defined, it is time to put the components together into the IRT models.

The 1-PL Model

The simplest model is the one-parameter logistic (1-PL) IRT model, which, as stated above, is mathematically identical to the Rasch model. This is a model for dichotomous items. It is called a one-parameter model because it uses only one item parameter—item difficulty—to explain the relationships among the item, the latent trait, and the client's response to the item. Item discrimination is assumed to be the same across all items. Figure 1 presents the 1-PL model both graphically and mathematically by showing the relationship between the three quantities: person ability (referred to as theta, θ), item difficulty (represented here by beta, β), and the probability of a correct response. The scale on the x-axis, reported in logits for all examples in this article,

[4] The parameters of item difficulty and item discrimination exist in CTT as well but are defined and/or calculated differently. Item difficulty is determined by the proportion of respondents who answer an item correctly. Item discrimination can be calculated through a variety of correlation coefficients that compare performance on an item to the total test score (Crocker & Algina, 1986).

FIGURE 1. Graphical and mathematical depiction of the one-parameter logistic (1-PL) dichotomous item response theory (IRT) model. The mathematical equation for the 1-PL model is: $P_{ni} = [e^{(\theta n - \beta i)}]/[1 + e^{(\theta n - \beta i)}]$. The 1-PL model says that the probability of a correct response by person *n* to item *i* (P_{ni}) is a function of the difference between person ability (represented by θ) and item difficulty (represented by β). The symbol *e* is the base of the natural logarithm. As the difference between person ability and item difficulty (i.e., θ – β) increases, the probability of a correct response increases. As this difference decreases, the probability of responding correctly decreases. This figure presents the item characteristic curves (ICCs) for two items and demonstrates the difference in item difficulty between two items analyzed with a 1-PL dichotomous IRT model.



represents both client ability and item difficulty for the construct (or trait) estimated by the instrument, in this example, word finding. Lower logit values represent lower word-finding ability and easier items (items that clients are more likely to answer correctly). The *y*-axis represents the probability of a correct response on the item. The two curves in Figure 1 are called item characteristic curves (ICCs). The ICC for each item demonstrates that as ability increases (*x*-axis), the probability of responding to the item correctly increases (*y*-axis). In Figure 1, the two ICCs represent two separate items from the Western Aphasia Battery (WAB; Kertesz, 1982) sentence completion task. The curve on the left is the ICC for the item "The grass is ____(green)"; the curve on the right is for the item "Where do nurses work?" These data are from a reanalysis of the WAB using a Rasch model (Hula, Donovan, Kendall, & Gonzalez-Rothi, 2010).

In Figure 1, the item difficulty for the item "The grass is ____ (green)" is 1.0 logit. Note that 1.0 on the *x*-axis corresponds to where the ICC crosses 50% on the *y*-axis. This means that a person with an ability level of 1.0 logit has a 50% probability of answering the item correctly.[5] This item is most sensitive for estimating trait levels at or very near 1.0 logit. This sensitivity is demonstrated by the nonlinear aspect of the ICC. Note how the slope of the ICC is steeper in the region around the item difficulty value and shallower at the ends of the curve. Greater differences in response probability are seen among clients who have trait values closer to the item difficulty value, and smaller changes in response probability across trait levels are seen more distant from the item difficulty value.

To explore the concept of item difficulty further, compare the two items in Figure 1. The first item ("The grass is ____") has an item difficulty of 1.0 logit, and the second item ("Where do nurses work?") has an item difficulty of 2.0 logits. Changes in item difficulty affect the probability of a correct response. Because the distance between these two items is exactly 1 logit, clients are 2.718 times more likely to respond correctly to the easier (lower logit) item. One advantage of IRT scaling is that if we have an estimate for a client's ability level, we can estimate the probability of that individual answering any given item correctly. Using Figure 1, a client with an ability level of 1.0 logit will respond correctly to items like "The grass is ____" 50% of the time and to items like "Where do nurses work?" approximately 27% of the time. Having this kind of information helps SLPs avoid administering items that are too difficult or too easy and to instead give items that will yield the most information about a client.

[5] This commonly used interpretation is actually an abbreviated version of the following more technically correct interpretation: 50% of individuals with an ability level of 1.0 will answer this item correctly. Also, given a set of items with difficulty equal to 1.0, a single individual with this ability level will correctly answer 50% of them.

The ICC can also be interpreted from the other direction using knowledge about the model and item difficulty to estimate a client's trait level. This is probably the more common situation. Suppose we are trying to estimate the ability level of a client with aphasia who answered "The grass is ____" correctly. Based on this single item, it is reasonable to assume that this client's ability estimate would be 1.0 logit, because only clients with $\theta \geq 1.0$ have a greater than 50% chance of answering the item correctly. This estimate of the person's ability level is not very precise, but we can improve its precision by considering the response to the second item.[6] Suppose that the client answered "Where do nurses work?" (item difficulty of 2.0 logits) incorrectly. A likely ability level that would have led to this combination of responses is 1.5 logits (half way between the difficulty levels of the two items). Because we have information from only two items, this estimate is still very imprecise. The precision of the estimate can be improved if we administer more items, especially items with difficulty values close to the client's ability estimate, and then update the estimate after each response.

The 2-PL Model

As discussed above, the 1-PL models assume that all items have the same discrimination, but in reality, item discrimination varies across items. To account for this, the two-parameter logistic (2-PL) model (Lord & Novick, 1968) includes both item difficulty and item discrimination, often designated as *a* or with the Greek letter alpha, α. Figure 2 demonstrates the concept of item discrimination by presenting the ICCs generated by a 2-PL model for the same two items from the WAB that were presented in Figure 1. Notice that the ICCs for the two items have different slopes. The slope of the ICC for "The grass is ____" is more shallow, and this means that it discriminates or differentiates less sharply between clients of high and low ability (clients with different ability levels are likely to respond more similarly to this item). The slope for "Where do nurses work?" is steeper, meaning that it discriminates more sharply, and a smaller change in ability will lead to a larger change in the odds of a correct response.

The 2-PL model has several advantages, including providing a closer statistical fit to the empirical data (i.e., the data from the population sample used for item calibration during instrument development) as well as more reliable ability estimates. For these reasons, many users of IRT prefer two-parameter models. However, these advantages come at a cost. Some of these costs are relevant only to instrument developers, and if the developers address these challenges adequately, SLPs can use the instruments with confidence. One of the main challenges is that during the instrument development process, larger sample sizes are needed for item calibration, making instrument development more complex and costly. The necessary sample size may vary depending on a variety of issues, including the model used, missing data, and characteristics of the variables, all of which are beyond the scope of this article (Muthén & Muthén, 2002). Recommendations have included sample sizes of 100 respondents for analyses using the 1-PL model (Linacre, 1994) and up to as many as 250 to 500 respondents for a 2-PL model analysis (Reise & Yu, 1990; Stone, 1992). The take-home message for SLPs is that when considering an IRT instrument, they may want to look at the sample size used during instrument development with the awareness that larger sample sizes are generally more favorable.

Other challenges of the 2-PL model are more visible to users such as SLPs. Inclusion of the discrimination parameter complicates the interpretation of the ICCs. Examine again Figures 1 and 2. In Figure 1, the ICCs are parallel, meaning that the difference in difficulty level between the two items is constant for all respondents at all trait levels. By contrast, in Figure 2, ICCs in the 2-PL model are not necessarily parallel. A consequence of this is that the ordering of items by difficulty is not consistent for all respondents. In Figure 2, the item "The grass is ___" is easier than "Where do nurses work?" for clients with ability estimates less than 2.6 (clients with ability levels less than 2.6 have a higher probability of answering the "grass" item correctly than of answering the "nurse" item correctly). The reverse is true for those with ability estimates higher than 2.6. For SLPs, 2-PL models will not seem much more difficult than 1-PL models in practical use. Instrument developers will provide computer programs or manual scoring algorithms that will guide the clinician through the scoring process.

Polytomous Models

The sections above describe 1-PL and 2-PL models for dichotomous items. Many SLPs, however, use instruments with items that have Likert scales (categorical rating scales) for their responses. These items are called polytomous items. A common example provided above involves asking clients to rate their communication experiences on a difficulty scale of 1 to 4. Different IRT models, usually extensions of the models described above, are used for polytomous items. The focus in polytomous models is on understanding the relationship between a client's trait level and the probability of that client choosing a particular response category. There are a number of different polytomous models, the details of which are beyond the scope of this article. The following paragraphs will focus on explaining two graphs that SLPs may need to interpret when using IRT instruments that have polytomous items. Both of these graphs explain the relationship between a client's latent trait level and the likelihood of choosing response categories for individual items.

One of the questions that may interest SLPs is the trait level at which a client's response moves from one response category to the next. For example, at what point might a client's response change from 2 (*somewhat difficult*) to 1 (*not difficult*) in the example above? The trait level at which a client is likely to move from one response category to the next is the category threshold. In polytomous models, the category thresholds are identified in a graph of operating characteristic curves (OCCs). Figure 3 shows the OCCs generated by a 1-PL polytomous model (the Rasch-based Partial Credit Model) using an item from the Communicative Participation Item Bank (CPIB; Baylor et al., 2009). This

[6] Measurement precision is inversely related to the standard error of measurement (Cook, Roddey, O'Malley, & Gartsman, 2005). Precision is related to the concept of reliability in that if measured repeatedly, very similar scores would result from each measurement.

FIGURE 2. This figure presents the same item as in Figure 1; however, the items have been analyzed with a two-parameter logistic (2-PL) IRT model. The mathematical equation for the 2-PL model is: $P_{ni} = \{e^{[\alpha i(\theta n - \beta i)]}\}/\{1 + e^{[\alpha i(\theta n - \beta i)]}\}$.The 2-PL model includes a second parameter to describe the items, discrimination, designated by α. The discrimination parameter functions as a multiplier for the difference between person ability and item difficulty [$(\theta - \beta)$ becomes $\alpha(\theta - \beta)$], permitting items to vary in the strength of their association with the underlying trait. This is reflected in the different slopes of the ICCs for different items with the steeper slope representing higher discrimination.

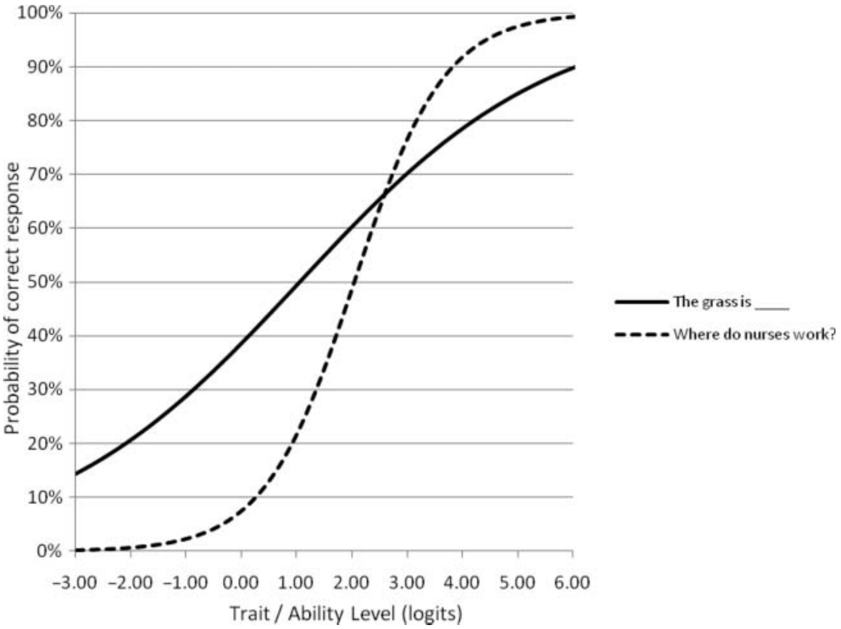
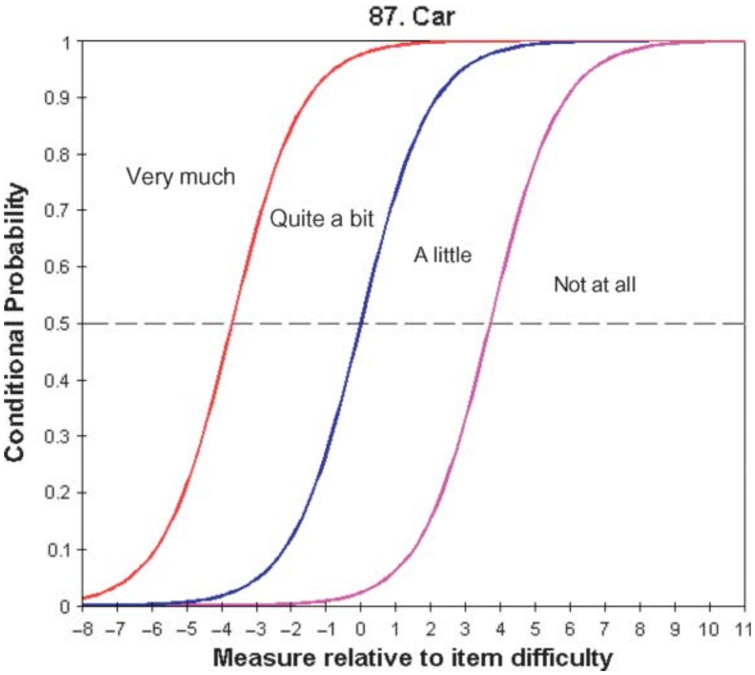


FIGURE 3. Operating characteristic curves for a polytomous item analyzed with a 1-PL IRT model. These curves show the threshold trait levels for responses crossing from one category to another based on theta.



self-report instrument asks adults with communication disorders to rate the interference they experience participating in a variety of everyday speaking situations on a categorical scale of *not at all, a little, quite a bit,* and *very much*. The item in Figure 3 asks clients to rate the interference they experience having a conversation while riding in a car. Each of the three curves in Figure 3 represents a category threshold or boundary between response categories for this item (there are three curves to separate the four response categories). The threshold for a category is identified as the logit value at which the probability of responding in the higher category reaches 0.50 (essentially a 50/50 chance of responding in the categories on either side of the threshold). In Figure 3, the OCC that separates the "a little" category from the "not at all" category crosses the 0.5 response probability on the *y*-axis at a value of approximately 3.5 logits on the *x*-axis. A client with a theta value of 3.5 logits has a 50% probability of choosing either the "a little" or "not at all" categories. Clients with theta values less than 3.5 are more likely to choose the "a little" (or lower) categories, while clients with theta values greater than 3.5 logits are likely to choose the "not at all" category.

The response categories in polytomous items can also be studied with a different type of graph—the category response curves. The OCCs described in the prior paragraph show the boundaries or thresholds for crossing from one category to the next based on latent trait levels. Category response curves show the probability of a client responding in each particular response category (*y*-axis) depending on his or her trait level (*x*-axis). Figure 4 shows the category response curves for the same item as in Figure 3. In this figure, a client with a trait level of –2.0 logits has the highest probability of endorsing the response category "quite a bit," and a client with a trait level of 2.0 logits is most likely to choose the response category of "a little." During instrument development, the category response curves are often examined to determine whether each category presents a unique response choice. Excessive overlap of categories or large gaps between categories on the category response curves may indicate that the number or content of response categories needs to be changed.

The previous sections have provided an introduction to some of the most common IRT models. As mentioned above, a much wider array of models exist. Table 1 presents a summary of IRT models that SLPs are most likely to encounter in the current literature for unidimensional item sets (instruments that measure a single construct or latent trait; Embretson & Reise, 2000; Hays et al., 2000). When administering IRT instruments, SLPs are not responsible for choosing the model to use. This decision is made by the instrument developers and is based on the types of items and response formats (i.e., yes/no vs. rating scales), as well as the complexity of the model needed to adequately estimate latent traits. SLPs should be aware, however, that the choice of models may affect ability estimation.

Developing IRT-Based Instruments

The following sections briefly trace the process of developing IRT-based instruments. This is not a comprehensive

FIGURE 4. Category response curves for a polytomous item. These are the category response curves for the same item depicted in Figure 3. The category response curves show the probability of a client choosing each category dependent upon trait level.

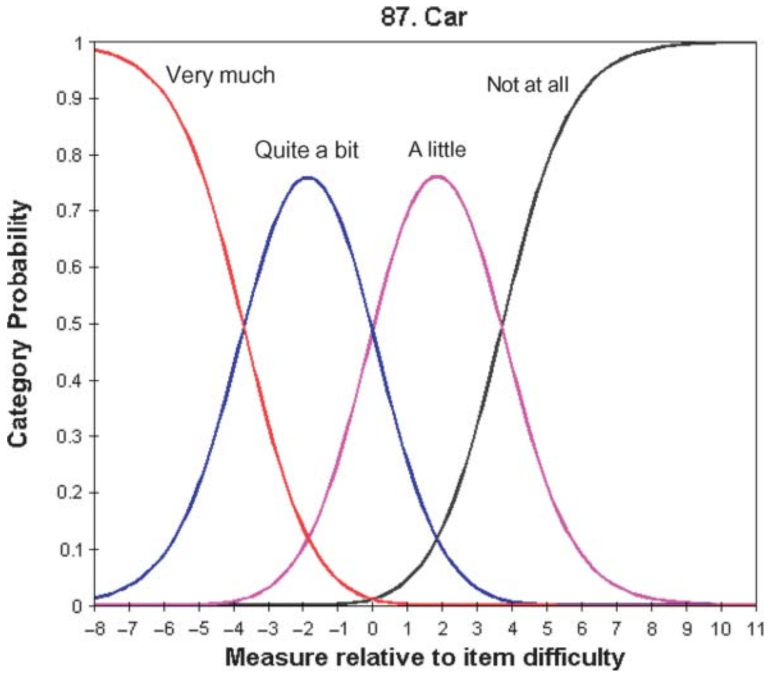


TABLE 1. Summary of common item response theory and Rasch models.

Model name (reference)	Parameters[a]	Response format[b]	Additional information
Rasch (Rasch, 1960)/1-PL	Item difficulty	Dichotomous	The Rasch and 1-PL models are mathematically equivalent but stem from different theoretical approaches. This model is described in the text.
Partial Credit Model (Masters, 1982; Masters & Wright, 1996)	Item difficulty	Polytomous	Originally designed for items that have multiple steps for which the respondent can get partial credit. Also appropriate for categorical rating scale items. Key feature is that Partial Credit Model makes no assumptions about the ordering of the response categories, meaning that the order and "distance" between response categories can vary across items. This allows developers to explore how the response format works for the items and to change the response formats if needed (i.e., add more categories or remove categories).
Rating Scale Model (Andrich, 1978a, 1978b)	Item difficulty	Polytomous	Similar to the Partial Credit Model except that this model assumes the response categories are in the same order with the same "distance" between categories for all items. This does not allow exploration of variations in and function of response categories. Cannot be used if different items have different response formats.
2-PL	Item difficulty and discrimination	Dichotomous	This model is described in the text.
Graded Response Model (Samejima, 1969, 1996)	Item difficulty and discrimination	Polytomous	Items can have different numbers of response categories or different response formats across the items.
3-PL	Item difficulty, discrimination, and guessing	Dichotomous	This model takes into account the possibility of clients answering items correctly due to guessing, particularly that very low ability clients can still score some items correctly due to guessing.

Note. 1-PL = one-parameter logistic; 2-PL = two-parameter logistic; 3-PL = three-parameter logistic.

[a] The parameters correspond to the 1-PL (item difficulty), 2-PL (item difficulty and discrimination), and 3-PL (item difficulty, discrimination, and guessing parameters) models.

[b] Dichotomous response formats have only two possible responses such as correct/incorrect, true/false, or yes/no. Polytomous response formats are multiple response formats such as Likert rating scales (i.e., *strongly agree, slightly agree, slightly disagree,* or *strongly disagree*).

guide to instrument development, and SLPs interested in taking on this task will need to consult more advanced references and the expert advice of a psychometrician experienced in IRT. However, while many SLPs may not be developing item banks, having insight into this process may be helpful when using IRT instruments. SLPs should be familiar with the basic concepts behind item bank construction to be informed "consumers" of the instruments. This section may help SLPs interpret information in test manuals, including instructions for administration and scoring.

In some instances, existing instruments developed using CTT methods are reanalyzed using IRT. IRT analyses of these instruments can shed additional light on the measurement properties of the instruments. However, it should be noted that reanalyzing CTT instruments using IRT does not provide all of the advantages of IRT. For example, item banks generated from their beginning through IRT typically contain a large number of items that are specifically chosen from an even larger set of candidate items to meet several criteria that will be discussed below. Test developers can choose the most desirable items from among the candidate set to create the optimal item bank for specific measurement purposes. When reanalyzing existing CTT instruments, the number of items has already been reduced and fixed during the original development of the instrument. If there are problems with the items in an IRT analysis such as multidimensionality, poor fit to the IRT model, or uneven distribution across the measurement range (all issues to be discussed further below), the test developers have limited options for improving an existing item set because there is no larger candidate item set from which to draw items that better fit the IRT model. Some researchers also use IRT to compile new item sets by combining existing CTT instruments measuring the same construct. This provides a larger pool of items from which to draw for the instrument but still has some of the limitations just described.

Meeting Key IRT Assumptions

When creating an IRT item bank, instrument developers begin by identifying the construct that they are interested in assessing (i.e., reading comprehension or verbal naming). The developers collect a large pool of candidate items that they think reflect that construct. They may use items that already exist in other instruments (with appropriate references to those instruments), or they may write their own items. When developers feel that they have a comprehensive set of candidate items, they will administer the items to a large sample of individuals in the target population. They will conduct a variety of analyses of that data to understand how the items function statistically. IRT analyses are conducted using statistical software designed for that purpose. Various software packages are available, examples of which include Winsteps (Linacre, 1991), which focuses on Rasch-based analyses; Multilog (Thissen, Chen, & Bock, 2003); Mplus (Muthén & Muthén, 1998); and Bilog (Zimowski, Muraki, Mislevy, & Bock, 2003). The information from these analyses guides the developers in retaining, removing, or revising items in the set of candidate items. The goal is to arrive at a final set of items, the item bank, which will serve as the reservoir of items for that instrument. Some key pieces of information that test developers use to make decisions about which items to include in the final item bank are discussed in the following sections. The first section presents information about two key assumptions in IRT that items must meet to be included in an item bank for most IRT models: unidimensionality and local independence.

Unidimensionality

One key assumption of most of the commonly used IRT models is unidimensionality.[7] Unidimensionality means that all of the items in an instrument represent the same underlying construct or latent trait. If items that represent different constructs are included in the same item bank, this will introduce different variables that will confound estimation of the latent trait of interest. For example, if developers are creating a test of reading comprehension, they may want to include items that test a variety of different types of reading tasks. Perhaps one of those tasks is a mathematical word problem that clients need to solve. It is possible that there are two variables tested by this item—reading comprehension and mathematical calculations. Including an item such as this in the item bank may introduce a second construct into the item set (mathematical skills), therefore violating the assumption of unidimensionality for reading comprehension. Statistical analyses are conducted in early stages of instrument development to provide quantitative evidence for unidimensionality and to guide the removal of items that represent different constructs. These statistical methods are exploratory and confirmatory factor analyses (Floyd & Widaman, 1995; Pett, Lackey, & Sullivan, 2003), which are used with both IRT and CTT. Factor analyses are used to explore the relationships among the items to find those that are most closely related to one another. Items that are unidimensional (all measure the same construct) will load onto a single factor that accounts for the observed item relationships. Various criteria may be used for determining dimensionality (i.e., the number of factors represented), but one common method is to examine the eigenvalues and to retain only those factors with the highest eigenvalues proportional to other factors. If more than one factor is needed to account for the observed item relationships, this is evidence of multidimensionality. To maintain unidimensionality, items that load onto factors other than the dominant factor are typically removed from the item pool. In practice, no item set is perfectly unidimensional. However, an instrument should have "essential" unidimensionality in that there is only one "major" or dominant dimension and the presence of other "minor" factors is minimal, will likely not affect measurement, and can be ignored (Stout, 1990).

Local Independence

A second key assumption of IRT is local independence. Local independence means the items are independent or uncorrelated after accounting for the trait being measured. Referring again to the example of a reading comprehension test, suppose that clients are asked to read a passage and then answer two questions about it. These two questions may be locally dependent because they relate to the same passage, and responses to both questions depend on how well that one passage was understood. In principle, whenever the response to an item depends on or is related to the response to another item, the assumption of local independence is violated. Local independence can be tested with a variety of statistical techniques. One procedure is to calculate the difference between the IRT model expectation and the actual observation for each response. The correlations of these differences, or residuals, among item pairs are then examined, and large correlations are taken as an indication of local dependence (Yen, 1984, 1993). At least one item in each locally dependent pair is typically considered for removal from the item set.

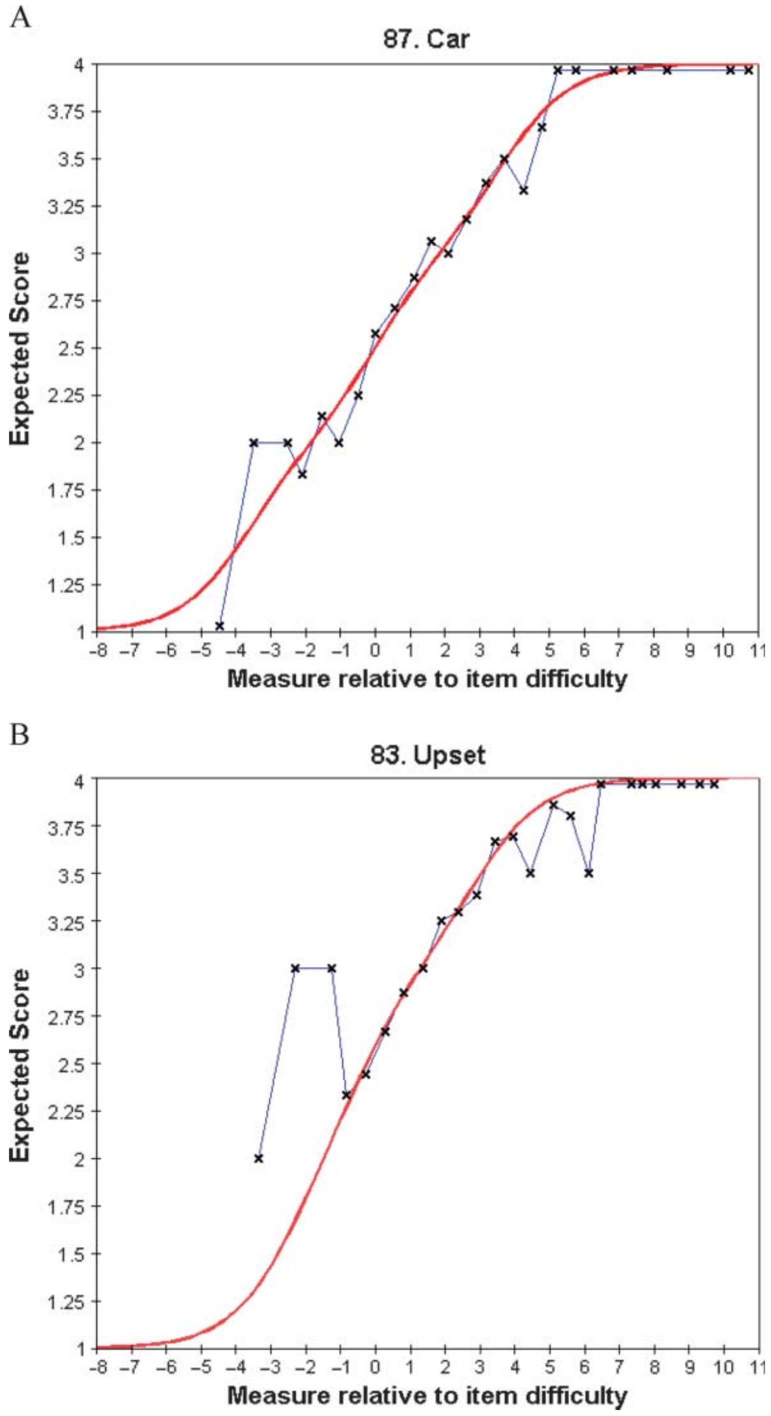
Once the two key assumptions of unidimensionality and local independence have been met, developers may need to prune additional items to arrive at an item bank that provides the measurement qualities desired in the instrument. At this point, instrument developers return to the particular IRT model they have chosen and conduct further analyses described below.

Item Fit

Let's return to Figure 1 and the basic concept of the IRT model. To use the IRT model for estimating latent traits, client responses must follow the pattern shown in the ICC. Clients with higher ability levels must show a higher likelihood of responding to an item correctly (or endorsing an opinion item). If responses obtained from clients do not follow this pattern, then the IRT model is not able to estimate the trait because the relationships among the items, the trait, and the person responses are no longer explained by the mathematical equation. For that reason, one of the first analyses conducted after items pass the tests of unidimensionality and local independence is that of item fit to the model. Item fit is evaluated by calculating the difference between an observed ICC (from actual responses obtained in the instrument development process) and the predicted (theoretical) ICC. Item fit is illustrated in Figure 5 using data from the CPIB (Baylor et al., 2009). In both parts of the figure, the smooth line is the expected ICC curve generated by the IRT model. The ×s represent the actual ICC curve generated from data collected from a sample of clients with spasmodic dysphonia. Notice in Part A (item: having a conversation while riding in a car) that the ×s follows the expected curve rather closely, although not perfectly. This is an example of adequate item fit to the model and shows

[7] There are also multidimensional IRT models that do not require unidimensionality and are used when an item set measures multiple constructs. These models are beyond the scope of this article.

FIGURE 5. Demonstration of item fit. Part A shows good item fit in that the empirical data (×s) follow the model or theoretical curve (smooth line) closely. Part B shows poor item fit to the model due to the deviation of the data from the theoretical curve at the lower end of the curve.



that clients do indeed seem to answer the item according to what the model would predict based on their trait level. Contrast this with Part B (item: getting your point across when you are upset) in which the pattern does not hold. In Part B, the data follow the predicted model for the upper portion of the curve but deviate notably from the model in the lower section of the curve. This is an example of poor item fit.

IRT analysis programs will often provide item fit statistics. Developers typically keep items whose fit statistics fall within a certain recommended range per the IRT model.

Items with poor item fit should be removed from the item bank or revised and retested. Poor item fit may occur for several reasons, including that a construct or variable other than the one of interest may be confounding measurement, the item may be ambiguous, or the item may demonstrate differential item functioning (DIF) across subgroups (discussed below).

DIF

Item bias is a concern in all instrument development. Developers should ensure that responses to the items are based on the latent trait of interest and that there is no systematic pattern of response to the items that is generated by other client characteristics such as age, gender, or socioeconomic status. In IRT, this issue can be addressed through tests of DIF. DIF analyses essentially ask whether item parameters (difficulty and discrimination) are the same across population subgroups (i.e., gender) after controlling for differences in ability. Various statistical or graphical methods are used to test for DIF. DIF investigations are most commonly used to identify and remove items that may introduce bias into the item set. However, substantive theoretical concerns may also drive DIF analyses. For example, Hula, Doyle, and Austermann Hula (2010) used DIF analysis between right-hemisphere and left-hemisphere stroke survivors in a study examining whether self-reported cognitive functioning in stroke can be measured as a single, undifferentiated construct, or whether communicative functioning constitutes a dimension that is separable from other kinds of cognitive functioning.

Measurement Range and Information Functions

One important decision to consider in instrument development is the range of the latent trait that the instrument is intended to estimate. To answer this, the developers need to know the purpose of the instrument. Many SLPs may want an item bank that covers a wide range of a trait because instruments that cover a broad range would be very practical for use with the wide range of clients that many SLPs will encounter in their settings. From this broad item bank, SLPs could draw smaller subsets of items to tailor to different clients, such as those with more or less severe symptoms. There might be situations in which SLPs would want an item bank that focuses exclusively on a narrow range of proficiency, for example, very high levels of cognitive function. This latter type of instrument might be very useful when working with high-functioning brain injury clients. One of the key advantages of IRT is the ability to generate item sets that target specific trait ranges.

To select items for the desired trait range in an instrument, developers return to the concept of item difficulty. It is often helpful to present the items visually along the logit scale, as is done in Figure 6, which is an item-person map often generated in Rasch and 1-PL analyses. The item-person map plots the distribution of both items and persons in the sample along the logit scale, which is oriented vertically in these maps. Figure 6 is representative of data from the CPIB (Baylor et al., 2009), but this map contains a highly reduced subset of items and clients for illustration purposes only. The ordering of items along the logit scale makes logical sense. Items that represent more difficult communication situations such as making a phone call to get information have higher item difficulty levels than easier tasks such as having a casual conversation with someone in a familiar setting.

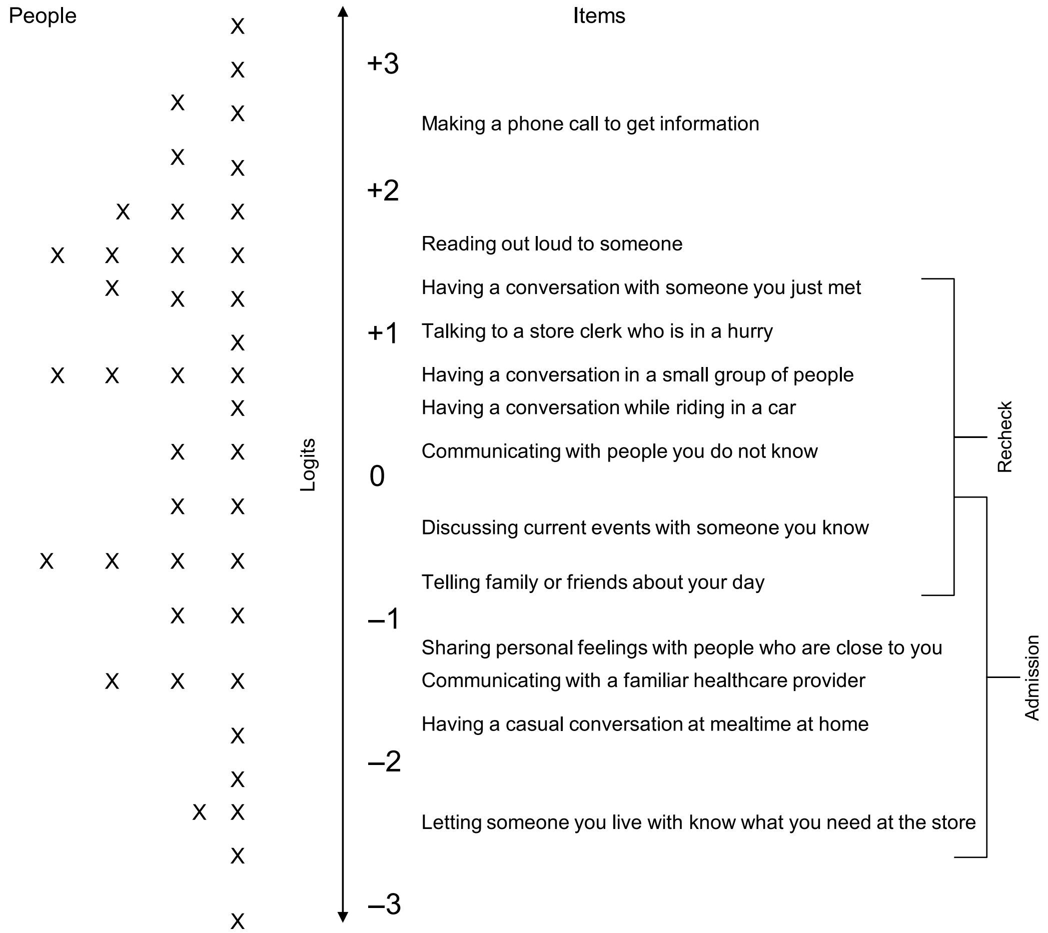
When an item bank is intended for use across a wide range of individuals (i.e., wide range of disorder severity levels), the items should be spread fairly evenly across that range. In Figure 6, the items are spread generally evenly across the range although there are gaps at the high and low ends. There are not many items between 2 and 3 logits or between –2 and –3 logits. Ideally, additional items would be created to fill in these gaps because the lack of items in these regions will lead to less precise measurement of individuals who fall at these trait levels. Sometimes multiple items will have the same or very similar item difficulty values. If instrument developers wish to reduce the number of items in the item bank, they may remove items with duplicate item difficulty levels as long as this does not omit desired content in the items.

In the other situation posed, the SLP may want an instrument that focuses on a narrow range of the trait, such as on very high-level cognitive tasks, to be able to identify individuals with more subtle impairments. For this type of instrument, the developers would construct the item bank to contain mostly items with high logit values. The developers may remove any items at lower difficulty levels and create a more dense collection of items at the high end of the range.

Another way to examine measurement range is to look at item and test information functions. These functions are particularly useful for two-parameter models that consider not only item difficulty but also item discrimination when determining the measurement properties of items. Recall from the section on 2-PL models above that the ordering of items according to difficulty may vary across trait levels when items have different discrimination values. For this reason, item-person maps are not as useful for 2-PL models. Information functions use data about both item difficulty and item discrimination to convey measurement range. An example of an individual item information function from the CPIB is depicted in Figure 7 for the item "Communicating with someone who is not paying attention to you." The *x*-axis is the trait range in logits, and the *y*-axis represents the amount of information available in the item. The term *information* refers to the psychometric information that will help to estimate the trait level and to differentiate among persons. Information is the inverse of measurement error or unexplained variability. Higher information values are associated with less measurement error and are more desirable.

Information functions are useful for understanding just how much of the trait range an item estimates and at what level of precision. Information functions reveal that items are not restricted to estimating only at their exact item difficulty value but instead can cover a broader range of the trait. For example, the item in Figure 7 has a peak in the region of –1.5 logits, suggesting it provides most precise estimation for individuals with ability levels in this region

FIGURE 6. This is a sample of an item-person map generated by a 1-PL model. This is a subset of data from the Communicative Participation Item Bank (CPIB; Baylor et al., 2009). Distributions of both client trait level and item difficulty are presented vertically along the logit scale. Each "×" on the left side of the vertical line represents a client at his or her ability level. High scores are better in that people with high logit scores do better with communicative participation (i.e., report lower interference). The items are on the right side of the vertical line. Items with higher logit values are the more difficult situations (situations in which more clients report interference).



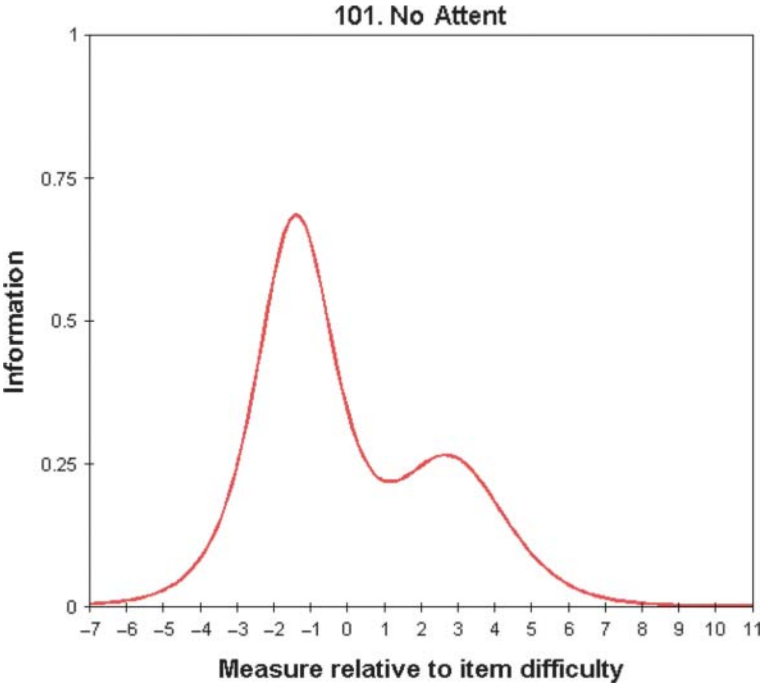
but substantially less precise estimation for clients outside this region. When constructing an item bank, developers can examine the item information functions to choose those items that provide the most information across the trait range for which the instrument is intended. When forming an item bank or a subset of items from a bank, the information functions from each item can be added together to form a curve that represents the measurement precision across the entire item set. This is the test information function. In test information functions, information (y-axis) values higher than 10 are associated with reliability coefficients ≥ .90 (Embretson & Reise, 2000), so measurement is considered to be most reliable in the ability range for which the test information function is above 10.

This section has described some of the key steps in IRT item bank development. Item banks can be changed over time. More items may be added, although this requires gathering responses from a large number of individuals and reanalyzing the entire item bank to ensure that any changes in the measurement properties introduced into the item bank with the new items are identified.

Other Steps in Instrument Development

Before leaving the topic of item bank development, we should note that using IRT in and of itself does not guarantee that an instrument is appropriate and relevant for the measurement purposes at hand. Instrument development requires many steps to ensure that the instrument addresses a relevant and meaningful construct, and estimates it in a valid and reliable manner. These steps should involve a combination of qualitative and quantitative processes, which are described elsewhere (Brod, Tesler, & Christensen, 2009; DeWalt, Rothrock, Yount, & Stone, 2007; Fries, Bruce, & Cella, 2005). The resources of the Patient Reported Outcomes Measurement Information System (www.nihpromis.org), the National Institutes of Health's roadmap initiative for the development of patient-reported outcomes measures, may

FIGURE 7. Item information function that demonstrates the measurement range and discrimination abilities of an item from the CPIB: "Communicating with people who are not paying attention to you." The item provides more precise estimates in the region of –1.5 logits, and the steep peak of the function suggests higher discrimination of clients around that logit region.



also be helpful. SLPs may also wonder about some of the "traditional" tests of validity and reliability that they are used to seeing in test manuals. Many of the IRT processes described above provide documentation of reliability and validity. Evidence of unidimensionality, good item fit, and lack of DIF, for example, all provide support that the instrument is valid and addresses only the intended trait without confounding variables. Developers will want to continue validity and reliability testing to document issues such as comparisons of the new instrument to current "gold standard" instruments, reliability of the instrument with time and retesting, and sensitivity of the instrument to change, particularly for the purpose of measuring treatment outcomes.

Administering the Item Bank

One of the key advantages of IRT—and one of the key differences between IRT and CTT instruments—is the option of adaptive assessment. CTT instruments typically require that all of the items in the scale and/or component subscales must be administered to obtain a valid, reliable, and interpretable score (although many CTT instruments do have stopping rules for establishing basals or ceilings). Instruments based in IRT may be used in the same way with all clients receiving the same set of test items. However, one of the major advantages of IRT is that it permits administration of different subsets of items to different clients (or to the same client at different times) with the ability to compare these scores directly. One of the most significant advantages of adaptive testing is that it increases measurement efficiency, meaning that a client's trait level can be estimated using fewer items, thus lowering response burden while maintaining measurement precision at a level that is commensurate with, or perhaps even more reliable than, longer fixed-length CTT instruments (Cook, O'Malley, & Roddey, 2005). Several new IRT-based instruments that utilize adaptive testing are on the horizon for speech-language pathology.

Adaptive testing begins with an IRT item bank calibrated using the methods of instrument development described above. Item banks can be of widely varying sizes, from 20 to more than 100 items. Ideally, SLPs want to administer as few items as necessary to reduce the burden on the client, particularly if communication impairments make responding to the items challenging. However, SLPs also want to administer enough items to ensure that they have a reasonably precise estimate of the trait level. Instrument developers may prepackage subsets of items in tailored short forms. The different short forms may be appropriate for different groups of clients, perhaps different severity levels of a communication disorder. The most flexible method of adaptive assessment, however, involves starting with the item bank in its entirety and selecting items based on the individual client's responses.

Administering an adaptive IRT-based instrument is somewhat like an audiologist conducting a hearing examination and "zeroing in" on the decibel level for the client's hearing threshold. As the audiologist progresses through the evaluation, he or she adjusts the intensity of the tones up or down depending on whether the client responds to the tones presented. The audiologist quickly tries to focus in on a narrow intensity range, presenting tones that are just slightly louder than tones the client does not hear or just slightly softer than tones the client can hear, until he or she identifies the decibel level where the client responds about half of the time—the client's threshold. As the audiologist narrows in on the client's threshold, the audiologist will not present tones that are much louder or much softer than this range because these tones are not going to help the audiologist zero in on the client's threshold.

In a similar manner, IRT supports adaptive administration, meaning that items are chosen for each client in response to the client's answers to preceding questions in order to zero in on the client's threshold or trait estimate. The first item administered may be an item of moderate difficulty, and after a response is given, the client's ability level is estimated based on this single response. A second item with a high information value for the estimated ability level is selected from the item bank and is administered. After this response, the ability level is re-estimated, and a third item is chosen. With each successive item administered, the estimate of the client's ability becomes more and more precise. Testing typically continues until a target level of reliability is reached or until a preset maximum number of items has been given. Computer algorithms for administration and scoring greatly facilitate adaptive testing and are referred to as Computerized-Adaptive Testing (Cook, O'Malley, & Roddey, 2005; Fries et al., 2005; Ware, Gandek, Sinclair, & Bjorner, 2005).

One of the advantages of adaptive testing is that items that are not going to quickly increase the precision of the trait estimate do not have to be administered. In so doing, the burden of time and energy is reduced for both clinician and client because they do not have to wade through items that are very easy for clients but do not provide much detail about their level of function (the "threshold" or boundary between what they can and cannot do). Also, if a client has a very low level of function, adaptive administration will not lead to presentation of high-difficulty items that can contribute to client frustration or discouragement. Adaptive testing that streamlines assessment is particularly important in our clinical populations because communication disorders may make participating in various assessments difficult.

Adaptive testing also allows for efficient retesting of clients at different times using different items from the bank, yet the scores can be directly compared to each other because all of the items are placed along the same logit scale. Returning to Figure 6, let's assume that when an SLP first administered the CPIB item bank to a new client, the client was experiencing many difficulties in daily communication situations. Adaptive administration led to presentation of items shown in the "admission" bracket in Figure 6. The client's trait level was estimated at –1.4 logits. Three months later, the item bank was readministered to document progress. Adaptive administration led to presentation of items in the "Recheck" section of Figure 6, and her trait level was estimated at 0.7 logits. Even though she was presented with different items at different times, her two trait estimates can be directly compared to measure change over time (an improvement of 2.1 logits). In the same manner, comparisons can be made across different individuals using the same item bank even if the individuals are administered different items. Comparisons can also be made across two different item banks measuring the same construct (i.e., two reading comprehension item banks), albeit with preliminary work to link the two item banks statistically.

Conclusion

IRT and Rasch instruments have long been used in educational and aptitude testing. IRT has not been as prevalent in SLP instruments, but new clinically relevant IRT-based instruments are emerging. SLPs will need to be familiar with the basic terminology and principles of IRT assessment to ensure appropriate implementation of these instruments. The purpose of this article was to provide an introduction to some of the key terminology and concepts to assist SLPs as they begin to work with IRT instruments. While some principles of IRT assessment including the logit scale, measurement models, and adaptive administration may require some "recalibrating" of our own ideas or understanding about measurement, the advantages of IRT and Rasch models in terms of measurement quality, efficiency, and accessibility for our clients make a strong case for including IRT-based instruments in SLP practice.

Acknowledgments

This project was supported in part by National Institute on Deafness and Other Communication Disorders Grant 1R03DC010044 (Baylor), VA Rehabilitation Research and Development Career Development Award C6210M (Hula), VA RR&D Merit Review Project C6098R (Doyle), and Louisiana Board of Regents Research Competitiveness Project LESQSF(2008-11)-RD-A-07 (Donovan). We gratefully acknowledge the assistance of Clement Stone, who provided comments on an earlier draft of this article.

References

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581–594.

Andrich, D. (1978b). Rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Baylor, C. R., Yorkston, K. M., Eadie, T. L., Miller, R. M., & Amtmann, D. (2009). Developing the Communicative Participation Item Bank: Rasch analysis results from a spasmodic dysphonia sample. *Journal of Speech, Language, and Hearing Research, 52,* 1302–1320.

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

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16,* 21–33.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Brod, M., Tesler, L. E., & Christensen, T. L. (2009). Qualitative research and content validity: Developing best practices based on science and experience. *Quality of Life Research, 18,* 1263–1278.

Cherney, L. R., Halper, A. S., Heinemann, A. W., & Semik, P. E. (1996). RIC evaluation of communication problems in right hemisphere dysfunction—revised (RICE-R): Statistical background. In A. S. Halper, L. R. Cherney, & M. S. Burns (Eds.), *Clinical management of right hemisphere dysfunction* (pp. 31–40). Gaithersburg, MD: Aspen.

Cook, K. F., O'Malley, K. J., & Roddey, T. S. (2005). Dynamic assessment of health outcomes: Time to let the CAT out of the bag? *Health Services Research, 40,* 1694–1711.

Cook, K. F., Roddey, T. S., O'Malley, K. J., & Gartsman, G. M. (2005). Development of a Flexilevel Scale for use with computer-adaptive testing for assessing shoulder function. *Journal of Shoulder and Elbow Surgery, 14*(1, Suppl. 1), S90–S94.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth/Thomson Learning.

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(5, Suppl. 1), S1–S10.

Donovan, N. J., Kendall, D., Young, M. E., & Rosenbek, J. C. (2008). The Communicative Effectiveness Survey: Preliminary evidence of construct validity. *American Journal of Speech-Language Pathology, 17,* 335–347.

Donovan, N. J., Velozo, C. A., & Rosenbek, J. C. (2007). The Communicative Effectiveness Survey: Investigating its item-level psychometric properties. *Journal of Medical Speech-Language Pathology, 15,* 433–447.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test—III*. Circle Pines, MN: AGS.

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

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Clinical and Experimental Rheumatology, 23*(Suppl. 39), S53–S57.

German, D. J. (1990). *Test of Adolescent/Adult Word Finding*. Austin, TX: Pro-Ed.

Gravatter, F. J., & Wallnau, L. B. (2000). *Statistics for the behavioral sciences* (5th ed.). Belmont, CA: Wadsworth/Thomson Learning.

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

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9, Suppl. 2), S28–S42.

Hula, W., Austermann Hula, S. N., & Doyle, P. J. (2009). A preliminary evaluation of the reliability and validity of a self-reported communicative functioning item pool. *Aphasiology, 23,* 783–796.

Hula, W., Donovan, N. J., Kendall, D., & Gonzalez-Rothi, L. J. (2010). Item response theory analysis of the Western Aphasia Battery. *Aphasiology, 24,* 1326–1341.

Hula, W., Doyle, P. J., & Austermann Hula, S. N. (2010). Patient-reported cognitive and communicative functioning: One construct or two? *Archives of Physical Medicine and Rehabilitation, 91,* 400–406.

Jette, A. M., & Haley, S. M. (2005). Contemporary measurement techniques for rehabilitation outcomes assessment. *Journal of Rehabilitation Medicine, 37,* 339–345.

Justice, L. M., Bowles, R. P., & Skibbe, L. E. (2006). Measuring preschool attainment of print-concept knowledge: A study of typical and at-risk 3- to 5-year-old children using item response theory. *Language, Speech, and Hearing Services in Schools, 37,* 224–235.

Kendall, D., del Toro, C., Nadeau, S., Johnson, J., Rosenbek, J. C., & Velozo, C. A. (2010, May). *The development of a standardized assessment of phonology in aphasia: Construct validity, sensitivity and test retest reliability*. Paper presented at the Clinical Aphasiology Conference, Isle of Palms, SC.

Kertesz, A. (1982). *The Western Aphasia Battery*. Philadelphia, PA: Grune & Stratton.

Linacre, J. M. (1991). Winsteps 3.63.0: Multiple choice, rating scale, and partial credit Rasch analysis [Computer software]. Chicago, IL: Mesa Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328. Retrieved from www.rasch.org/rmt/rmt74m.htm.

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

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

Masters, G. N., & Wright, B. D. (1996). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York, NY: Springer.

Maxwell, S. E., & Delaney, H. D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97,* 85–93.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88,* 355–383.

Michell, J. (2003). Measurement: A beginner's guide. *Journal of Applied Measurement, 4,* 298–308.

Milman, L. H., Holland, A., Kaszniak, A. W., D'Agostino, J., Garrett, M., & Rapcsak, S. (2008). Initial validity and reliability of the SCCAN: Using tailored testing to assess adult cognition and communication. *Journal of Speech, Language, and Hearing Research, 51,* 49–69.

Muthén, L. K., & Muthén, B. O. (1998). Mplus (Version 4.2) [Computer software]. Los Angeles, CA: Author.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Thousand Oaks, CA: Sage.

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (2nd ed.). Chicago, IL: University of Chicago Press.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . PROMIS Cooperative Group. (2007). Psychometric evaluation and calibration of health-related quality of life item banks. *Medical Care, 45*(5, Suppl. 1), S22–S31.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using Multilog. *Journal of Educational Measurement, 27,* 133–144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (*Psychometrika* Monograph No. 17). Richmond, VA: The Psychometric Society.

Samejima, F. (1996). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.

Smith, E. V., Jr., & Smith, R. M. (2004). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

Stevens, S. S. (1946, June 7). On the theory of scales of measurement. *Science, 103,* 677–680.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of Multilog. *Applied Psychological Measurement, 16,* 1–6.

Stout, W. F. (1990). A new item response theory modeling approach with applications to undimensionality assessment and ability estimation. *Psychometrika, 55*(2), 293–325.

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). Multilog (Version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.

Ware, J. E., Gandek, B., Sinclair, S. J., & Bjorner, J. B. (2005). Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology, 50*(1), 71–78.

Willmes, K. (1981). A new look at the token test using probabilistic test models. *Neuropsychologia, 19,* 631–645.

Willmes, K. (2003). Psychometric issues in aphasia therapy research. In I. Papathanasiou & R. De Bleser (Eds.), *The sciences of aphasia: From theory to therapy* (pp. 227–244). Amsterdam, the Netherlands: Pergamon.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Erlbaum.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). Bilog-MG 3 [Computer software]. Lincolnwood, IL: Scientific Software International.