An Introduction to Modern Measurement Theory

This tutorial was written as an introduction to the basics of item response theory (IRT) modeling and its applications to health outcomes measurement for the National Cancer Institute's Cancer Outcomes Measurement Working Group (COMWG). In no way this tutorial is meant to replace any text on measurement theory, but only serve as a stepping-stone for health care researchers to learn this methodology in a framework that may be more appealing to their research field. An illustration of IRT modeling is provided in the appendix and can provide better insight into the utility of these methods. This tutorial is only a draft version that may undergo many revisions.

Please feel free to comment or ask questions of the author of this piece:

Bryce B. Reeve, Ph.D.
Outcomes Research Branch
Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
Tel: (301) 594-6574
Fax: (301) 435-3710
Email: reeveb@mail.nih.gov

Outline

Reference: Some Terms You Will See in this Tutorial as well as in the Literature

Assumption of Local Independence – A response to a question is independent of responses to other questions in a scale after controlling for the latent trait (construct) measured by the scale.

Assumption of Unidimensionality - the set of questions are measuring a single continuous latent variable (construct).

Construct (a.k.a. trait, domain, ability, latent variable, or theta) - see definition of theta below.

Information Function (for a scale or item) - an index, typically displayed in a graph, indicating the range of trait level θ over which an item or test is most useful for distinguishing among individuals. The information function characterizes the precision of measurement for persons at different levels of the underlying latent construct, with higher information denoting more precision.

Item – Question in a scale

Item Characteristic Curve (ICC, a.k.a. item response function, IRF, or item trace line) – The ICC models the relationship between a person's probability for endorsing an item category and the level on the construct measured by the scale.

Item Difficulty (Threshold) Parameter *b* – point on the latent scale θ where a person has a 50% chance of responding positively to the scale item.

Item Discrimination (Slope) Parameter *a* - describes the strength of an item's discrimination between people with trait levels (θ) below and above the threshold *b*. The *a* parameter may also be interpreted as describing how an item may be related to the trait measured by the scale.

Scale (measure, questionnaire, or test) – A scale in this tutorial is assumed to measure a single construct or domain.

Slope Parameter – See Item Discrimination

Test Characteristic Curve (TCC, a.k.a. Test Response Function) - The TCC describes the expected number of scale items endorsed as a function of the underlying latent variable.

Theta (θ)– The unobservable (or latent) construct being measured by the questionnaire. These constructs or traits are measured along a continuous scale. Examples of constructs in health outcomes measurement are depression level, physical functioning, anxiety, or social support.

Threshold Parameter – See Item Difficulty

An Introduction to Modern Measurement Theory

Each year new health outcomes measures are developed or revised from previous measures in the hope of obtaining instruments that are more reliable, valid, sensitive, and interpretable. This increasing need for psychometrically-sound measures calls for better analytical tools beyond what traditional measurement theory (or classical test theory, CTT) methods can provide. In the past decade, applications of item response theory (IRT) in health research measurement have increased considerably because of its utility in item and scale analysis, scale scoring, instrument linking, and adaptive testing. IRT is a model-based measurement in which trait level estimates (e.g., level of physical functioning or level of depression) depend on both persons' responses and on the properties of the questions that were administered (Embretson and Reise, 2000).

IRT has a number of advantages over CTT methods to assess health outcomes. CTT statistics such as item difficulty (proportion of correct responses), item discrimination (corrected item-total correlation), and reliability are contingent on the sample of respondents to whom the questions were administered. IRT item parameters are not dependent on the sample used to generate the parameters, and are assumed to be invariant (within a linear transformation) across divergent groups within a research population and across populations. In addition, CTT yields only a single estimate of reliability and corresponding standard error of measurement, whereas IRT models measure scale precision across the underlying latent variable being measured by the instrument (Cooke & Michie, 1997; Hays, Morales, & Reise, 2000). A further disadvantage of CTT methods is that a participant's score is dependent on the set of questions used for analysis, whereas, an IRT-estimated person's trait level is independent of the questions being used.

Because the expected participant's scale score is computed from their responses to each item (that is characterized by a set of properties), the IRT estimated score is sensitive to differences among individual response patterns and is a better estimate of the individual's true level on the trait continuum than CTT's summed scale score (Santor & Ramsay, 1998).

Because of these advantages, IRT is being applied in health outcomes research to develop new measures or improve existing measures, to investigate group differences in item and scale functioning, to equate scales for cross-walking patient scores, and to develop computerized adaptive tests. This handbook provides an overview of IRT and the more commonly used models in health outcomes research, along with a discussion of the applications of these models to develop valid, reliable, sensitive, and feasible endpoint measures.

A Brief History of Item Response Theory (incomplete)

While many think of item response theory as modern psychometric theory, the concepts and methodology of IRT has been developed for over three-quarters of a century. L. L. Thurstone (1925) laid down the conceptual foundation for IRT in his paper, entitled "A Method of Scaling Psychological and Educational Tests." In it, he provides a technique for placing the items of the Binet and Simon (1905) test of children's mental development on an age-graded scale. Plots of the proportions of children in successive age cross-sections succeeding on successive Binet tasks and the effective location of each item on chronological age reflect many of the features suggestive of IRT (Bock, 1997).

Thurstone dropped his work in measurement to pursue the development of multiple factor analysis, but his colleagues and students continued to refine the theoretical bases of IRT (Steinberg & Thissen, in draft). Richardson (1936) and Ferguson (1943) introduced the normal ogive model as a means to display the proportions correct for individual items as a function of

normalized scores. Lawley (1943) extended the statistical analysis of the properties of the normal ogive curve and described maximum-likelihood estimation procedures for the item parameters and linear approximations to those estimates. Fred Lord (1952) introduced the idea of a latent trait or ability and differentiated this construct from observed test score. Lazarsfeld (1950) described the unobserved variable as accounting for the observed interrelationships among the item responses.

Considered a milestone in psychometrics (Embretson & Reise, 2000), Lord and Novick's (1968) textbook entitled *Statistical Theories of Mental Test Scores* provides a rigorous and unified statistical treatment of classical test theory. The remaining half of the book, written by Allen Birnbaum, provides an equally solid description of the IRT models. Bock, and several student collaborators at the University of Chicago, including David Thissen, Eiji Muraki, Richard Gibbons, and Robert Mislevy developed effective estimation methods and computer programs such as Bilog, Multilog, Parscale, and Testfact. Along with Aitken (Bock & Aitken, 1981), Bock developed the algorithm of marginal maximum likelihood method to estimate the item parameters that are used in many of these IRT programs.

In a separate line of development of IRT models, Georg Rasch (1960) discussed the need for creating statistical models that maintain the property of *specific objectivity*, the idea that people and item parameters be estimated separately but comparable on a similar metric. Rasch inspired Gerhard Fischer (1968) to extend the applicability of the Rasch models into psychological measurement and Ben Wright to teach these methods and help to inspire other students to the development of the Rasch models. These students, including David Andrich, Geoffrey Masters, Graham Douglas, and Mark Wilson, helped to push the methodology into education and behavioral medicine (Wright, 1997).

Item Response Theory Model Basics

IRT is a model for expressing the association between an individual's response to an item and the underlying latent variable (often called "ability" or "trait") being measured by the instrument. The underlying latent variable in health research may be any measurable construct such as physical functioning, risk for cancer, or depression. The latent variable, expressed as theta (θ), is a continuous unidimensional construct that explains the covariance among item responses (Steinberg & Thissen, 1995). People at higher levels of θ have a higher probability of responding correctly or endorsing an item.

IRT models use item responses to obtain scaled estimates of θ, as well as to calibrate items and examine their properties (Mellenbergh, 1994). Each item is characterized by one or more model parameters. The item difficulty, or threshold, parameter *b* is the point on the latent scale θ where a person has a 50% chance of responding positively to the scale item (question). Items with high thresholds are less often endorsed (Steinberg & Thissen, 1995). The slope, or discrimination, parameter *a* describes the strength of an item's discrimination between people with trait levels (θ) below and above the threshold *b*. The *a* parameter may also be interpreted as describing how an item may be related to the trait measured by the scale and is directly related, under the assumption of a normal θ distribution, to the biserial item-test correlation ρ (Linden & Hambleton, 1997). For item *i* the relationship is:

$$a_i = \frac{\rho_i}{\sqrt{1-\rho_i^2}}.$$

The slope parameter is often thought of and is linearly related (under some conditions) to the variable loading in a factor analysis. Some IRT models, in education research, include a lower-

asymptote parameter or guessing parameter *c* to possibly explain why people of low levels of the trait θ are responding positively to an item.

To model the relation of the probability of a correct response to an item conditional on the latent variable θ, trace lines, estimated from the item parameters, are plotted. Most IRT models in research assume that the normal ogive or logistic function describes this relationship accurately and fits the data. The logistic function is similar to the normal ogive function, and is mathematically simpler to use and, as a result, is predominately used in research. The trace line (or sometimes called the item characteristic curve, ICC) can be viewed as the regression of item score on the underlying variable θ (Lord, 1980, p. 34). The left graph in Figure 1 models

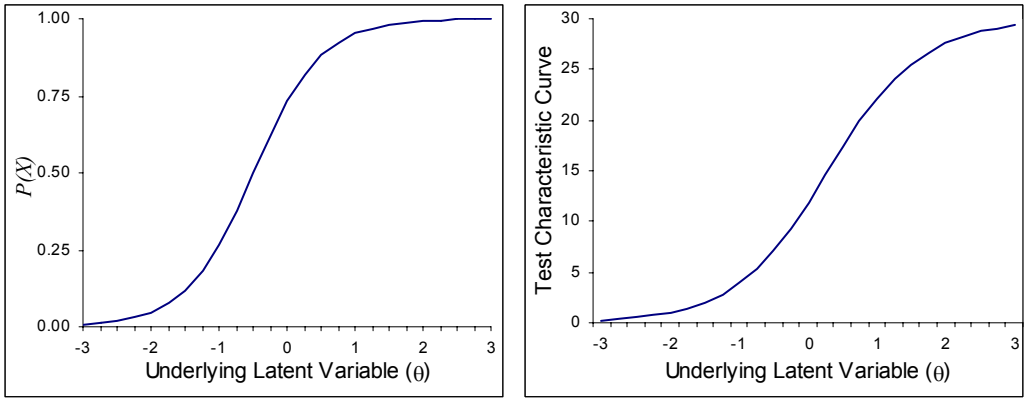


Figure 1: Left figure is an IRT item characteristic curve (trace line) for one item. Right figure is test characteristic curve for thirty items.

the probability for endorsing an item conditional on the level on the underlying trait. The higher a person's trait level (moving from left to right along the θ scale), the greater the probability that the person will endorse the item. For example, if a question asked, "Are you unhappy most of the day?", then the left graph of Figure 1 shows that people with higher levels of depression (θ) will have higher probabilities for answering *yes* to the question. For dichotomous items, the probability of a negative response is high for low values of the underlying variable being measured, and decreases for higher levels on θ. The collection of the item trace lines forms the

scale; thus, the sum of the probabilities of the correct response of the item trace lines yields the test characteristic curve (TCC). The TCC describes the expected number of scale items endorsed as a function of the underlying latent variable. The right figure of Figure 1 presents a TCC curve for 30 items. When the sum of the probabilities is divided by the number of items, the TCC gives the average probability or expected proportion correct as a function of the underlying construct (Weiss, 1995).

Another important feature of IRT models is the information function, an index indicating the range of trait level θ over which an item or test is most useful for distinguishing among individuals. In other words, the information function characterizes the precision of measurement for persons at different levels of the underlying latent construct, with higher information denoting more precision. Graphs of the information function place persons' trait level on the horizontal *x*-axis, and amount of information on the vertical *y*-axis (left graph in Figure 2).

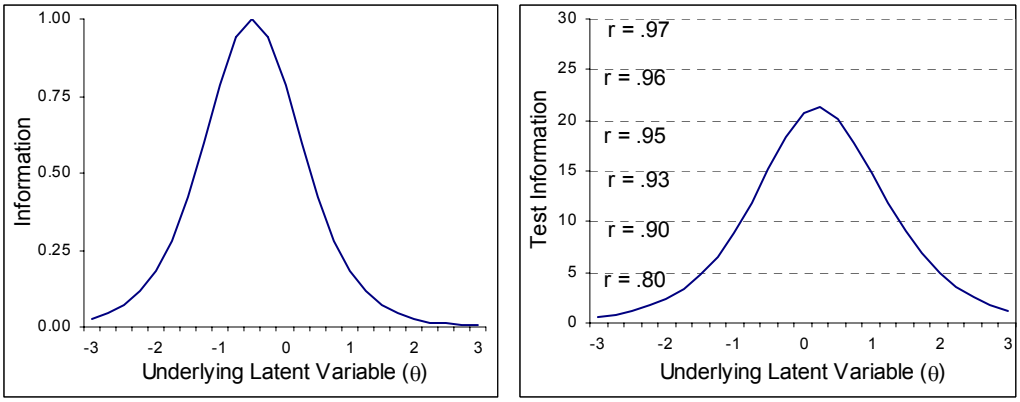


Figure 2: Item information curve on left side. Test information curve with approximate test reliability on right side.

The shape of the item information function is dependent on the item parameters. The higher the item's discrimination, the more peaked the information function will be; thus, higher discrimination parameters provide more information about individuals whose trait levels (θ) lie near the item's threshold value. The item's difficulty parameter(s) determines where the item

information function is located (Flannery, Reise, & Widaman, 1995). With the assumption of local independence (reviewed below), the item information values can be summed across all of the items in the scale to form the test information curve (Lord, 1980).

At each level of the underlying trait θ, the information function is approximately equal to the expected value of the inverse of the squared standard errors of the θ-estimates (Lord, 1980). The smaller the standard error of measurement (SEM), the more information or precision the scale provides about θ. For example, if a measure has a test information value of 16 at θ = 2.0, then examinee scores at this trait level have a standard error of measurement of $\left(1/\sqrt{16}\right)$ = .25, indicating good precision (reliability approximately .94) at the level of theta (Flannery et al., 1995). The right graph in Figure 2 presents a test information function. Most information (precision in measurement) is contained within the middle of the scale (-1.0 < θ < 1.5) with less reliability (labeled *r* in graph) at the high and low ends of the underlying trait. To observe the conditional standard error of measurement for a given scale, the inverse of the square root of the test information function across all levels of the θ continuum is graphed.

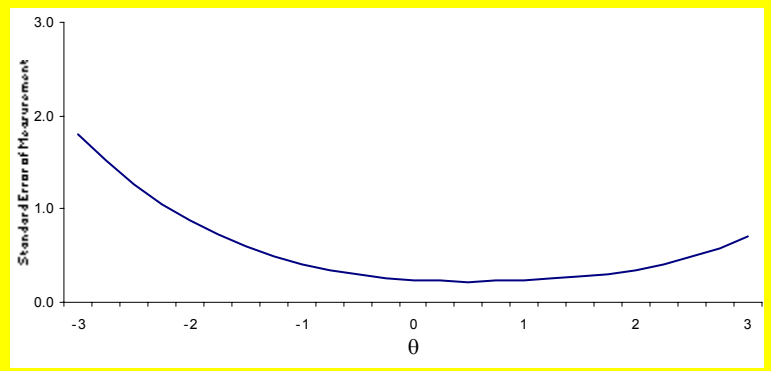


Figure 3: Test standard error of measurement

Figure 3 presents the test SEM with little error (more precision) in measurement for middle to high levels of the underlying trait, and high error in measuring respondents with low levels of θ. If the scale, presented in Figures 2 and 3, measures physical functioning from poor (θ = -3) to

high ($\theta = +3$), then the scale lacks precision in measuring physically disabled patients ($\theta < -1.5$), but adequately captures the physical functioning of average to healthy (ambulatory) individuals.

Scale scoring in item response theory has a major advantage over classical test theory. In classical test theory, the summed scale score is dependent on the difficulty of the items used in the selected scale, and therefore, not an accurate measure of a person's trait level. The procedure assumes that equal ratings on each item of the scale represent equal levels of the underlying trait (Cooke & Michie, 1997). Item response theory, on the other hand, estimates individual latent trait level scores based on all the information in a participant's response pattern. That is, IRT takes into consideration, which items were answered correctly (positively) and which ones were answered incorrectly, and utilizes the difficulty and discrimination parameters of the items when estimating trait levels (Weiss, 1995). Persons with the same summed score but different response patterns may have different IRT estimated latent scores. One person may answer more of the highly discriminating and difficult items and receive a higher latent score than one who answers the same number of items with low discrimination or difficulty. IRT trait level estimation uses the item response curves associated with the individual's response pattern. A statistical procedure, such as maximum likelihood estimation, finds the maximum of a likelihood function created from the product of the population distribution with the individual's trace curves associated with each item's right or wrong response. A full discussion of trait scoring follows the overview of the IRT models.

Assumptions of Item Response Theory Models

The IRT model is based on the assumption that the items are measuring a single continuous latent variable θ ranging from $-\infty$ to $+\infty$. The unidimensionality of a scale can be evaluated by performing an item-level factor analysis, designed to evaluate the factor structure

underlying the observed covariation among item responses. The assumption can be examined by comparing the ratio of the first to the second eigenvalue for each scaled matrix of tetrachoric correlations. This ratio is an index of the strength of the first dimension of the data. Similarly, another indication of unidimensionality is that the first factor accounts for a substantial proportion of the matrix variance (Lord, 1980; Reise & Waller, 1990). For tests using many items, the assumption of unidimensionality may be unrealistic; however, Cooke and Michie (1997) report that IRT models are moderately robust to departures from unidimensionality. If multidimensionality exists, the investigator may want to consider dividing the test into subtests based on both theory and the factor structure provided by the item-level factor analysis. Multi-dimensional IRT models do exist, but its models as well as informative documentation and user-friendly software are still in development.

In the IRT model, the item responses are assumed to be independent of one another: the assumption of local independence. The only relationship among the items is explained by the conditional relationship with the latent variable θ. In other words, local independence means that if the trait level is held constant, there should be no association among the item responses (Thissen & Steinberg, 1988). Violation of this assumption may result in parameter estimates that are different from what they would be if the data were locally independent; thus, selecting items for scale construction based on these estimates may lead to erroneous decisions (Chen & Thissen, 1997). The assumptions of unidimensionality and local independence are related in that; items found to be locally dependent will appear as a separate dimension in a factor analysis.

For some IRT models, the latent variable (not the data response distribution) is assumed to be normally distributed within the population. Without this assumption, estimates of θ for

some response patterns (e.g., respondents who do not endorse any of the scale items) have no finite values resulting in unstable parameter estimates (Chen & Thissen, 1997).

Item Response Theory Models

Two Theories towards Measurement

Thissen and Orlando (2001) discuss two approaches to model building in item response theory. One approach is to develop a well-fitting model to reflect the item response data by parameterizing the ability or trait of interest as well as the properties of the items. The goal of this approach is item analysis. The model should reflect the properties of the item response data sufficiently and accurately, so that the behavior of the item is summarized by the item parameters. The philosophy is that the items are assumed to measure as they do, not as they should (Thissen & Orlando, 2001). This approach to model building believes the theory of measurement is to explain (i.e., model) the data.

Another approach of IRT model building is to obtain specific measurement properties defined by the model to which the item response data must fit. If the item or a person does not fit within the measurement properties of the IRT model, assessed by analysis of residuals (i.e., item and person fit statistics), the item or person is discarded. This approach follows that of the Rasch (1960) models, and in the cases where the data fits the model, offers a simple interpretation for item analysis and scale scoring. This approach to model building believes optimal measurement is defined mathematically, and then the class of item response models that yield such measurement is derived.

The two approaches described above yield a division in psychometrics. Those who believe health research measurement should be about describing the behaviors behind the response patterns in a survey will use the most appropriate IRT model (e.g., Rasch/One-

Parameter Logistic Model, Two-Parameter Logistic Model, Graded Model) to fit the data. The choice of the IRT model is data dependent. Researchers from the Rasch tradition believe that the only appropriate models to use are the Rasch family of models, which retain strong mathematical properties such as specific objectivity (person parameters and item parameters estimated separately) and summed score simple sufficiency (no information from the response pattern is needed) (see model descriptions below). Several advantages of the Rasch model include: the ability of the model to produce more stable estimates of person and item properties when there is a small number of respondents, when extremely non-representative samples are used, and when the population distribution over the underlying trait is heavily skewed.

Embretson and Reise (2000) suggest one should use the Rasch family of models when each item carries equal weight (i.e., each item is equally important) in defining the underlying variable, and when strong measurement model properties (i.e., specific objectivity, simple sufficiency) are desired. If one desires fitting an IRT model to existing data or desires highly accurate parameter estimates, then a more complex model such as the Two-Parameter Logistic Model or Graded Model should be used.

The IRT models

Table 1 presents seven common IRT models with potential application to health-related research. The table also indicates if the model is appropriate for dichotomous (binary) or polytomous (3 or more options) responses, and some characteristics associated with each model. Models noted with an asterisk are part of the Rasch family of models and, therefore, retain the unique properties of summed score simple sufficiency and specific objectivity. Each of the models are discussed below. Because of the separate development of the Rasch and One-Parameter Logistic models, they are discussed individually.

Table 1

Model (* = belongs to Rasch Family)	Item Response Format	Model Characteristics
Rasch Model* / One Parameter Logistic Model	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
Two Parameter Logistic Model	Dichotomous	Discrimination and threshold parameters vary across items.
Three Parameter Logistic Model	Dichotomous	Includes psuedo-guessing parameter
Graded Model	Polytomous	Ordered responses. Discrimination varies across items.
Nominal Model	Polytomous	No pre-specified item order. Discrimination varies across items.
Partial Credit Model*	Polytomous	Discrimination power constrained to be equal across items.
Rating Scale Model*	Polytomous	Discrimination equal across items. Item threshold steps equal across items.

The Rasch Simple Logistic Model

Rasch (1960) was the first to develop the one-parameter logistic model (sometimes referred to as the simple logistic model), however this model differed from models discussed below. In the Rasch Model, a person is characterized by a level on a latent trait ξ, and an item is characterized by a degree of difficulty δ. The probability of an item endorsement is a function of the ratio of a person's level on the trait to the item difficulty ξ/δ (Tinsley, 1992).

Given that the data adequately fit the Rasch model, one can make simple comparisons of the items and respondents according to the principles of *specific objectivity*. *Specific objectivity* means that comparison of two items' difficulty parameters are assumed to be independent of any group of subjects being surveyed, and the comparison of two subjects' trait levels does not depend on any subset of items being administered (Mellenbergh, 1994). The Rasch model assumes that the items are all equal in discrimination (weight equally on a factor) and that chance factors (e.g. guessing) do not influence the response.

For a particular item, Rasch proposed a simple trace line (probability) function, that increases from zero to one with trait level, as:

$$T = \frac{\xi}{\xi + \delta}$$

(Thissen & Orlando, 2001). The model in this form has the interpretation of the probability of a positive response being equal to the value of the person parameter ξ relative to the value of the item parameter δ (Linden & Hambleton, 1997). If we use current item response theory notation, substituting exp θ for ξ and exp b for δ, we have:

$$T = \frac{\exp\theta}{\exp\theta + \exp b} = \frac{1}{1 + \exp[-(\theta - b)]}$$

(Thissen & Orlando, 2001). As before, theta (θ) represents a person's trait level, and b represents the item threshold. This model shows the dependent variable, the probability of endorsing an item, as a function of the difference between two independent variables, the person's level on the underlying trait θ and the item threshold b (difficulty). Rasch constrained the sum of the difficulty parameters for all scale items to be equal to zero ($\sum b = 0$), thus setting the scale of the θ parameter. Given this constraint, the population distribution of θ is unspecified. The distribution "has some mean, relative to the average item difficulty, and some variance, relative to the unit slope of the trace lines.... The shape of the population distribution [of θ] is unknown; it is whatever shape it has to be to produce the observed score distribution" (Thissen & Orlando, 2001, p. 76-77).

Figure 4 presents seven item trace lines for varying ranges of threshold parameters (-1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5). Items with higher threshold parameters are less often endorsed and require high levels of the underlying trait θ to endorse the item.

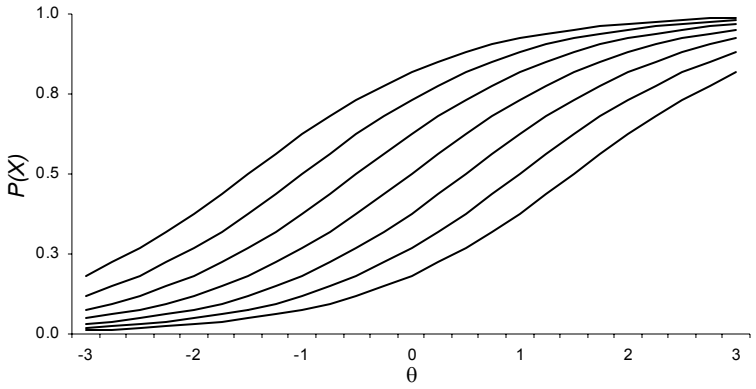


Figure 4 Rasch / One Parameter Logistic IRT Model (7 items)

The One-Parameter Logistic Model

The development of the Rasch model was independent of the development of the one-parameter logistic model, but both have similar features and are mathematically equivalent. The one-parameter logistic (1PL) model trace line for a given item *i* is:

$$T_i(u_i = 1|\theta) = \frac{1}{1 + \exp[-a(\theta - b_i)]}.$$

$T_i(u_i = 1|\theta)$ traces the conditional probability of a positive $(u_i = 1)$ response to item *i* as a function of the trait parameter θ, the threshold or difficulty parameter b_i, and the discrimination parameter *a*. Where the Rasch model had a fixed slope of one for all items, the 1PL model only requires the slope to be equal for all items (Thissen & Orlando, 2001).

The population distribution of the underlying variable θ for the one-parameter logistic model (as well as the two and three-parameter logistic models) is usually specified to have a population mean of zero and a variance of one. The threshold (or difficulty) parameters b_i are located relative to zero, which is the average trait level in the population, and the slope parameter *a* takes some value relative to the unit standard deviation of the latent variable (Thissen & Orlando, 2001). Thus, it is the latent variable θ the model is assuming to be normally distributed, not the categorical item responses (Thissen & Steinberg, 1988).

The Two-Parameter Logistic Model

The two-parameter logistic model (2PL; Birnbaum, 1968) allows the slope or discrimination parameter *a* to vary across items instead of being constrained to be equal as in the one-parameter logistic or Rasch model. The relative importance of the difference between a person's trait level and item threshold is determined by the magnitude of the discriminating power of the item (Embretson & Reise, 2000). The two-parameter logistic model trace line for the probability of a positive response to item *i* for a person with latent trait level θ is:

$$T_i(u_i = 1|\theta) = \frac{1}{1 + \exp[-1.7a_i(\theta - b_i)}.$$

The constant, 1.7, is added to the model as an adjustment so that the logistic model approximates the normal ogive model. Approximately half of the literature includes the adjustment and half does not (Thissen & Steinberg, 1988).

Figure 5 presents 2PL trace lines for five items with varying threshold/difficulty *b* and discrimination *a* parameters. The item marked by a dashed line represents an item with little relationship with the underlying variable being measured by the survey. Items with steeper sloped trace lines have more discriminating power.

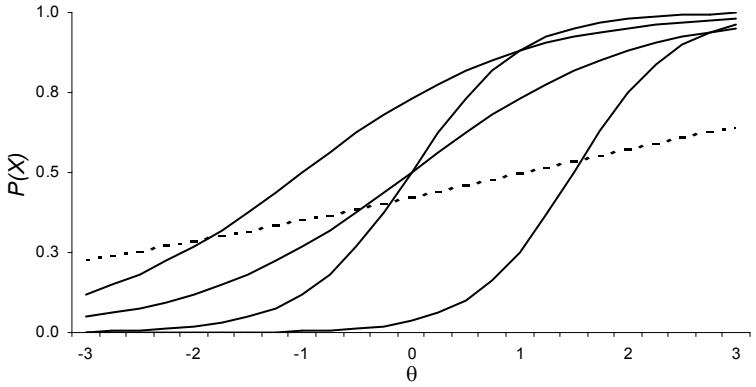


Figure 5 Two Parameter Logistic Trace Lines (5 items)

The Three-Parameter Logistic IRT Model

The three-parameter logistic model (3PL; Lord, 1980) was developed in educational testing to extend the application of item response theory to multiple choice items that may elicit guessing. For item *i*, the three-parameter logistic trace line is:

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

The guessing parameter *c* is the probability of a positive response to item *i* even if the person does not know the answer. When *c* = 0, the three-parameter model is equivalent to the 2PL model. Including the guessing parameter changes the interpretation of other parameters in the model. The threshold parameter *b* is the value of theta at which respondents have a (.5 + .5*c*)*(100)% chance of responding correctly to the item (Thissen & Orlando, 2001).

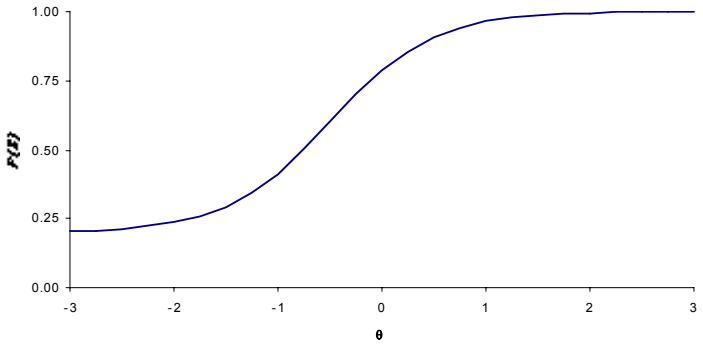


Figure 6 Three Parameter Logistic Model (1 item)

Figure 6 presents a 3PL model trace line for one item. Survey respondents low on the underlying trait have a 20 percent probability (*c* = .2) of endorsing the item.

The interpretation of the guessing parameter for multiple choice tests in educational measurement is straightforward, but can be vague in health measurement. In most health-related measurement research, the guessing parameter is left out, thus opting for the 2PL model. However, the guessing parameter may provide insightful information to understanding the

behavior of participants in the questionnaire. There may be a considerable proportion of participants in the survey who may respond positively to an item for other reasons besides the trait being measured. In ability testing (in an educational context), it is usually assumed that respondents will usually inflate their abilities by guessing at the right answer. However in self-report surveys, respondents may over report desirable behaviors or attitudes, and underreport painful or embarrassing behaviors (Schaeffer, 1988). In other words, persons may be motivated to conceal their true trait level by claiming to have or not have symptoms.

The Graded Model

For questions with three or more response categories, Samejima (1969) proposed a model for graded or ordered responses. A response may be graded on a range of scores, as an example, from poor (0) to excellent (9). For survey measurement, a subject may chose one option out of a number of graded options, such as a five-point Likert-type scale: "strongly-disagree", "disagree", "neutral", "agree", and "strongly agree" (Mellenbergh, 1994).

From dichotomously-scored items to polytomously-scored items, item response theory adapts to the transition more easily than classical (or traditional) test theory by needing only to make changes to the trace line models themselves (Thissen, Nelson, Billeaud, & McLeod, 2001). Samejima's (1969) graded model is based on the logistic function giving the probability that an item response will be observed in *category k or higher*. For ordered responses $u = k$, $k = 1,2,3,$ ..., m, where response m reflects the highest θ value, the graded model trace line is:

$$T_i\left(u_i = k|\theta\right) = \frac{1}{1+\exp[-a_i(\theta - b_{ik})]} - \frac{1}{1+\exp[-a_i(\theta - b_{ik+1})]}$$

$$T\left(u = k|\theta\right) = T*(k|\theta) - T*(k+1|\theta)$$

(Thissen, Nelson, et al., 2001). The trace line models the probability of observing each response alternative as a function of the underlying construct (Steinberg & Thissen, 1995). The slope a_i varies by item *i*, but within an item, all response trace lines share the same slope (discrimination). This constraint of equal slope for responses within an item keeps trace lines from crossing, thus avoiding negative probabilities. The threshold parameters b_{ik} varies within an item with the constraint $b_{k-1} < b_k < b_{k+1}$. At each value $\theta = b_k$, the respondent has a 50% probability of endorsing the category. $T^*(k|\theta)$ is the trace line describing the probability that a respondent at any particular level of θ will respond in that scoring category or a higher category. The graded model trace line $T(u = k|\theta)$ represents the proportion of participants responding to that category across θ which will be a nonmonotonic curve, except for the first and last response categories (Thissen, Nelson, et al., 2001).

For the first response category $k = 1$, $T^*(1|\theta) = 1$; therefore, the trace line $T(u = 1|\theta)$ will have a monotonically decreasing logistic function with the lowest threshold parameter:

$$T_i(u_i = 1|\theta) = 1 - \frac{1}{1 + \exp[-a_i(\theta - b_{i2})]}.$$

(Thissen, Nelson et al., 2001). For the last response category $k = m$, $T^*(m+1|\theta) = 0$; therefore, the trace line $T(u = m|\theta)$ will have a monotonically increasing logistic function with the highest threshold parameter:

$$T_i(u_i = m|\theta) = \frac{1}{1 + \exp[-a_i(\theta - b_{im})]}$$

(Thissen, Nelson et al., 2001).

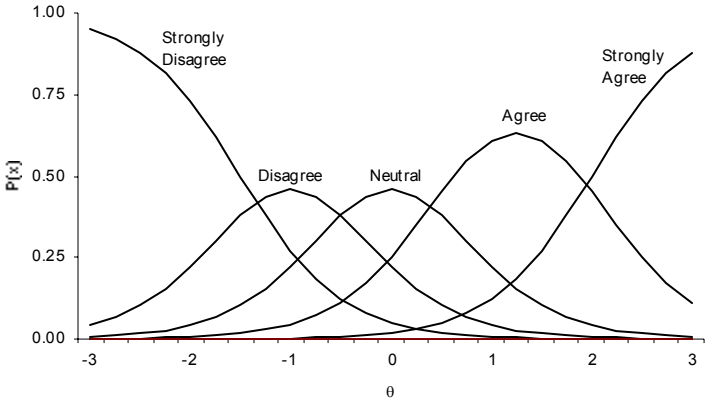


Figure 7 Graded Model (1 item with 5 response categories)

Figure 7 displays a graded item with five response categories. The model presents over what levels of the underlying trait θ a person is likely to endorse one of the response options.

The Nominal Model

Proposed by Bock (1972), the Nominal Model is an alternative to the Graded Model for polytomously scored items, not requiring any *a priori* specification of the order of the mutually exclusive response categories with respect to θ. The nominal model trace line for scores $u = 1, 2, \ldots, m_i$, for item i is:

$$T_i(u_i = x|\theta) = \frac{\exp[a_{ix}\theta + c_{ix}]}{\sum_{k=1}^{m} \exp[a_{ik}\theta + c_{ik}]}.$$

Where θ is the latent variable, a_ks are discrimination parameters, and c_ks are the intercepts. To identify the model, additional constraints are imposed. The sum of each set of parameters must equal zero, i.e.

$$\sum_{k=0}^{m-1} a_{ik} = \sum_{k=0}^{m-1} c_{ik} = 0$$

(Thissen, Nelson et al., 2001).

The Nominal Model is used when no pre-specified order can be determined among the response alternatives. In other words, the model allows one to determine which response alternative order is associated with higher levels on the underlying trait. This model has also been applied to determine the location of the neutral response in a Likert-type scale in relation to the ordered responses. Once the item order has been confirmed, the Graded IRT Model is often fit to the data.

The Partial Credit Model

For items with two or more ordered responses, Masters (1982) created the partial credit model within the Rasch model framework, and thus the model shares the desirable characteristics of the Rasch family: simple sum as a sufficient statistic for trait level measurement, and separate persons and item parameter estimation allowing specifically objective comparisons. The partial credit model contains two sets of location parameters, one for persons and one for items, on an underlying unidimensional construct (Masters & Wright, 1997).

The partial credit model is a simple adaptation of Rasch's model for dichotomies. The model follows that from the intended order $0 < 1 < 2, \ldots, < m$, of a set of categories, the conditional probability of scoring x rather than x -1 on an item should increase monotonically throughout the latent variable range. For the partial credit model, the expectation for person j scoring in category x over x -1 for item i is modeled:

$$\frac{\exp(\theta_j - \delta_{ix})}{1 + \exp(\theta_j - \delta_{ix})},$$

where δ_{ix} is an item parameter governing the probability of scoring x rather than x -1. The δ_{ix} parameter can be thought of as an item step difficulty associated with the location on the underlying trait where categories x-1 and x intersect. Rewriting the model, the response function

for the probability of person *j* scoring *x* on one of the possible outcomes 0, 1, 2, . . ., *m* of item *i* can be written:

$$T_i\left(u_i = x \middle| \theta_j\right) = \frac{\exp \sum_{k=0}^{x} (\theta_j - \delta_{ik})}{\sum_{h=0}^{m_j} \exp \sum_{k=0}^{h} (\theta_j - \delta_{ik})}, \quad x = 0, 1, \ldots, m_i,$$

(Masters & Wright, 1997). Thus, the probability of a respondent *j* endorsing category *x* for item *i* is a function of the difference between their level on the underlying trait and the step difficulty $(\theta - \delta)$.

Thissen, Nelson et al. (2001, p. 148) note the Partial Credit Model to be a constrained version of the Nominal Model in which, "not only are the *a*'s constrained to be linear functions of the category codes, all of those linear functions are constrained to have the same slope [for all the items]"

The Generalized Partial Credit Model is a generalization of the Partial Credit Model that allows the discrimination parameter to vary among the items.

The Rating Scale Model

The Rating Scale Model (Andrich, 1978a, 1978b) is another member of the Rasch family because the model retains the elegant measurement property of simple score sufficiency. The Rating Scale Model is derived from the Partial Credit Model with the same constraint of equal discrimination power across all items. The Rating Scale Model differs from the Partial Credit Model in that the distance between difficulty steps (or levels) from category to category within each item is the same across all items. The Rating Scale Model includes an additional parameter λ_i, which locates where the item *i* is on the underlying construct being measured by the scale. The response function for the unconditional probability of person *j* scoring *x* on one of the possible outcomes 0, 1, 2, . . ., *m* of item *i* can be written:

$$T_i\left(u_i = x\middle|\theta_j\right) = \frac{\exp\sum_{k=0}^{x}(\theta_j - (\lambda_i + \delta_k))}{\sum_{h=0}^{m_j}\exp\sum_{k=0}^{h}(\theta_j - (\lambda_i + \delta_k))}, \quad \text{where } \sum_{k=0}^{0}\left(\theta_j - (\lambda_i + \delta_k)\right) = 0$$

(Embretson & Reise, 2000). The constraint that a fixed set of rating points are used for the entire item set requires the item formats to be similar throughout the scale (e.g., all items have four response categories).

Trait Scoring

In classical test theory, scales are scored typically by summing the responses to the items. This summed score may then be linearly transformed to a scaled score estimate of a person's trait level. For example, if a respondent endorses 20 out of 50 items, he receives a score of 40%. On the other hand, item response theory uses the properties of the items (i.e., item discrimination, item difficulty) as well as knowledge of how item properties influence behavior (i.e., the item trace line) to estimate a person's trait score based on their responses to the items (Embretson & Reise, 2000).

IRT models are used to calculate a person's trait level by first estimating the likelihood of the pattern of responses to the items, given the level on the underlying trait being measured by the scale. Because the items are locally independent, the likelihood function *L* is

$$L = \prod_{i=1}^{nitems} T_i\left(u_i|\theta\right)$$

which is simply the product of the individual item trace lines $T_i\left(u_i|\theta\right)$. $T_i\left(u_i|\theta\right)$ models the probability of the response *u* to the item *i* conditional on the underlying trait θ. Often, information about the population is included in the estimation process along with the information of the item response patterns. Therefore, the likelihood function is a product of the IRT trace

lines for each individual item *i* multiplied by the population distribution of the latent construct $\phi(\theta)$:

$$L = \prod_{i=1}^{nitems} T_i(u_i|\theta)\ \phi(\theta).$$

Next, trait levels are estimated typically by a maximum likelihood method; specifically, the person's trait level maximizes the likelihood function given the item properties. Thus, a respondent's trait level is estimated by a process that 1) calculates the likelihood of a response pattern across the continuous levels of the underlying trait θ, and 2) uses some search method to find the trait level at the maximum of the likelihood (Embretson & Reise, 2000). Often times, this search method uses some form of the estimate of the mode (highest peak) or the average of the likelihood function. These estimates can be linearly transformed to have any mean and standard deviation a researcher may desire.

As an illustration of trait scoring, Figure 8 presents graphs of the population distribution of the underlying variable, four items' two-parameter logistic IRT trace lines, and the likelihood function for a person's response pattern of endorsing (i.e., true response) the first two items and not endorsing (i.e., false response) the last two items of a four item scale. In Figure 8a, the population distribution of θ is assumed to be normally distributed, and the scale is set to a mean of zero and variance of one. Other distributions can be used or no population information can be provided in the estimation process. However, the distributions of trait levels are often assumed to be normally distributed in the population. The trace lines for item 1 (a = 2.33, b = -0.14; see Figure 8b) and item 2 (a = 2.05, b = -0.02; see Figure 8c) represent the probability of a respondent endorsing the item given their level on the underlying trait. The trace lines for item 3 (a = 3.47, b = 0.26; see Figure 8d) and item 4 (a = 2.41, b = 1.27; see Figure 8e) represent the probability of a respondent not endorsing the item given their level on the underlying trait.

These trace lines for non-endorsement are represented by a monotonically decreasing curve to reflect that respondents high on the latent trait are less likely to respond false to (or not endorse) the items.

The likelihood function (see Figure 8f) describes the likelihood of a respondent's trait level given the population distribution of θ and the person's response pattern of endorsing the first two items and not endorsing the last two items, as represented by the following equation:

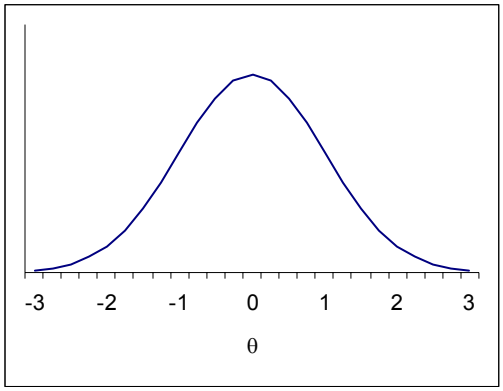
$$L = \phi(\theta) * T_1(u_1 = true|\theta) * T_2(u_2 = true|\theta) * T_3(u_3 = false|\theta) * T_4(u_4 = false|\theta)$$

The maximum of the likelihood (i.e., the respondent's trait level) for this response pattern is estimated to be $\hat{\theta} = 0.14$. The average of the distribution can also be used as an estimate of the respondent's trait level, $\hat{\theta} = 0.12$.

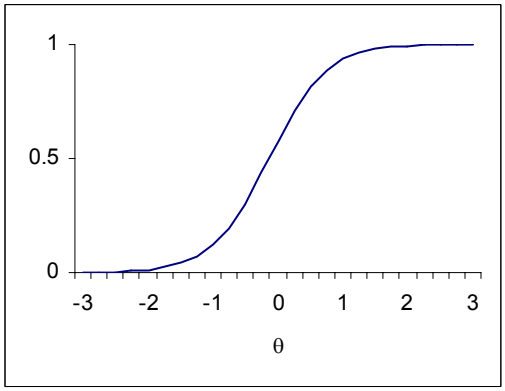
As another example, Figure 9 presents graphs of the population distribution, four item trace lines, and likelihood function for a respondent who endorses the first, second, and fourth item, but responds negatively to the third item. The maximum-likelihood estimate for this response pattern is $\hat{\theta} = 0.72$. As expected, this response pattern yields a higher estimated score because the respondent endorsed one additional item than the prior response pattern. The likelihood function for the second response pattern is smaller (has less area under the curve) than the likelihood function for the first response pattern. The decreased area reflects an inconsistency of item responses in the second pattern. The items are ordered by thresholds (i.e., item 1 is the least difficult and item 4 is the most difficult item to endorse; see *b* parameter estimates), therefore, one would expect that if a person endorses item four, then they should also endorse the first three items with lower difficulties. As an example, if these four items represent physical tasks of harder complexity from walking ten steps (item 1), walking to your mailbox (item 2), walking a block (item 3), to walking a mile (item 4), then you would expect someone

Figure 8

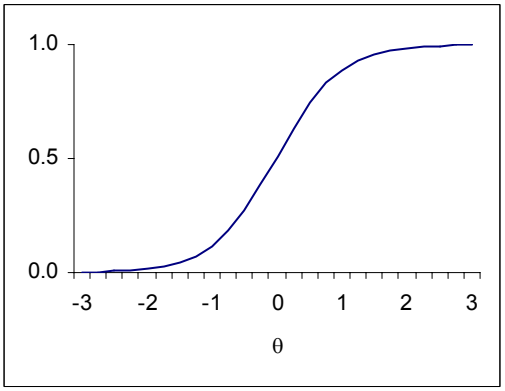
a) Population Distribution (Normal; mean = 0; variance = 1)



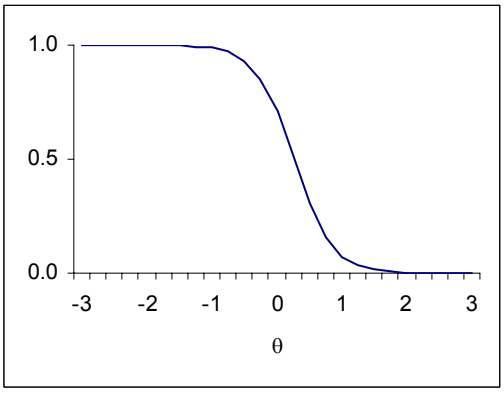
b) Item 1 (a = 2.33, b = -0.14) true response



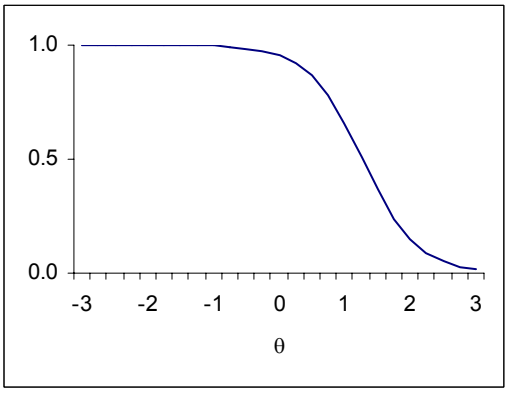
c) Item 2 (a = 2.05, b = -0.02) true response



d) Item 3 (a = 3.47, b = 0.26) false response



e) Item 4 (a = 2.41, b = 1.27) false response



f) Likelihood function (mode = 0.14, average = 0.12)

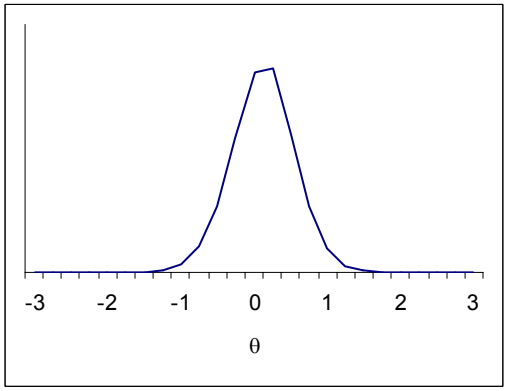
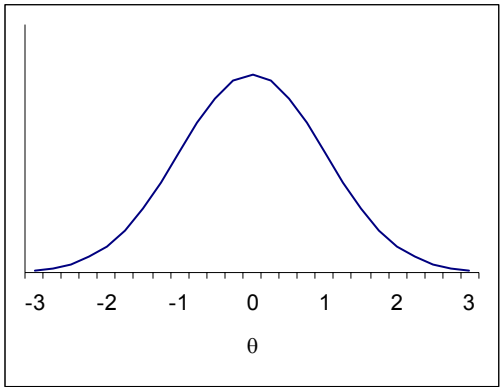
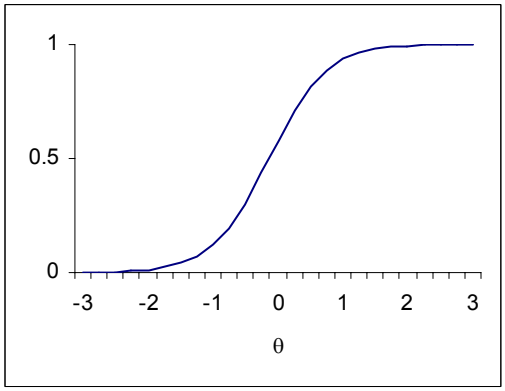


Figure 9

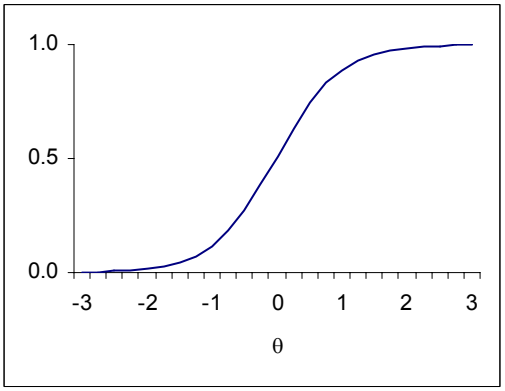
a) Population Distribution (Normal; mean = 0; variance = 1)



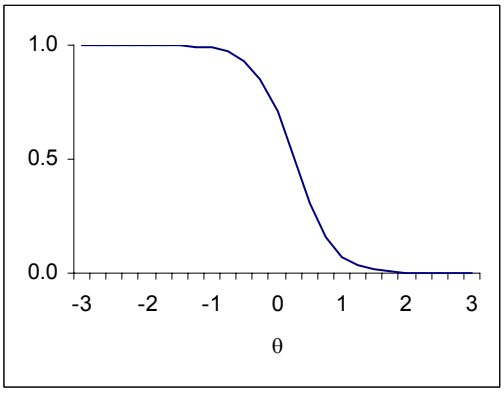
b) Item 1 (a = 2.33, b = -0.14) true response



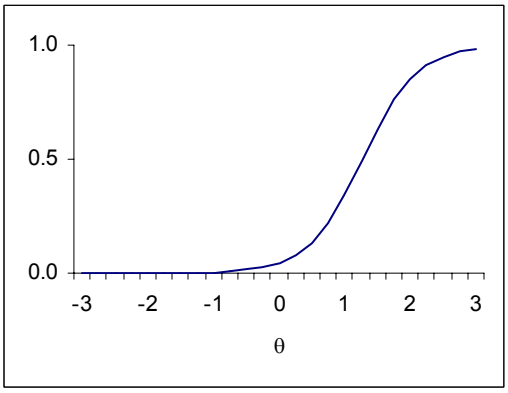
c) Item 2 (a = 2.05, b = -0.02) true response



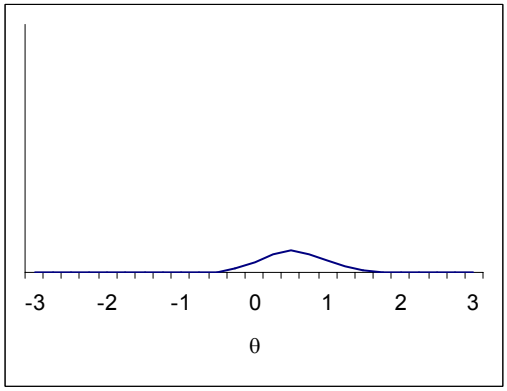
d) Item 3 (a = 3.47, b = 0.26) false response



e) Item 4 (a = 2.41, b = 1.27) true response



f) Likelihood function (mode = 0.72, average = 0.80)



who endorsed the item *walking a mile* to also endorse the item *walk a block*.

The one-parameter logistic model or Rasch model constrains the discrimination parameter, *a*, to be equal across the four items in the example. Therefore any response pattern with the same number of endorsed items will have the same estimated trait level. Knowledge of an individual's particular response pattern is not needed and, by the same token, information about that individual's trait level that might be derived from the response pattern is ignored. Thus, the total score is a sufficient statistic for estimating trait levels. Table 2 lists all possible response patterns for the four binary items ($2^4 = 16$ patterns), the summed score of the response pattern, and the associated maximum-likelihood estimates of the trait level calculated by using either the two-parameter logistic (2PL) IRT model or the Rasch model. The table shows that response patterns with two item endorsements (#6 to #11) are estimated by the Rasch model to have the same trait level ($\hat{\theta} = 0.22$). For these types of item response patterns with equal number of total responses, the 2PL IRT model will give higher estimates of the trait level for patterns with higher item thresholds (difficulties), and likewise, lower estimated trait levels for patterns with lower item thresholds. Thus, the 2PL IRT model will estimate a higher latent trait score for a person who gets two harder items correct than a person who endorses two easier items. Those who exclusively use the Rasch model view this property of the 2PL IRT model as a weakness. It is inconsistent for a person to endorse a harder item and not an easier item.

The Pearson correlation among the summed score, the Rasch model score, and the 2PL IRT model score, for the above example, are above .97. Despite the presented data is only an example, adding more items still yield correlations among the scores above .9. So the question is often asked, *Why not use the summed score, it is easier to compute?* Summed scores are on an ordinal scale that assumes the distance between any consecutive scores is equal. IRT model

Table 2

#	Item Response Pattern 0 = not endorse, 1 = endorse	Summed Score	2 PL IRT Model Maximum Likelihood Estimate	1 PL IRT / Rasch Model Maximum Likelihood Estimate
1	0 0 0 0	0	-0.82	-0.84
2	1 0 0 0	1	-0.27	-0.22
3	0 1 0 0	1	-0.21	-0.22
4	0 0 1 0	1	-0.19	-0.22
5	0 0 0 1	1	-0.01	-0.22
6	1 1 0 0	2	0.14	0.22
7	1 0 1 0	2	0.15	0.22
8	0 1 1 0	2	0.19	0.22
9	1 0 0 1	2	0.31	0.22
10	0 1 0 1	2	0.36	0.22
11	0 0 1 1	2	0.37	0.22
12	1 1 1 0	3	0.52	0.71
13	1 1 0 1	3	0.72	0.71
14	1 0 1 1	3	0.74	0.71
15	0 1 1 1	3	0.80	0.71
16	1 1 1 1	4	1.35	1.36

Note: The four items are ordered by difficulty with the last item having the highest threshold.

scores are on an interval scale and distance between scores vary depending on the difficulty (and sometimes discriminating power) of the question. For example, the difference in physical ability to endorse two items that ask if one can walk 10 feet and 20 feet is certainly different than the ability to endorse two items that ask if one can walk 100 feet and two miles. Ability scores should be close together for the first two items (walking 10 feet and 20 feet), and scores should be farther apart for answering the last two items (walking 100 feet and 2 miles).

Classical Test Theory and Item Response Theory

The past and most of the present research in health measurement has been grounded in classical test theory (CTT) models; however works by Embretson and Reise (2000) and Reise (1999) point out several advantages to moving to IRT modeling. Table 3 provides several key differences between CTT and IRT models.

Precision of measurement statistics such as standard error of measurement (SEM) and reliability indicate how well an instrument measures a single construct. The SEM describes an expected score fluctuation due to error in the measurement tool (Embretson and Reise, 2000).

Table 3

Classical Test Theory	Item Response Theory
Measures of precision fixed for all scores	Precision measures vary across scores
Longer scales increase reliability	Shorter, targeted scales can be equally reliable
Test properties are sample dependent	Test properties are sample free
Mixed item formats leads to unbalanced impact on total test scores	Easily handles mixed item formats.
Comparing respondents requires parallel scales	Different scales can be placed on a common metric
Summed scores are on ordinal scale	Scores on interval scale
	Graphical tools for item and scale analysis

Reliability is the fraction of observed score variance that is true score variance, or the proportion that is not error variance (Wainer & Thissen, in press). In CTT, both SEM and reliability (such as Cronbach's α, internal consistency) measures are fixed for all scale scores. In other words, CTT models assume that measurement error is distributed normally and equally for all score levels (Embretson & Reise, 2000). In IRT, measures of precision are estimated separately for each score level or response pattern, controlling for the characteristics (e.g., difficulty) of the items in the scale. Precision of measurement is best (low SEM, high reliability/information) typically in the middle of the scale range (or trait continuum), and precision is least at the low and high ends of the continuum where items do not discriminate well among respondents.

In CTT, scale reliability is a function of the number of items in the scale. Higher reliability requires longer scales. Many times, redundant or similar items are included in such instruments. In IRT, shorter and equally reliable scales can be developed with appropriate item placement. Redundant items are discouraged and actually violate the assumption of local independence of the IRT model. These short, reliable scales are often accomplished through the use of adaptive tests that chose a set of items that target in on a respondent's level on an underlying trait.

CTT scale measures such as reliability (Cronbach's α), item-total score correlation (point-biserial correlation), standard error of measurement, and difficulty (proportion) are sample dependent, meaning that, these measures vary across samples, especially for non-representative samples. IRT item properties are assumed to be sample-invariant within a linear transformation. This property of IRT makes the model very attractive for researchers investigating population differences.

Mixed item formats are surveys that include items scored as true/false, Likert-type/graded scales, or open-ended responses. In CTT, mixed item formats have an unbalanced impact on the total scale score. Items are unequally weighted leading to some items, with a high number of response options, to drive the survey score. Methods to correct for mixed item formats are limited because CTT's statistics are sample dependent. IRT has models for both dichotomously scored items (e.g., true/false), and polytomously scored questions (e.g., 5 category Likert-type scale). IRT item parameters are set to relate responses to the underlying trait (Embretson & Reise, 2000), thus, IRT can easily model the mixed item formats included in many surveys.

In health care research, there is a great need to compare respondents who take different surveys. CTT requires instruments to have a parallel form (e.g., equal means, variances, and covariances) to equate scores. This is difficult, almost impossible, to accomplish given the multitude of existing surveys in health research. Error in equating scores is influenced by any survey form differences (e.g., number of responses for each item, number of items). IRT models control for differences in item properties. Using a set of anchor items, IRT can place new items or items with different formats on a similar metric to link respondent scores. Once IRT item parameters have been estimated with an IRT model, investigators may calculate comparable

scores on a given construct for respondents from that population who did not answer the same questions, without intermediate equating steps (Orlando, Sherbourne, & Thissen, 2000).

A wonderful feature of IRT models are the graphical descriptions of item functioning and test functioning. IRT models allow you to graph the probability of a person endorsing an item given the person's level on the underlying trait (see trace lines in Figure 1 for binary data and Figure 7 for multiple-response data), the reliability or precision of the scale at all levels of the underlying construct (see information curves in Figures 2 and 3), response pattern scoring (see Figures 8 and 9), and group differences in responding to items (see Figure 15 and 16). These graphical tools are valuable for decision making, presentation of results, and allowing researchers unfamiliar with IRT to better understand its methodology.

Applications of Item Response Theory in Health Care Research

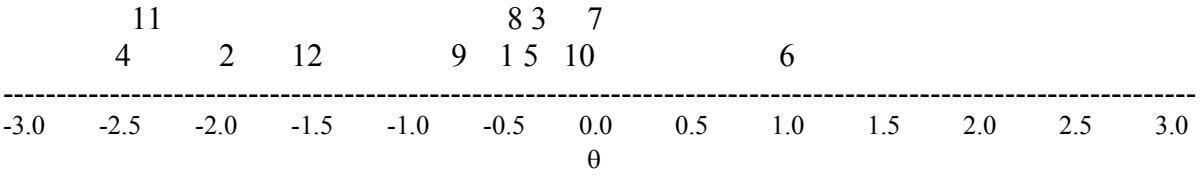
Item and Scale Analysis

Central to the development or evaluation of health outcomes measures is the analysis of the item's function within each scale. Item response theory provides a methodology to investigate each item's properties to assess what levels on the underlying trait is measured by the item (i.e., item threshold or location on theta) and the extent to which the item is related to the underlying construct measured by the instrument (i.e., item slope or discrimination). Knowledge of the individual item properties within a scale allows instrument developers to remove items that may be uncorrelated to the trait, or add items to measure respondents at levels on the underlying trait not addressed by the instrument. Another important tool in item analysis is the investigation of item bias or differential item functioning (DIF), in which one group responds differently to an item than another group after controlling for differences between the groups. A discussion of DIF follows this section.

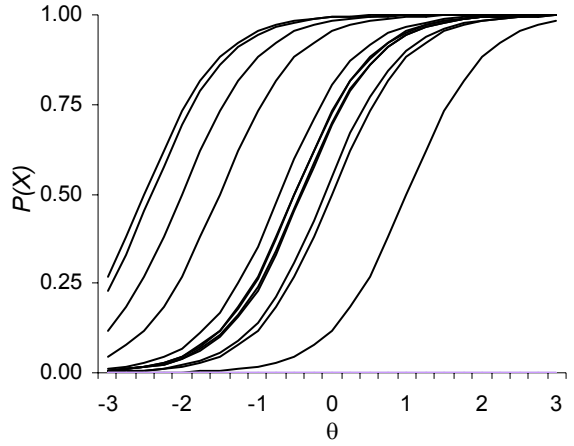
Information about the items' location on the underlying trait allows researchers the opportunity to identify trait levels not measured or overly measured by the instrument. After estimating item properties, there are many graphical resources to view the items' location along the underlying trait. One common and interpretable graphical tool to model the item locations for the Rasch model is an "item map". The Rasch or 1PL IRT model constrains the items' discrimination power (item slope) to be equivalent, and estimates the location of the items on the underlying trait θ. Figure 10-a presents an item map for twelve items numbered 1 – 12 above the horizontal axis line. Except for item #6, the items are located at the lower end of the latent scale, meaning that the items are measuring low levels of the trait θ being measured by the instrument. Items # 1, 3, 5, and 8 measure people at similar trait levels, suggesting that an instrument developer may wish to remove one or two of the items because they provide redundant information. Figure 10-b displays the item characteristic curves (trace lines) for the same set of twelve items. The Rasch model's constraint of equal item discriminations (i.e., equal slopes) is easily observed in this plot of parallel curves. Items # 1, 3, 5, and 8 are bunched together around theta levels of –0.47. For an overall view of the functioning of the twelve items in the scale, the test information function (see Figure 10-c) displays the precision of measurement of respondents at different levels of the underlying latent construct θ. A scale developer can easily observe that the instrument lacks information for measuring respondents with high levels of the trait θ.

The two-parameter logistic IRT model estimates an item's location along the continuous scale, but also estimates the discrimination power or an item's relationship with the underlying variable. Figure 11 presents item characteristic curves for twelve items with the same location (threshold) parameters as items in Figure 10, but with varying levels of discriminative power.

a.



b.



c.

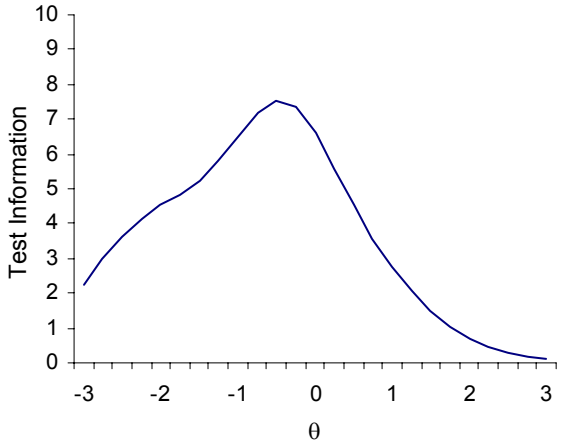


Figure 10 a. Item map for twelve items along the latent trait θ. b. Item characteristic curves for the same twelve items. c. Test information curve for the twelve-item scale.

For the redundant item set (# 1, 3, 5, and 8), items #5 and #8, identified by dotted lines, provide less precision in measurement of θ than items #1 and #3, and thus, may be removed from the scale.

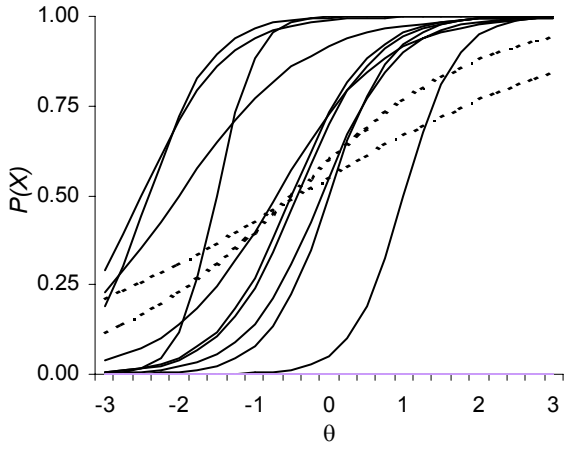


Figure 11 2PL IRT model trace lines for twelve items.

Placement of items within a scale may differ depending on the purpose of the instrument. A diagnostic instrument, with the intent to identify respondents high on a trait for purposes of treatment or to divide respondents into one of several groups along a continuous measure, will require items with location parameter values at the cut-off points on the scale. Thus, accuracy in person measurement is most precise at the category thresholds. However, instrument developers wishing to measure respondents at all levels of the underlying trait will choose items that are evenly distributed across the continuum of theta. A helpful tool for assessing areas on the underlying trait measured by an instrument is the test information curve (see Figure 10-c). Diagnostic instruments should have high information (i.e., precision in measurement) at the cut-off levels on theta and general instruments should have a high uniform distribution of information across theta.

To obtain the measurement properties of the Rasch family of models, instrument developers must carefully choose an item set with equal discriminatory power across all items. That is, each item is equally related to the underlying trait measured by the scale. The software programs that fit the Rasch model (as well as any other IRT model) will provide a series of item fit statistics or residuals analysis. For the Rasch model, items with extreme discrimination power both at the low as well as the high values will be identified as misfit and will be removed from the scale.

Item analysis also occurs before an IRT model is fit to the data. Classical test theory statistics provide important information such as item-scale correlations (e.g., point-biserial correlation) and item difficulties (e.g., proportion correct). As a test for the assumption of unidimensionality for the IRT model, factor analyses will provide the magnitude to which an item is related to the construct measured by the instrument. Investigation of the assumption of

item local independence will reveal redundant items. Redundant items essentially ask the same question of the respondent. Such locally dependent items typically have similarly-worded formats or sentence stems.

Together, the applications of both classical test theory and item response theory provide instrument developers the tools to understand how each of the items function within a scale. IRT can increase scale precision with fewer items by the selection of items that are highly related to the underlying dimensions (Steinberg & Thissen, 1995). Also, developers can create shorter scales by choosing questions (items) that target the population of interest to measure. Shorter instruments with high reliability are attractive for both instrument responders and graders.

Differential Item Functioning

Differential item functioning (DIF) is a condition when an item functions differently for respondents from one group to another. In other words, respondents, with similar levels on a latent trait θ (e.g., physical functioning) but who belong to different populations, have a different probability of responding to an item. DIF items are a serious threat to the validity of the instruments to measure the trait levels of members from different populations or groups. Instruments containing such items may have reduced validity for between-group comparisons, because their scores may be indicative of a variety of attributes other than those the scale is intended to measure (Thissen, Steinberg, & Wainer, 1988).

In the past, differential item functioning has been called "item bias" in the literature because such an item biases one group to have a higher scale score than another group. For example, the item “I cry easily” on the Minnesotta Multiphasic Personality Inventory (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) unfavorably *biases* females to have higher depression scores than males who as a group are culturally condoned not to display such

emotions. Item bias indicates that groups are treated unfairly; however, Angoff (1993) cautions that differential item functioning can occur without judgements of unfairness to one group or another. For example, when two different national groups were compared on a common measure as part of a research effort to study linguistic differences (e.g., Alderman & Holland, 1981), researchers were not interested in item bias, but were investigating cultural differences. Therefore, the term "item bias" has either been replaced by the expression "differential item functioning" or the terms are used interchangeably in the current literature.

DIF is most frequently defined in the context of item response theory (IRT). Item trace lines provide a means for comparing the responses of two different groups, say groups *reference* (e.g., control) and *focal* (e.g., treatment), to the same item. In the framework of IRT, item parameters are assumed to be invariant to group membership (in contrast to classical test theory where parameter estimates and statistics vary with the sample being measured). Therefore, differences between the trace lines, estimated separately for each group, indicates that respondents from the reference group and focal group at the same level of the underlying trait have different probabilities of endorsing the item. More precisely, DIF is said to occur whenever the conditional probability, *P(X)*, of a correct response or endorsement of the item for the same level on the latent variable differs for two groups (Camilli & Shepard, 1994).

The key decision that must be made for DIF analysis is the selection of the appropriate IRT model (Camilli & Shepard, 1994). Different models allow a different number of item parameters (i.e., *b*, *a*, *c* parameters) to be estimated from the data of item responses, and thus, allow for the evaluation of DIF for different item properties.

Analyses based on the one-parameter logistic IRT model or the Rasch dichotomous model investigate DIF in the threshold or location parameter *b*. In other words, does a reference

and focal group have different item endorsement rates after controlling for group differences on the latent variable being measured by the instrument. Figure 12 presents two item trace curves estimated separately for each group, with the focal group having the higher threshold level. This DIF example shows that, over all the levels on the underlying variable θ, the reference group has a higher probability *P(X)* of endorsing the item than the focal group. That is, the item is easier for the reference group to answer than the focal group.

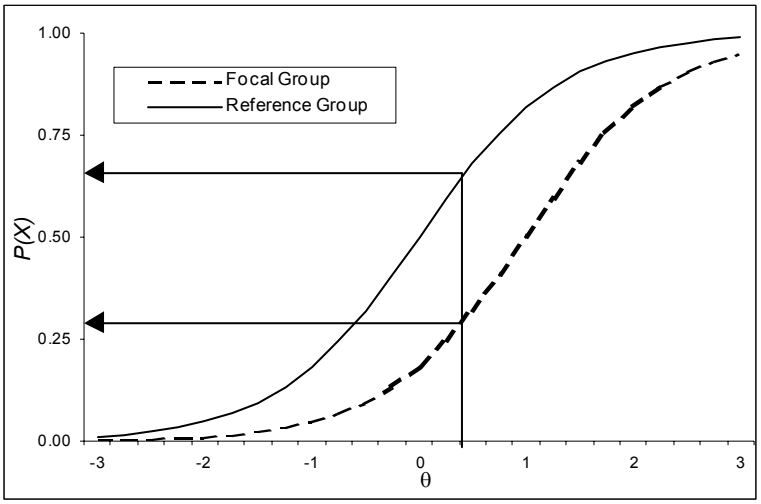


Figure 12. One-parameter logistic (or Rasch) model trace lines for reference ($b = 0$) and focal groups ($b = 1$). At equal levels of the underlying variable, respondents from the reference group are estimated to have a higher probability of endorsing the item than respondents from the focal group.

For binary data, the two-parameter logistic IRT model allows for the investigation of DIF between groups in the item's threshold parameter *b*, slope parameter *a*, or both parameters. DIF in the slope parameter represents an interaction between the underlying measured variable and group membership (Teresi, Kleinman, & Ocepek-Welikson, 2000). The degree to which an item relates to the underlying construct depends on the group being measured. Figure 13 presents the reference and focal groups' two-parameter logistic IRT trace lines for the same item. DIF analyses suggest that the item is more endorsed (or answered correctly) by the reference group

(i.e., lower *b* parameter) than the focal group, and the item is more discriminating among respondents in the reference group (i.e., higher *a* parameter) than the focal group.

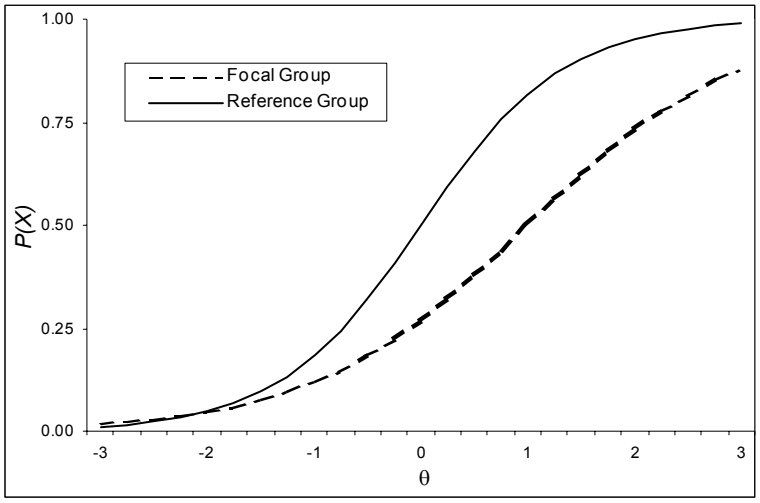


Figure 13. Two-parameter logistic IRT model trace lines for the focal group (a = 1.5, b = 0) and reference group (a = 1.0, b = 1.0).

Like the two-parameter IRT model, the three-parameter logistic IRT model allows for the investigation of DIF in the discrimination parameter and threshold parameter, but also evaluates DIF in the pseudo-guessing parameter. Figure 14 presents three-parameter logistic IRT curves for the reference and focal groups. DIF in the pseudo-guessing parameter indicates that groups

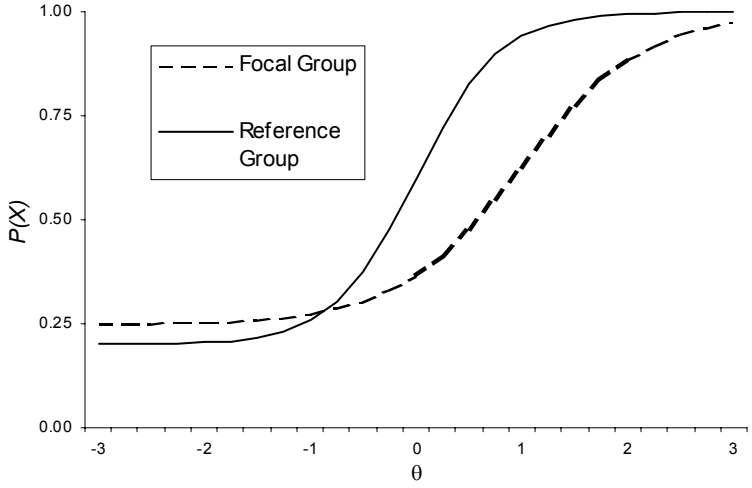


Figure 14. Three-parameter logistic IRT curves for the reference group (a = 1.5, b = 0, c = .20) and focal group (a = 1.0, b = 1.0, c = .25).

are differently affected by some confounding variable that explains why respondents low on the underlying variable measured by the instrument may endorse the item.

Rasch modelers investigate differential item functioning only in the threshold (location) parameter. This methodology has strict requirements to maintain the elegance of the Rasch model (e.g., sum score sufficiency). Any item that differs in its ability to discriminate among respondents compared to other items in a measure is considered a misfitting item to the Rasch model (Smith, 1991). Thus, if an item has different estimated slopes (i.e., discrimination ability) between the reference and focal groups, the item is not considered biased, but is considered misfit and is removed from the measure.

As an illustration of the application of DIF analysis in the framework of Rasch measurement, Tennant (2000, June) investigated differences in responding patterns between respondents from the United Kingdom and Italy on items in the quality of life measurement scale LMSQoL. Study results found significant differential item ordering (i.e., different threshold parameter estimates) between the two groups of respondents. Figure 15 presents the LMSQoL scale items ordered by their estimated threshold parameters for respondents from Italy and the UK. Item ordering represents differences in groups' attitudes towards quality of life for their culture. Tennant's (2000, June) results suggest a strong need to investigate DIF for all health measures to create a core set of outcome measures that can be used across different cultures.

Many researchers (e.g., Angoff, 1993; Camilli & Shepard, 1994) believe investigation of DIF in the framework of Rasch measurement is limited. Non-consideration of the possible differences in respect to discrimination power or differences in respect to pseudo-guessing will result in undetected DIF items and will lead to the ultimate removal of the most useful items in a measure (Angoff, 1993). Therefore, applications of the Rasch models limit researchers'

understanding of group differences in responding to items in a measure. IRT models that allow the level of discrimination to vary from item to item describe the data more accurately than models that constrain the slope parameter to be equal across items.

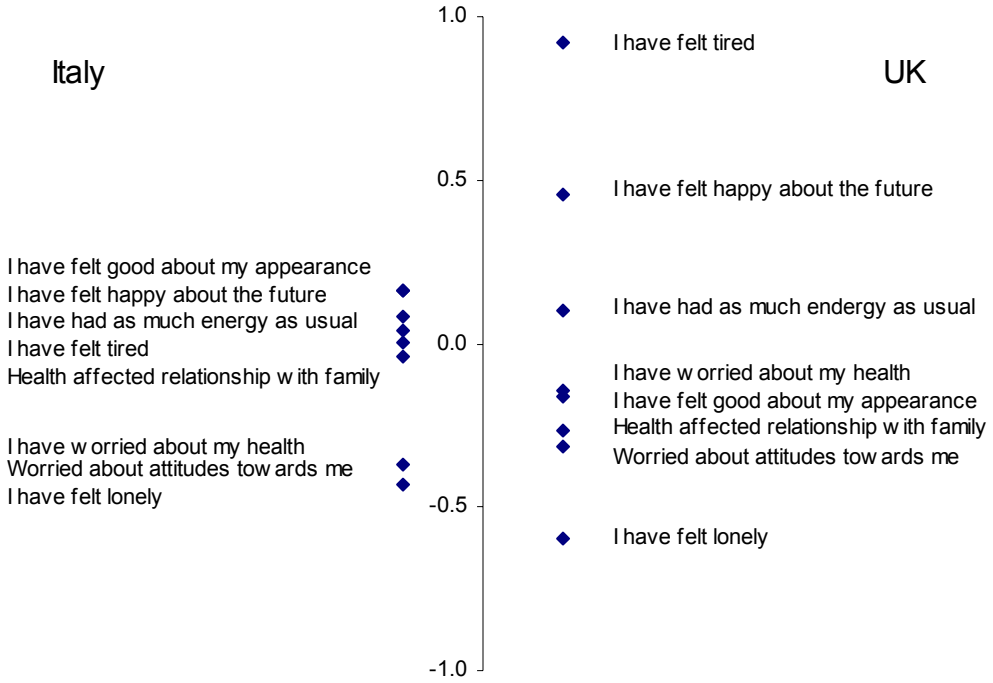


Figure 15. Hierarchical ordering of items in LMSQoL scale for respondents from Italy and UK. Location values are threshold parameter estimates.

To illustrate the use of the two-parameter logistic model, Reeve (2000) investigated DIF between non-clinical and clinical[1] responses to the Physical Malfunctioning Subscale of the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, et. al., 1989). DIF analyses revealed differential functioning between clinical and non-clinical respondents for the item: *During the past few years, I have been well most of the time*. Figure 16 shows that the item has a higher level of discrimination power for the non-clinical respondents than the clinical respondents; that is, the item has a higher correlation with the construct *physical malfunctioning* for the non-clinical group. DIF results suggest that the two groups' definition of "well" is

differently perceived. It may be that the non-clinical respondents interpret the item to mean "well" in the physical sense, where as the clinical respondents may interpret the item to

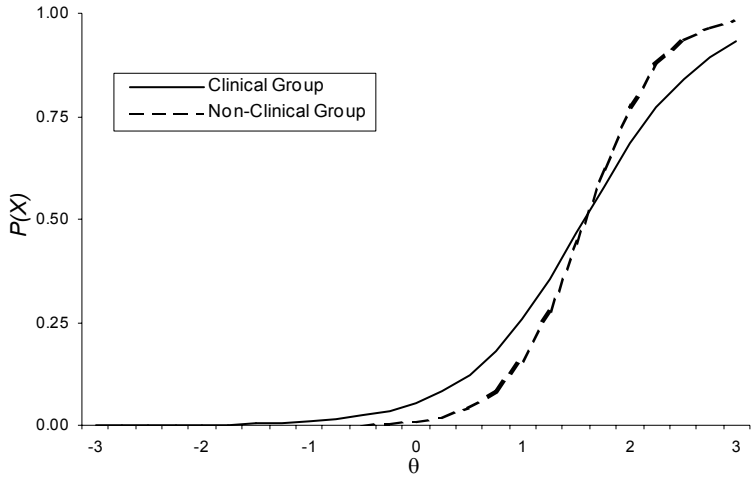


Figure 16. Two-parameter logistic trace lines for the item *During the past few years, I have been well most of the time* modeled for a clinical (a = 1.82, b = 1.58) and non-clinical (a = 2.91, b = 1.58) group of respondents to the MMPI-2.

mean "well" in the mental sense. Because the threshold parameter of this item is found to be invariant across the two study groups, applications of the Rasch model would not identify DIF and therefore would likely not discover the discrepancy between the groups' understanding of the item content. The Rasch model may find the item to misfit in one group and not the other, possibly leading to further investigation.

As a goal for any researcher to create a core set of endpoint measures that could be used in a variety of trial-based and observational studies and that is invariant to sub-groups, it is necessary to investigate the psychometric properties of the items within the scales. Essential to this investigation is the analysis of differential item functioning to observe if a group such as cancer patients interprets or treats the item content of a measure in the same manner as a control group. DIF analysis based on the IRT model can not only provide identification and removal of

[1] The clinical sample is a group of 2,672 females from both inpatient and outpatient mental services collected across

biased items, but also provide aid to interpreting psychological or physiological processes underlying group differences in responding to particular items.

Instrument Equating and Computerized Adaptive Tests

With the large number of health care measures for an even larger set of outcome variables, there is a great need to have some standardization of the concepts and metrics to allow comparisons of instruments and respondents administered those instruments (Ware, Bjorner, & Kosinski, 2000). The goal to either develop or identify a core set of health care measures (or outcome domains) can be accomplished by linking scales together or creating an item pool to place all items of a similar construct on a common metric. The applications of IRT are ideal for such a methodology because of the property of item invariance. Items calibrated by the appropriate IRT model(s) are linked together on a common metric allowing the creation of questionnaires that may use a different set of items depending on the target audience of responders. In other words, two responders administered two different assessments (i.e., sets of questions) can have scores comparable on a similar metric.

The goal of instrument linking or equating is not to change the content, but to adjust for the differences in difficulties of the measures (McHorney & Cohen, 2000). That is, the scales are measuring the same construct, but tap into different levels of the underlying construct. Several methodologies exist for instrument linking. Ideally, one would administer both instruments to a sample and calibrate (obtain the properties of) the items simultaneously. However, this method can be a burden on respondents especially for long scales. As an alternative, a set of items that are common to both instruments can be selected as anchors. The anchor items are used to set the metrics to which items not common to both instruments are

---

the United States from 1990 to 1997.

scaled. Therefore, instruments with a different number or difficulty of items can be linked by responses to a common set of anchor items.

As an example, McHorney and Cohen (2000) linked items that measure overall physical functioning from scales that focus on a respondent's activities of daily living (ADLs: independent human functioning) and instrumental activities of daily living (IADLs: domestic and community independence). Pooling from multiple ADL and IADL measures yielded 1,588 items. After consolidating similar items, the researchers were left with 206 items that cut across several domains of physical functioning (toileting, bathing, cooking/eating, dressing, mobility, and household and community activities). To reduce respondent burden, three instrument forms were created that each contains a set of 89 common items (anchors) and 39 unique items in each form. The items were calibrated using the IRT graded response model for the six-category response format. The success of the study is indicated by the ability to compare respondent scores on any subset of items taken from the 206-item bank. Using item response theory, McHorney and Cohen (2000) were also able to identify deficiencies within each of the domains as it pertains to the precision or accuracy of measurement.

The new direction of modern measurement is in the applications of computerized adaptive testing (CAT). IRT, with its properties of item invariance and group invariance, make this technology feasible. To develop a CAT, a subset of items are chosen from an item pool that experts agree to best measure the construct(s) of interest. This item bank is calibrated using the appropriate IRT model based on responses to the item set. Once an item bank has been created and calibrated, a computer then proceeds through an algorithm of choosing an item most appropriate for the trait level of a respondent, estimating the respondent's trait level based on their response to the item, and then choosing the next most appropriate item from the bank until

some precision in measurement is met. The benefits of a CAT methodology to measurement is the reduction in respondent burden to answering many items without a loss in reliability or precision in measurement.

Ware, Bjorner, and Kosinski (2000) developed an online CAT for measuring the impact of headaches. With an item bank of 53 items that cover the range of headache suffering (θ), study findings indicate as many as 98 percent of respondents in a controlled simulation test and 70 percent of respondents in an internet pilot test require five or fewer items to obtain an accurate measure of their suffering. Validation tests of the CAT instrument found high correlations with three other headache instruments and a moderate correlation with one other instrument.

Conclusion

There is a great need in the health care industry to have tools to monitor population health on a large scale as well as more precise tools to identify those who need and are most likely to benefit from treatment (Ware, Bjorner, & Kosinski, 2000). The applications of item response theory modeling can help to create these tools. Item and scale analysis within the framework of IRT will ensure reliable, valid, and accurate measurement of respondent trait levels. Identification of items that are informative or problematic help investigators to understand the domains they are measuring as well as the populations they measure.

Furthermore, there is a need in the health care industry to standardize the concepts and metrics of healthcare measurement to allow comparisons of results across assessment tools and across diverse populations (Ware, Bjorner, & Kosinski, 2000). In a reference to measures of physical functioning but that applies to all health outcome instruments, McHorney and Cohen (2000) find that the vast amount of measures differ in their source of items, breadth and depth of

measurement, item difficulty, response format, scaling method, and psychometric rigor. Because of the limitations with classical test theory in regards to sample and test dependencies, these variations make it nearly impossible to compare respondent scores across different measures. However, researchers such as McHorney and Cohen (2000) demonstrat that linking instruments within the framework of IRT modeling can allow comparisons of instruments and respondents.

Item banking is one method that will place multiple measures on a common metric to allow cross-walking of scores. From the item bank, any number of instruments can be tailor-made to measure the population of interest without the worry of score comparability with other groups that may be taking an alternative assessment developed from the same item bank. On top of that, item banking allows for the development of computerized adaptive tests that reduce respondent burden and increases reliable measurement by using a methodology that targets in on a respondent's true score.

So, why are the methodologies of item response theory slow to be adopted into the health care measurement field? Item response theory was developed within the framework of educational testing and so most of the literature and terminology is oriented towards that discipline (Hambleton & Swaminathan, 1985). Some researchers (e.g., Zickar & Drasgow, 1996) also argue that while ability testing, such as verbal or mathematical skills, in educational measurement can be conceptualized, it is unclear if a latent trait such as quality of life or patient satisfaction can be modeled in the same manner. Another limitation of the modern measurement theory is the complexities of the mathematical IRT models. Most researchers have been trained in classical test theory and are comfortable with reporting statistics such as summed scale scores, proportions correct, and Cronbach's alpha. Beyond the mathematical formulas, there are the complexities of the numerous IRT models themselves as to what circumstances are appropriate

to use IRT and which model to use. There is not even a consensus among psychometricians as to the definition of measurement and which IRT models fit that definition. Adding to the burden of confusion, the numerous available IRT software in the market are not user-friendly and often yield different results (parameter and trait estimates) because of the different estimation processes used by the software.

Despite these limitations, the practical applications of IRT cannot be ignored. Knowledge of IRT is spreading as more and more classes are being taught within the university disciplines of psychology, education, and public health, and at seminars and conferences throughout the world. Along with this, more books and tutorials are being written on the subject as well as more user-friendly software is being developed. Research applying IRT models are appearing more frequently in health care journals, and much of their concluding comments is directed towards discussing the benefits and limitations of using the methodology in this field. Together, a better understanding of the models and applications of IRT will emerge and IRT will be as commonly used as the methodology of classical test theory. This effort will result in instruments that are shorter, reliable, and targeted towards the population of interest.

One further note is that item response theory is only one step towards the goal of the creation of reliable and valid health care measures. Hambleton (2000, p. 63) states quite well that IRT is not "the solution to all of our instrument and measurement problems. It is a mathematical model only, and when it can be demonstrated that (1) the model fits the data of interest, (2) model parameters are properly estimated, and (3) the model is used correctly, the model has many useful features. But, *none of the IRT models* [paraphrased] are magic wands to wave over vague instrument specifications and poorly constructed items to make reliable and valid measurements. Hard and thoughtful work is still required in defining constructs and related

domains of content, drafting items to measure the constructs, field testing, test norming, and conducting reliability and validity studies…If these steps are not handled well, bad measurements will follow."

References

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language* (Research Rep. No. 81-16). Princeton, NJ: Educational Testing Service.

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

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du nieveau intellectuel anormoux {New methods for the diagnosis of levels of intellectual abnormality]. *Annee Psychologique, 11*, 191-244.

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

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*, 443-459.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Camilli, G., & Shepard, L. A. (1994). *MMSS Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist - Revised. *Psychological Assessment, 9*, 3-14.

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Ferguson, G. A. (1943). Item selection by the constant process. *Psychometrika, 7*, 19-29.

Fischer, G. (1968). *Psychologische test theorie* [Psychological test theory]. Bern: Huber.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Journal of Research in Personality*, *29*, 168-188.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38* (9 Supplement), 60-65.

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

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 Supplement), 28-42.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 62-A, Part 1, 74-82.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and Prediction* (pps. 362-412). New York: Wiley.

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

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, Whole No. 7.

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

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. (1997). 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-Verlag Inc.

McHorney, C. A., & Cohen, A. S. (2000). Equating health status measures with item response theory: Illustrations with functional status items. *Medical Care, 38* (9 Supplement), 43-59.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*, 223-236.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: application to depression measurement. *Psychological Assessment, 12*, 354-359.

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

Reeve, B. B. (2000). *Item- and scale-level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219-242). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika, 1*, 33-49.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 34*, (4, Pt.2, Whole No. 17).

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: applications of item response models. *Psychological Assessment, 10*, 345-359.

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. In C. C. Clogg (Ed.), *Sociological Methodology*, (Vol. 18, pp. 271-307). Washington DC: American Sociological Association.

Smith, R. M. (1991). *IPARM: Item and person analysis with the Rasch model*. Chicago, IL: Mesa Press.

Steinberg, L., & Thissen, D. (in draft). Chapter 1 – An intellectual history of item response theory: Models for binary responses," from a draft of *Item response theory for psychological research*.

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: a festschrift honoring Donald W. Fiske* (pp. 161-181). Hilldale, NJ: Erlbaum.

Tennant, A. (2000, June). Rasch perspectives on cross cultural validity. *Measurement of healthcare outcomes*. Symposium conducted at the Third International Outcomes Measure Conference, Chicago, IL.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651-1683.

Thissen, D. (1991). *Multilog User's Guide. Multiple, categorical item analysis and test scoring using Item Response Theory*, Version 6.3. Chicago, IL: Scientific Software International.

Thissen, D., Nelson, L., Billeaud, K. & McLeod, L. (2001). Chapter 4-Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D. & Orlando, M. (2001). Chapter 3-Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring*. Hillsdale, NJ: Erlbaum.

Thissen, D. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433-451.

Tinsley, H. E. A. (1992). Psychometric theory and counseling psychology research. In S. D. Brown & R. W. Lent (Eds.), *Handbook of counseling psychology* (2nd ed., pp. 37-70). New York, NY: Wiley.

Wainer, H., & Thissen, D. (2001). Chapter 2 – True score theory. In D. Thissen & H. Wainer (Eds.), *Test Scoring*. Hillsdale, NJ: Erlbaum.

Ware, J. E., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Medical Care*, 38 (9 Supplement), 73-83.

Weiss, D. J. (1995). Improving individual differences measurement with item response theory and computerized adaptive testing. In D. J. Lubinski, & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 49-79). Palo Alto, CA: Davies-Black Publishing.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*, 21-33.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

Illustration of IRT Modeling

Scale

The ten items listed in Table A-1 make up the Harris-Lingoes Brooding Subscale of the Minnesotta Multiphasic Personality Inventory (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Respondents who score high on the brooding subscale appear to "lack energy to cope with problems and may have concluded that life is no longer worthwhile. They brood, cry, and ruminate, and they may feel they are losing control of their thought processes." (Butcher, et al., p. 45, 1989).

Table A-1 Harris-Lingoes Brooding Subscale

Item Number	Content	Response
1	Periods when I couldn't "get going."	**true** false
2	I wish I could be as happy as others.	**true** false
3	I don't seem to care what happens to me.	**true** false
4	Criticism or scolding hurts me terribly.	**true** false
5	I certainly feel useless at times.	**true** false
6	I cry easily.	**true** false
7	I am afraid of losing my mind.	**true** false
8	I brood a great deal.	**true** false
9	I usually feel that life is worthwhile.	true **false**
10	I am happy most of the time.	true **false**

Note: Responses in bold typeface represent higher levels of brooding.

The developers selected items for the MMPI using an empirical-keying approach (Hathaway & McKinley, 1983). Items were not chosen because of their theoretical import or content, but because they maximally discriminate a target group (e.g., a clinically-depressed population) from a control group (e.g., a non-clinical population) (Steinberg & Thissen, 1995). Because this self-report instrument was originally intended as a diagnostic tool, item functioning (IRT location parameter *b*) in this scale should occur in the high levels of the brooding trait.

Sample

The data are responses to the depression scales of the MMPI-2 sampled from a clinical population. The clinical sample contains responses from 26,397 females and 26,109 males from both inpatient and outpatient services collected across the nation from 1990 to 1997.[2] For the

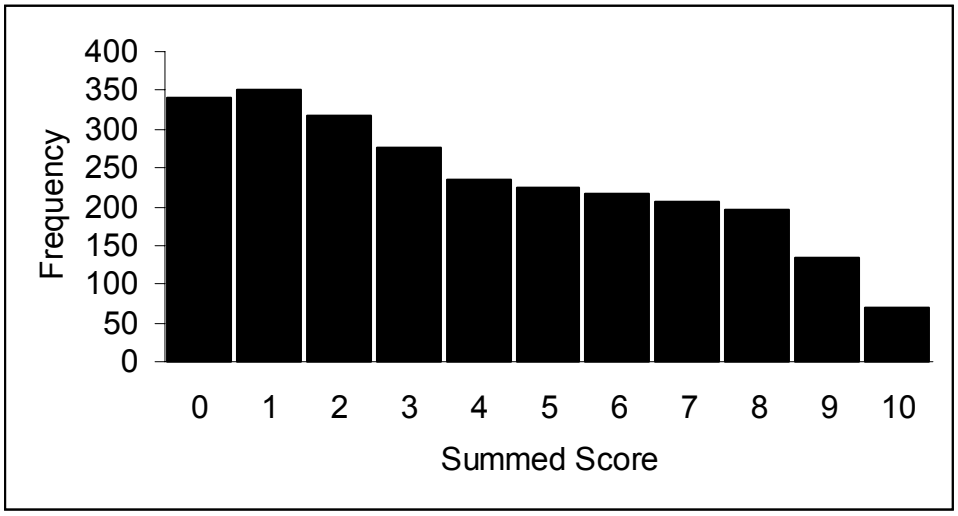
[2] The data is provided by the Caldwell-Greene archive. The data are kept in the L. L. Thurstone Psychometric Laboratory and maintained by Reeve, Hancock, Dahlstrom, and Panter.

IRT analyses, ten percent random subsamples of 2,569 females and 2,698 males were drawn from the full sample to reduce computational burden. This illustration will focus on responses from the clinical female sample, but will also discuss results from the analyses on the clinical male sample responses in the investigation of differential item functioning.

Classical Test Theory Analysis of Data

The items in the data set were formatted so that a score of 0 indicates the respondent answered the item in a manner consistent with lower levels of brooding and a score of 1 indicates the respondent answered the item in a manner consistent with higher levels of brooding. The average summed score for the females in the clinical sample is 3.94 and a standard deviation of 2.94. The histogram in Figure A-1 displays the frequency of all summed scores among the

Figure A-1 Histogram of the Brooding Scale Summed Score of the Clinical Females



2,569 females in the clinical group. The sample distribution of summed scores has a slight negative skew, but respondents are well represented across the entire range of possible scores. A measure of reliability, the Kuder-Richardson KR-20 estimate of internal consistency (i.e., coefficient alpha) for the ten item subscale, is 0.83.

Table A-2 lists descriptive statistics of the individual items in the brooding subscale. The column titled "average (when item endorsed)" shows the average summed score for those respondents who endorsed the item in a direction consistent with higher levels of brooding. Larger differences between this score and the scale average of 3.94 indicate the respondent is likely to endorse more of the other items in the scale; thus, the respondent reports more "brooding" behaviors when they endorse the item. The proportion of the 2,569 females in the

Table A-2 Classical Test Theory Item Analyses

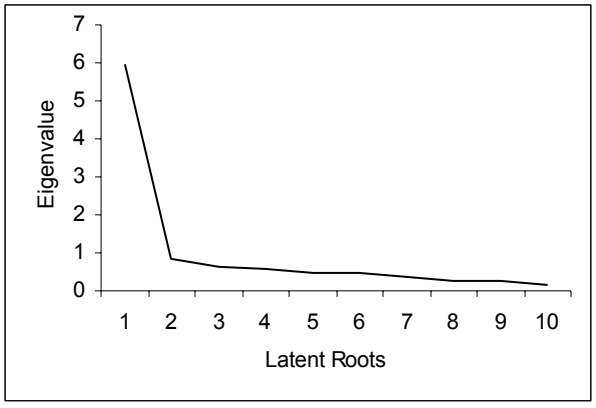
#	Item	ten item sum scale Average	Average (when item endorsed)	proportion endorsed (n = 2569)	point-biserial correlation (item-scale)
1	Periods when I couldn't "get going"	3.94	5.92	0.501	0.675
2	I wish I could be as happy as others	3.94	5.87	0.540	0.713
3	I don't seem to care what happens to me	3.94	7.85	0.151	0.563
4	Criticism or scolding hurts me terribly	3.94	5.34	0.552	0.530
5	I certainly feel useless at times	3.94	6.03	0.504	0.720
6	I cry easily	3.94	5.41	0.546	0.548
7	I am afraid of losing my mind	3.94	6.77	0.296	0.626
8	I brood a great deal	3.94	7.13	0.248	0.624
9	I usually feel that life is worthwhile	3.94	7.41	0.186	0.565
10	I am happy most of the time	3.94	6.56	0.413	0.748

clinical sample who endorsed the item in a direction consistent with higher levels of brooding is provided in the column marked "proportion endorsed". The proportion of respondents endorsing an item is an indication of item difficulty or the amount of the trait measured by the test that must be needed to endorse the item. In the last column, the point-biserial correlation is a measure of the relationship between the item endorsement and scale score. Items 4 (*Criticism or scolding hurts me terribly*) and 6 (*I cry easily*) are least related to the total scale score and items 5 (*I certainly feel useless at times*) and 10 (*I am happy most of the time*) have the highest relationship with the total scale score.

Modern Measurement Theory

Before an IRT model can be fit to the data, the assumptions of unidimensionality and local independence must be assessed. Analysis of the factor structure within a scale is often carried out by researchers using the classical test theory methods. However, factor analysis of binary or graded data carried out by traditional methods are problematic due to the violation of non-normality and non-interval measure of the data. To overcome these problems, a full information factor analysis was performed on the data using a multidimensional IRT model to model the factor structure. Results from the full-information factor analysis suggest that the ten-item brooding subscale is approximately unidimensional. The scree plot in Figure A-2 is one methodology for determining the number of factors to retain in an exploratory analysis, and presents the amount of variation (eigenvalue) explained by each factor (latent root) extracted from the covariation among item responses. Examination of the scree plot reveals a first factor accounting for much of the variability (59%) observed among the response variables.

Figure A-2 Scree Plot



The assumption of local independence asserts that responses to an item are independent of responses to another item once controlling for the underlying variable measured by the scale. A follow-up analysis of the assumption of local independence within the Brooding subscale identified highly correlated response patterns among two-item pairs. Results suggest items 4 (*Criticism or scolding hurts me terribly*) and 6 (*I cry easily*), and items 3 (*I don't seem to care*

what happens to me) and 9 (*I usually feel that life is worthwhile*) to be locally dependent. Solutions to remove local dependence in the subscale will be reviewed later in the discussion of recommended scale modifications.

Choosing the Appropriate Item Response Theory Model

Because of the binary (dichotomous) response format of the items, either the one-parameter logistic IRT (Rasch) model, the two-parameter logistic (2PL) IRT model, or the three-parameter logistic (3PL) IRT model may be appropriate for the data. Goodness-of-fit statistics may be used to test for the amount of improvement in model fit to the data. The simpler Rasch model is nested within the 2PL model, which is nested within the more complex 3PL IRT model. The degrees of freedom for the test of the difference between the goodness-of-fit statistics between nested models is the difference in additional parameters needed to be estimated for the more complex model. For example, the test of difference in model fit between the 2PL and Rasch model has nine degrees of freedom representing the removal of the constraint of equal discrimination across the ten items by the 2PL IRT model. The difference in fit statistics suggests that the 2PL IRT model provide a significantly improved fit ($G^2(9) = 374.8$, $p < .01$) to the data than the Rasch model. Evidence for allowing the discrimination power to vary among scale items in the 2PL model may also be observed in the variation of item factor loadings from the factor analysis and the variation of item-scale correlations (point-biserial correlations) provided in Table A-2. Tests for the improvement in model fit for the 3PL model did not find any additional explanatory power for using the more complex model. Also, some may argue theoretically that the 3PL model's pseudo-guessing parameter may not provide any insight into the respondent's behavior in health care research.

Item and Scale Analysis using Item Response Theory

Table A-3 displays the two-parameter logistic IRT model estimated slope and threshold parameters for the ten items in the brooding subscale. The high slope parameters (*a*) indicate the items are highly related to the latent trait measured by the scale. For example, *I brood a great deal* along with *I am happy most of the time* and *I certainly feel useless at times* all have a high

Table A-3 IRT Estimated Parameters for the Brooding Subscale

		clinical females (n = 2569)	
#	item	*a*	*b*
1	Periods when I couldn't "get going"	1.95	-0.02
2	I wish I could be as happy as others	2.46	-0.15
3	I don't seem to care what happens to me	2.20	1.33
4	Criticism or scolding hurts me terribly	1.03	-0.26
5	I certainly feel useless at times	2.42	-0.03
6	I cry easily	1.11	-0.23
7	I am afraid of losing my mind	1.71	0.75
8	I brood a great deal	1.84	0.93
9	I usually feel that life is worthwhile	1.84	1.24
10	I am happy most of the time	2.83	0.25

Note: Multilog was used to calculate the two-parameter logistic model.

loading (*a* parameter), as would be expected for a scale measuring brooding. The location of the items across the underlying trait are scattered from $\theta = -0.26$ to $\theta = 1.33$. The item with the lowest threshold is *I cry easily*, suggesting that it does not take high levels of the brooding trait for a clinical female to endorse this item. On the other hand, it takes high levels of brooding for a person to respond "false" to the item *I usually feel that life is worthwhile*.

Figure A-3 displays each of the item's trace line (top graph) and information curve (bottom graph) for the clinical female sample. From the item trace lines, one can determine how discriminating each item is relative to each other and over what range of depressive severity

Figure A-3 Item Characteristic Curves and Information Curves for Each of the Items in the Harris-Lingoes Brooding Subscale.

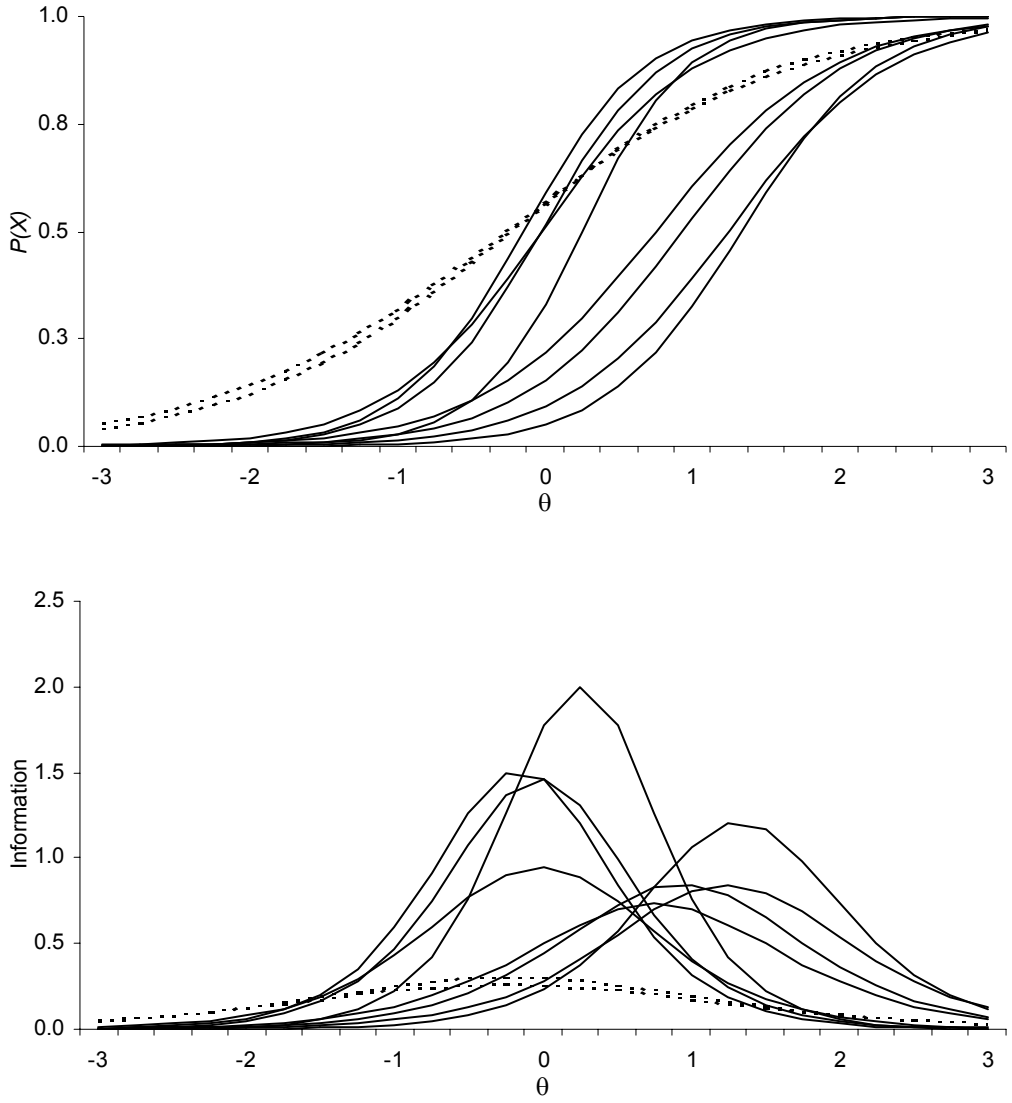
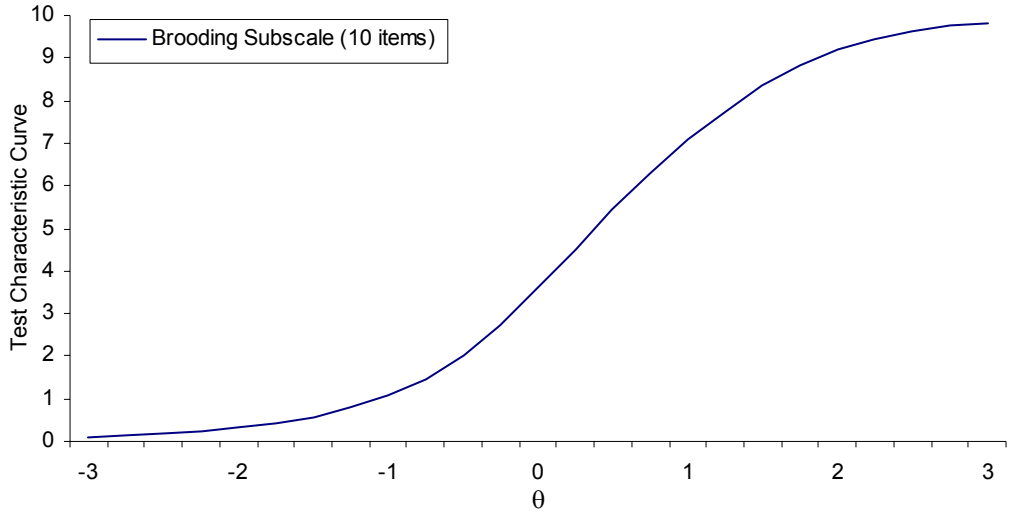
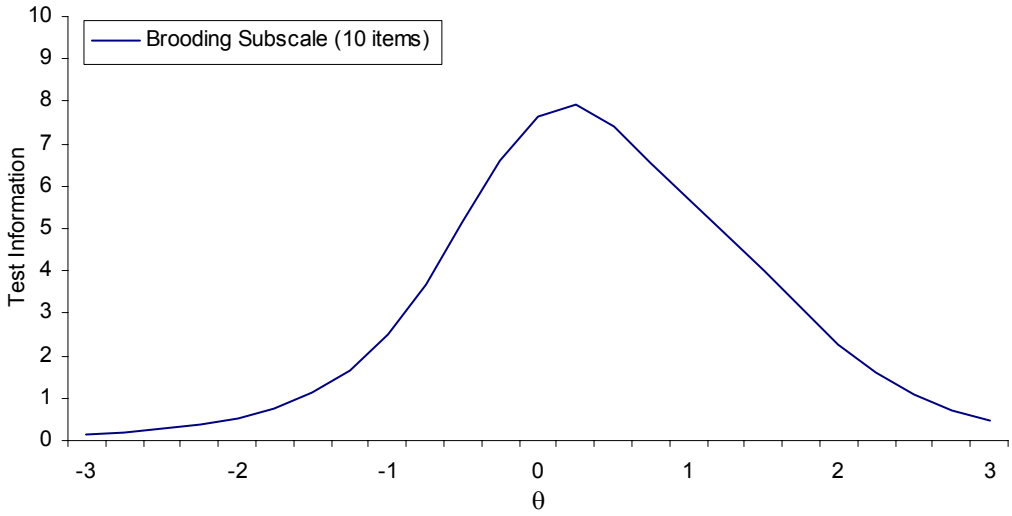
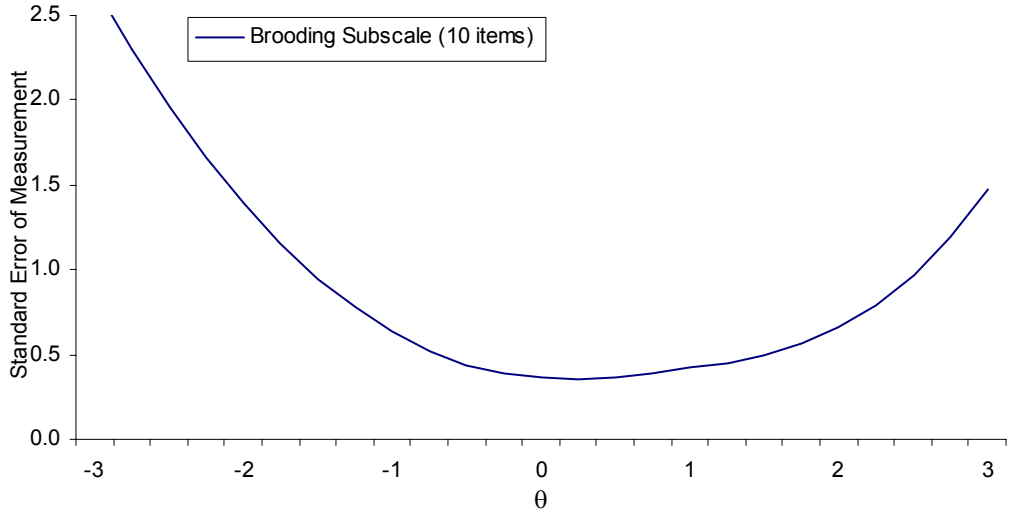


Figure A-4 Test Characteristic Curve, Test Information Curve, and Standard Error of Measurement Curve for the Harris-Lingoes Brooding Subscale.







items best discriminate among individual trait levels. The two items marked by dash lines (*I cry easily* and *Criticism or scolding hurts me terribly*) are the least related to the underlying trait in relation to the other items in the scale. The item information curves are an index indicating the range of the brooding trait over which an item is most useful for distinguishing among individuals. Information curves with high peaks denote items with high discrimination, thus providing more information over the trait levels around the item's estimated threshold. The two items marked by dashed lines (*I cry easily* and *Criticism or scolding hurts me terribly*) provide little precision to estimating trait levels around the items' thresholds.

Figure A-4 displays the brooding subscale's test characteristic curve (top), test information curve (middle), and test conditional standard error of measurement curve (bottom) for the clinical female group. The test characteristic curve describes the expected number of items that will be endorsed in a direction consistent with brooding, conditional on the level on the underlying trait. The test information curve shows what level of the underlying variable the scale is most accurate for measuring levels of brooding. An inverse square-root transformation of the test information curve yields the test standard error of measurement curve, which describes the precision in measurement of trait levels across theta. Together, these diagrams suggest that the subscale measures respondents from middle to high levels ($-1 < \theta < 2$) of the latent brooding trait. No items in the subscale appear to measure low levels of the brooding trait. This finding is expected given that the MMPI-2, a diagnostic instrument, is designed to identify individuals scoring high on depressive traits.

Analysis of Differential Item Functioning

Differential Item Functioning (DIF) is a condition when an item functions differently for respondents from one group to another, after controlling for differences between the groups on

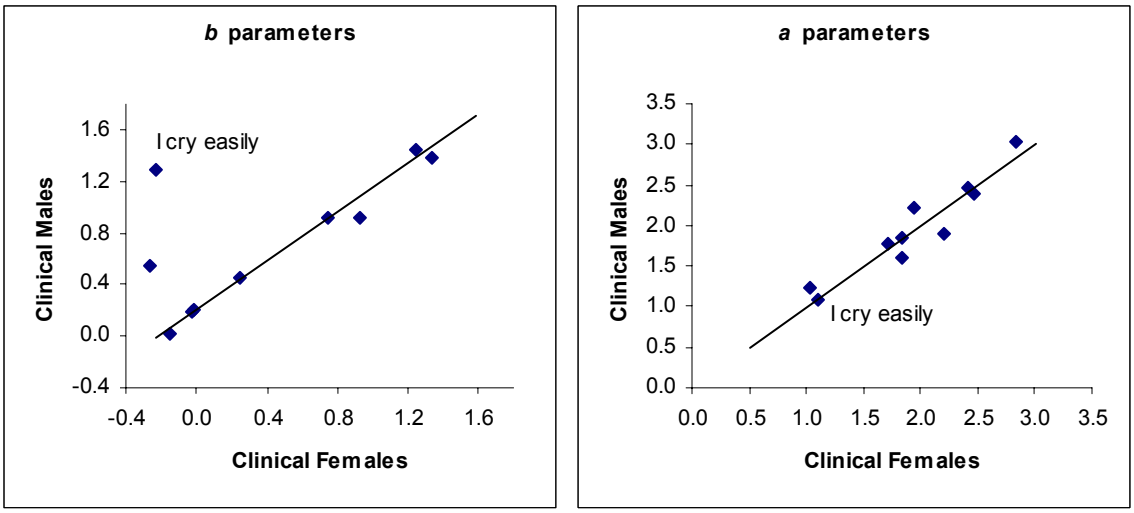
the underlying trait. To illustrate the analysis of DIF within the framework of IRT, responses from the clinical female sample as well as responses from a clinical male sample will be considered. This analysis will focus on how items within the brooding subscale may function differently between females and male respondents in a clinical population.

The clinical male sample responses were analyzed in a similar manner to the methodology used above for the clinical female respondents. The scale appears to be unidimensional with some possible problems with item local independence. The 2PL IRT model was fit to the clinical male data and item functioning appears to be similar to the findings from the clinical females.

Once parameters have been estimated separately for each group, the next step is to identify items that may function differently between the groups and a set of items that can be used as "anchors" for testing DIF. Item anchors are items thought to function similarly between the groups and will be used to link the two groups together on a similar metric. To identify item candidates for DIF and anchors for linking the two populations, item parameters for the entire scale are plotted on a bivariate graph for a pairwise comparison among groups (Angoff, 1982). This step is performed separately for each estimated parameter (e.g., item *b* parameters estimated from clinical females versus estimated from clinical males, see Figure A-5). Unbiased items will present a linear relationship in the scatterplot, while outlying or divergent items may suggest differential item functioning between the groups (Mackinnon, Jorm, Christensen, Scott, Henderson, & Korten, 1995). Even when the groups differ in the level of the latent variable, the items will still present a linear relationship, but will be displaced vertically or horizontally depending on which group is higher on theta (Angoff, 1982). In the left diagram of Figure A-5,

the item *I cry easily* appears to have a higher threshold for clinical males than females. In the right diagram, discrimination power for this item appears to be equal across the two groups

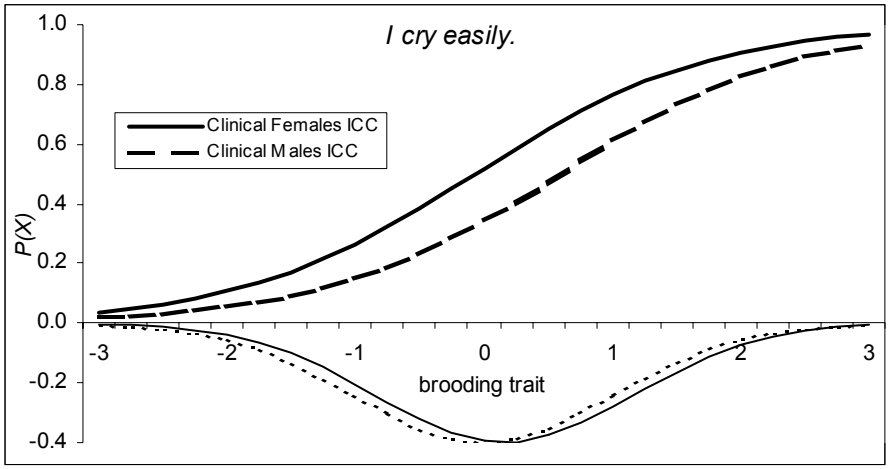
Figure A-5 Bivariate plots of item parameters between female and male groups



Holding the other items in the scale as anchors to link the female and male groups on the same metric, goodness-of-fit tests find that that males need higher levels of the brooding trait than females to endorse the item *I cry easily*. Figure A-6 presents the item trace lines for both the clinical females and the clinical males (dashed line) for the item *I cry easily*. Over all levels of the underlying trait, females have a higher probability of endorsing the item than males. Below the horizontal axis in Figure A-6, the estimated population distribution curves for both the females and males (dashed line) show that based on the item anchors, the females are estimated to have a population mean approximately 0.15 standard units above the males on brooding.

The differences in the item's threshold level across groups may reflect society's different expectations of the gender types. Crying is viewed by society as a normal reaction to a stressful situation for females, but abnormal for males (Shaeffer, 1988; Santor, Ramsay, & Zuroff, 1994). Men are taught that crying is a sign of weakness, but women are given leeway to show their emotions (Sprock & Yoder, 1997; Shaeffer, 1988).

Figure A-6 Item Characteristic Curves for Females and Males



Scale Summary

The analyses find that the ten-item scale functions well for measuring respondents high on the latent trait (see test information curve in Figure A-4). This is consistent with the purpose of the MMPI-2, to identify persons high on problematic personality traits that may need clinical help. If one wanted to design an instrument that would accurately measure respondents at all levels of the latent trait, one would need to add items that measure respondents at the low levels of the scale.

Early in the analyses, two item pairs were found to be locally dependent, that is, responses to the items are highly correlated after controlling for the underlying trait measured by the scale. For the locally dependent item pair, *Brooding or scolding hurts me terribly* and *I cry easily*, the latter item was found to function differently between female and male respondents. This item has been identified by many as a characteristically poor item because of the differences in society's attitude regarding gender roles and emotional display. Keeping this item in the scale unfavorably biases females to have higher estimated levels of depression. Removal of the item will remove local dependence and gender bias. Crying is certainly an indicator of depressive

symptoms, and if test constructors wish to include the item on the depression scales, it may be beneficial to reword the item to "I spend a great deal of time crying" or "Most days I cry."

The other item pair found to be locally dependent is *I don't seem to care what happens to me* and *I usually feel that life is worthwhile*. Both items provide precision in measurement in the higher levels of the brooding trait. Removal of one of the items will certainly decrease reliability. Therefore, another option to remove item dependence is to combine the items into a testlet. A testlet combines one or more items into a single item with multiple categories. This single item preserves the unidimensionality of the scale without removal of any information from the items. Thus, for parameter estimation and scoring, a testlet can combine both of these two dependent items into an item with ordered levels of: 0 – endorsed neither item, 1 – endorsed the item *I usually feel that life is worthwhile*, 2 –endorsed the item *I don't seem to care what happens to me*, or 3 – endorsed both items. IRT can easily model this testlet with a graded model along with the other dichotomously scored items in the scale without any complications.