7-2-2012

Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications.

Geoffrey L. Thorpe

Andrej Favia

Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications

Geoffrey L. Thorpe and Andrej Favia
University of Maine

July 2, 2012

INTRODUCTION

There are two approaches to psychometrics. ***CLASSICAL TEST THEORY*** is the traditional approach, focusing on test-retest reliability, internal consistency, various forms of validity, and normative data and standardization. Modern test theory or ***ITEM RESPONSE THEORY (IRT)*** focuses on how specific test items function in assessing constructs. IRT makes it possible to scale test items for difficulty, to design parallel forms of tests, and to provide for adaptive computerized testing (DeMars, 2010). "(T)he basic concepts of item response theory rest upon the individual items of a test rather than upon some aggregate of the item responses such as a test score" (Baker, 1985/2001, p. 6).

Using IRT methodology in data analysis can be challenging because "IRT programs are still much more traditional and 'user-unfriendly' than many commercially-available statistical packages" (Kline, 2005, p. 107). Because there is something of a learning curve involved when one first starts to use programs like these, this paper outlines some of the basic procedures involved in using two representative programs: ***MULTILOG*** (MULTILOG 7; du Toit, 2003; Thissen, Chen, & Bock, 2003) and ***PARSCALE*** (PARSCALE 4; du Toit, 2003; Muraki & Bock, 2003). A third program, ***WINSTEPS*** (WINSTEPS 3.61.2; Bond & Fox, 2007; Linacre, 2006), is also noted briefly. Also provided is some of the essential background material on IRT theory and rationale, and on its requirements and assumptions. Readers are encouraged to consult the software manuals, books, chapters, and articles in the reference list for more detailed information on technique and for authoritative and definitive theoretical coverage.

ITEM RESPONSE THEORY (IRT): A BRIEF HISTORY

The following information is drawn from Baker (1985/2001), du Toit (2003), and Embretson and Reise (2000). A paper by D. N. Lawley of Edinburgh University introduced IRT as a measurement theory in 1943. By 1960 Georg Rasch had developed IRT models in Denmark to measure reading ability and to devise tests for the military. His name is associated with one of the best-known IRT models. Allan Birnbaum contributed four chapters on IRT to Lord and Novick's (1968) *Statistical Theories of Mental Test Scores.* Lord worked at the Educational Testing Service (ETS), and therefore had access to huge databases of test scores. In 1969 Fumiko Samejima pioneered graded response models in IRT, and her name is associated with the polytomous IRT models that deal with Likert-scale data and other tests with ordered multiple response options for each item (DeMars, 2010). Bock, Thissen, and others drew from her work in developing IRT parameter estimation models such as the marginal maximum likelihood method used in MULTILOG. In 1974 Gerhard Fischer extended Rasch's binary or dichotomous model so as to handle polytomous data in the linear logistic latent trait model. Benjamin Wright directed doctoral dissertations in education based on IRT methodology at the University of Chicago in the 1970s and 1980s. Wright influenced Richard Woodcock, who developed the Woodcock-Johnson Psycho-Educational Battery (Woodcock, McGrew, & Mather, 2001).

But we find the following figure to be an excellent representation of some of the IRT essentials, one that predates all of the above citations:

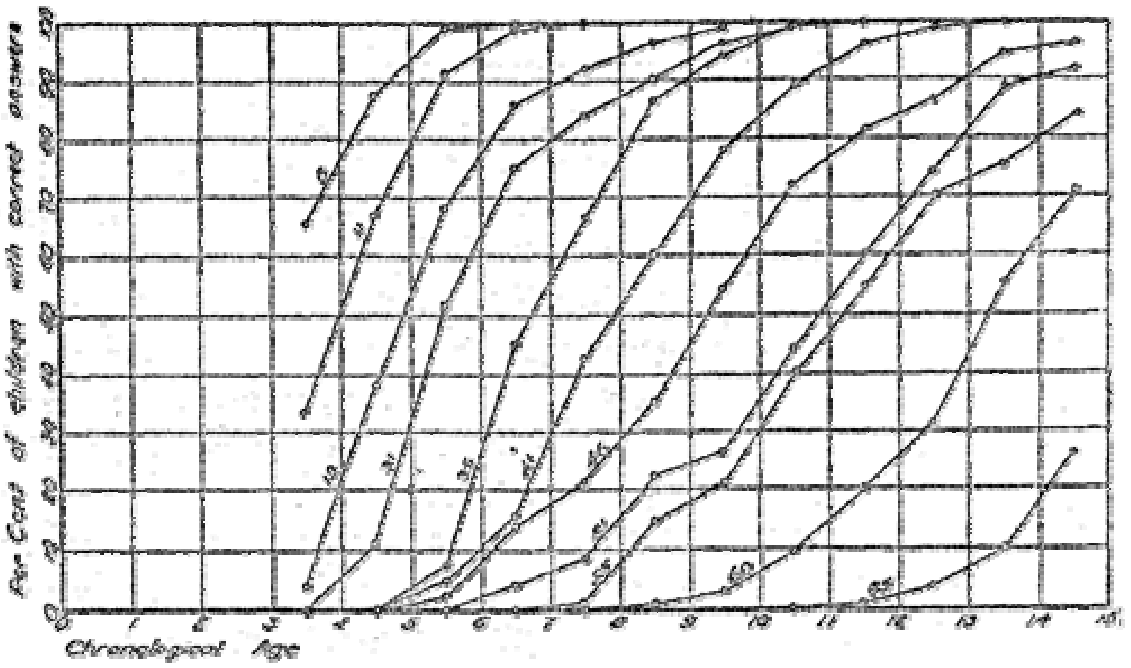


FIG. 5.

Figure 1: Proportions of correct response to selected items from the Binet-Simon test among children in successive age groups (Thurstone, 1925, p. 444). This material was originally published by the American Psychological Association, and is now in the public domain

In Figure 1 (Fig. 5 in the original article) each of the 11 trace lines represents a test item. The "percent of children with correct answers" on the y-axis easily translates into an index of the probability of a correct response to an item by children of different ages. Arbitrarily, the point at which the trace line for an item crosses the .5 probability level on the y-axis (the 50% mark) serves as an index of item difficulty as represented by student age levels. The corresponding location on the x-axis is the difficulty level, as expressed by the chronological age of 50% of those who pass the test item. In many current IRT applications, a complex iterative process incrementally refines the estimates of examinees' ability levels and item difficulties from responses to the test as a whole, but in this example chronological age is the index of ability. IRT methodology in a nutshell!

LIMITATIONS OF CLASSICAL TEST THEORY

Interchangeable Test Items

A potential problem with classical test theory methods is the often untested assumption that the items within a test are interchangeable in contributing to a total test score, the aggregate of the scores on each item. The Beck Depression Inventory – II (BDI-II; Beck, Steer, & Brown, 1996), for example, has 21 items, each scored on a 4-point scale; respondents choose one of the response options 0, 1, 2, or 3 for each item. A respondent's score on the BDI-II is simply the aggregate of the scores on each of the 21 items. Scores of 14 and above signify at least mild depression when the BDI-II is used as a screening measure. But there are many, many different ways of scoring 14 or higher on the BDI-II. (In fact the total number of possible response patterns on the BDI-II is 4^{21}, or 4.39 trillion!) Patients with the same test score may have widely varying patterns of specific problems or symptoms.

A counterargument is that the items within a test typically intercorrelate substantially, perhaps undermining concerns about the assumed equivalence of test items. It is also true that test items may be assigned different weightings before one summates the item scores to produce a test score. However, that is not the case in many of the tests that are familiar to clinical and other applied psychologists. To take another example, the Self-Harm Inventory (SHI; Sansone, Wiederman, & Sansone, 1998) presents 22 true/false items to identify self-destructive behaviors and borderline personality disorder. Any five (or more) item endorsements are significant for psychopathology. Item 2 asks if respondents have cut themselves on purpose, Item 3 asks if respondents have burned themselves on purpose, Item 10 asks about making medical situations worse, such as by skipping medication, and Item 20 asks if respondents have tortured themselves with self-defeating thoughts. Although we might hypothesize that such items reflect different levels of psychopathology, all SHI items are given equal weighting. It is unlikely that all test response patterns producing scores of five or above on the SHI have been studied, because the total number of possible patterns is 2^{22}, and the number of patterns producing scores of 5 or higher is also very large. In estimating item difficulties, an IRT analysis of SHI data could reveal whether skipping medication or burning oneself is associated with a higher trait level for self-destructiveness.

Concerns about item equivalence are redoubled in tests with more than two ordered response options for each item. The next example derives from a test used in forensic psychology to determine a criminal suspect's understanding of the 4 elements of the famous Miranda warning (Frumkin, 2010; Miranda v. Arizona, 1966; Rogers, 2011). The comprehension of Miranda Rights test (Grisso, 1998) has four items, one for each element (e.g., "You have the right to remain silent"). The interviewee paraphrases the element to indicate his or her level of understanding. Each response is rated 0, 1, or 2. A score of 0 signifies no comprehension; 1, partial comprehension; and 2, full comprehension. Imagine two defendants, each of whom scores 6 on the test. One has 2 points for each of two items and 1 for each of the remaining two items; the other has 2 points for each of three items and 0 for the remaining item. The first defendant has at least partial understanding of all elements, but the second has no understanding at all of one of them; yet both have the same score (Frumkin, 2010).

The Metrics of Likert Scales

A typical Likert scale provides five response options for each survey item, such as Strongly Disagree (SD), Disagree (D), Neutral (N), Agree (A), and Strongly Agree (SA). Each option is coded numerically; for example, SD=1, D=2, N=3, A=4, and SA=5. But it has been argued that traditional methods for analyzing Likert data are inappropriate because they assume interval or even ratio measurement; "the relative value of each response category across all items is treated as being the same, and the unit increases across the rating scale are given equal value" (Bond & Fox, 2007, p. 102). Some would describe Likert scales as representing at best nominal or categorical measurement, but even if they are viewed as involving ordinal measurement they "do not have origins or units of measurement and should not be treated as though they are continuous" (Linacre, 2005). The importance of this is that the assumptions of many traditional statistics may not be met. For example, Linacre (2005) points out that product-moment correlation coefficients are only appropriate for continuous variables.

Furthermore, researchers seem to choose the metric for Likert scale response options quite arbitrarily, with significant implications for measurement. This example is drawn from Bond and Fox (2007, pp. 102 ff.):

Suppose the categories are:
SD D N A SA
and the coding is:
1 2 3 4 5

Five endorsements of SA give 25, five times the score given by five endorsements of SD (i.e., 5), and about twice the score of endorsing two Ns and three Ds (i.e., 12).

But suppose the categories are the same:
SD D N A SA
but the coding is:
0 1 2 3 4

Five endorsements of SA give 20, five endorsements of SD give 0, and endorsing two Ns and three Ds gives 7 (about one third of the score given by five SAs).

Differences in Item Difficulty

To give a very obvious illustration of the problems of uncritically giving equivalent weightings to all test items in the context of Likert scales, suppose the items on a phobic anxiety scale (our example) include:

1. I am so anxious that I have not left my house for five years; and

2. I feel uncomfortable in large crowds, though I do not avoid them.

Presumably the first item represents much higher anxiety than the second. Bond and Fox (2007, p. 103) argue that the responses to two items like these might, if verified empirically, more appropriately line up like this:

1. (stay at home)

SD D N A SA

2. (crowds)

SD D N A SA

and thus the scoring should be adjusted accordingly. By providing information on item difficulty and on how the different response options function within each item, graded response models from IRT can inform the allocation of appropriate item and response- option weightings.

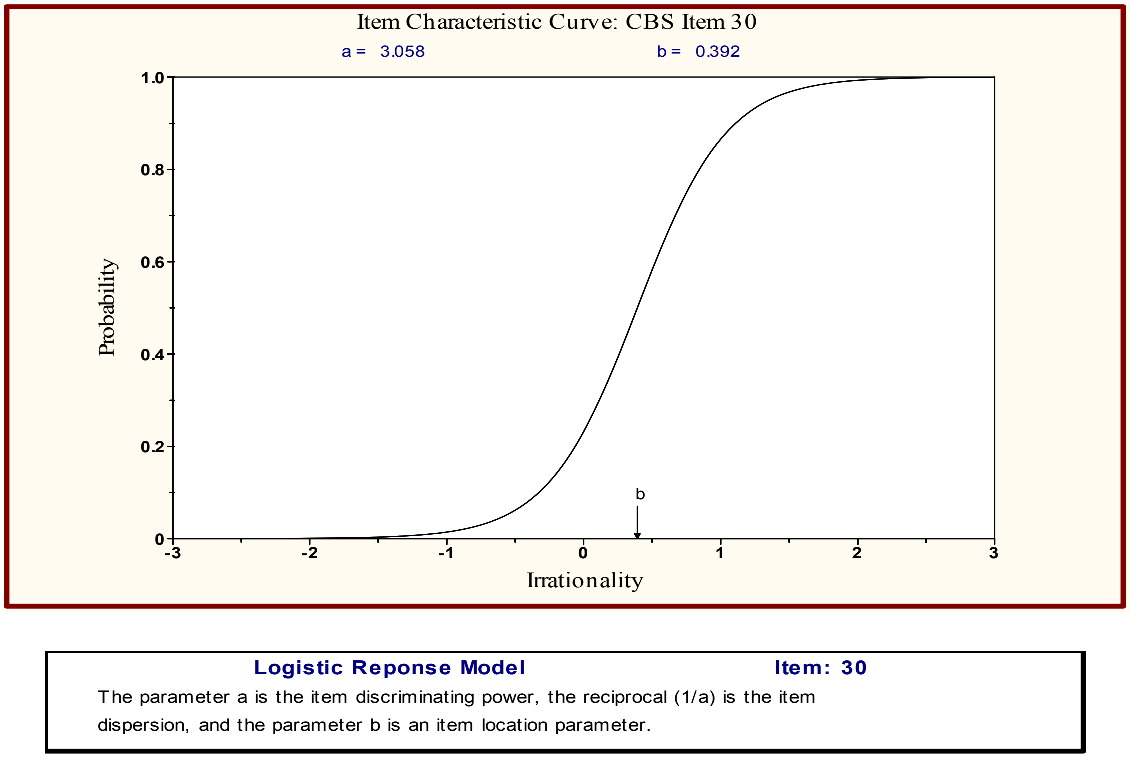
SELECTED IRT CONCEPTS

The Item Characteristic Curve (or item response function): Dichotomously-Scored Items

The *item characteristic curve* (ICC) is central to any presentation of item response theory. In the simplest case, item characteristic curves plot participants' responses to dichotomously-scored test items.

An assumption is that each test respondent possesses a certain level of the implicit trait or ability measured by the test, often assessed initially by computing a respondent's total score on the test. The level of the trait is designated by theta (θ). The curve plots $P(\theta)$ (the probability of a respondent with a given level of θ making the designated response) as a function of trait level (ability). Trait levels are plotted along the x-axis of the graph, sometimes using z-scores ranging from -3 to $+3$ as the arbitrary measurement scale, as in MULTILOG and PARSCALE (but not WINSTEPS). Lower trait levels are represented towards the left on the x-axis, and higher trait levels towards the right. The probability that respondents with a given trait level will endorse the item can be estimated. Again, the curve plots $P(\theta)$ as a function of trait level (ability). The probability that a respondent with a certain level of the trait will agree with the item, $P(\theta)$, is plotted against the y-axis.

This sample ICC (Figure 2) was taken from the responses of 605 individuals to Item 30 of an irrational beliefs inventory, the Common Beliefs Survey - III: "The influence of the past is so strong that it is impossible to really change" (Thorpe, McMillan, Sigmon, Owings, Dawson, & Bouman, 2007, p. 178). Responses were recoded dichotomously from a Likert scale to reflect agreement versus disagreement with this item, which is worded in the direction of irrationality. It is assumed that we are using unidimensional IRT models (with a single latent trait) with dichotomous or binary data (e.g., true/false, agree/disagree, correct/incorrect). The ICC is a smooth, S-shaped curve, the cumulative form of the logistic function, derived in turn from the normal ogive.



Logistic Reponse Model **Item: 30**

The parameter a is the item discriminating power, the reciprocal (1/a) is the item dispersion, and the parameter b is an item location parameter.

Figure 2: Sample Item Characteristic Curve from MULTILOG

The ***one-parameter logistic*** (1PL) model is informally known as the Rasch model (although there are technical differences between the 1PL and the Rasch models). In the 1PL model items differ only in difficulty; the slopes of the curves are equal (are held constant).

The ***two-parameter logistic*** (2PL) model estimates two parameters: difficulty and discrimination. Figure 2 above was produced by the 2PL model program in MULTILOG.

The ***three-parameter logistic*** (3PL) model estimates the difficulty and discrimination parameters, and includes guessing as a pseudo-parameter.

Item Difficulty

Item difficulty (designated by b) describes where the item functions along the ability scale. An easy item functions among the low-ability examinees; a difficult item functions among the high-ability examinees (Baker, 1985/2001). Thus difficulty is a location index along the x-axis, i.e. how far to the right or left the curve is displaced. The index of an item's location is the point on the x-axis at which the curve crosses the 0.5 probability value on the y-axis.

Item Discrimination

Item discrimination (designated by a) indicates how well it separates respondents with abilities below (to the left of) the item location from those with abilities above (to the right of) the item location. Discrimination is shown by the *steepness of the curve* in its middle section – the steeper the curve, the better the discrimination of the item. A useful mnemonic for recalling that ***a*** denotes discrimination and ***b*** difficulty is that “a” appears in the word “discrimination.”

Table 1: Interpretation of Values for Discrimination (a); from Baker (1985/2001, p. 34)

0.01 – 0.34	very low
0.35 – 0.64	low
0.65 – 1.34	moderate
1.35 – 1.69	high
1.70 and above	very high

Test Information

The traditional index of the utility of a test from classical test theory is its standard error of measurement ($\mathrm{SE_M}$). It is assumed that raw scores on tests and test items are a composite of the true score and random error. The $\mathrm{SE_M}$ refers to “the distribution of random errors around the true score” (Kline, 2005, p. 92). Thus, the lower the $\mathrm{SE_M}$ value, the more dependable is the test score. A single value for the $\mathrm{SE_M}$ is given for the test as a whole. By contrast, “in IRT, the concept of test and item information is used. Information is [inversely related to] the $\mathrm{SE_M}$, and is calculated separately for different ability levels. The test information function indicates how well each ability level is being estimated by the test” (Thorpe et al., 2007, p. 179).

The Logistic Function

“(T)he standard mathematical model for the item characteristic curve is the cumulative form of the logistic function” (Baker, 1985/2001, p. 21). Logistic functions have often been used to model “resource limited exponential growth” (Mueller, 2008). “In much of the . . . literature, the parameter a [*discrimination* in dichotomous models, *slope* in graded response models] is reported as a **normal** [emphasis added] ogive model value that is then multiplied by 1.7 to obtain the corresponding **logistic** [emphasis added] value” (Baker, 1985/2001, p. 22), but that is not the usage in Baker’s book, or in the MULTILOG program: “All models estimated with MULTILOG are truly ‘logistic,’ which means that there is no $D = 1.7$ scaling factor. This means that item-discrimination (slope) parameters will be approximately 1.7 times higher than they would be if reported in the normal metric” (Embretson & Reise, 2000, p. 337).

It is important to be aware that logistic is the default in MULTILOG, whereas in

PARSCALE the user selects either normal ("NORMAL, SCALE=0") or logistic ("LOGISTIC, SCALE=1.7") on the >CALIB command.

Computing ICC Curves

For computing the points on the ICC in the 2PL model, Kline (2005) gives the basic equation in a form similar to the following (p. 111):

$$P(\theta) = \frac{1}{1 + e^{-L}} = \frac{1}{1 + e^{-a(\theta - b)}} \tag{1}$$

$e =$ the constant 2.718
$L = a(\theta - b)$, the negative of the exponent, is the logistic deviate (logit)
$a =$ item discrimination
$b =$ item difficulty
$\theta =$ the ability level

$$\textit{Note:} \quad e^{-L} = \frac{1}{e^{L}}$$

In Equation 1, the part of the equation in which ***1 is divided by 1 plus*** the e^{-L} part is simply a device for ensuring that the values for $P(\theta)$ fall between 0.00 and 1.00 on the y-axis of the graph. Also: when $a = 1$, the equation is effectively that of the 1PL model.

Other authors (e.g., Embretson & Reise, 2000; Liao, 2006; Ostini & Nering, 2006) give the equation for the 2PL model in this form:

$$P(\theta) = \frac{e^{-L}}{1 + e^{-L}} = \frac{e^{-a(\theta - b)}}{1 + e^{-a(\theta - b)}} \tag{2}$$

Equation 2 is modified from Liao (2006, p. v), who gives the equation for the 3PL model as follows (with slight modifications):

$$P(\theta) = c + (1 - c) \cdot \frac{e^{-L}}{1 + e^{-L}} = c + (1 - c) \cdot \frac{e^{-a(\theta - b)}}{1 + e^{-a(\theta - b)}} \tag{3}$$

Note. $c =$ the pseudo-guessing parameter. When $c = 0$, the equation gives the same results as for Equation 2.

Group Invariance

With a given latent trait, estimates of item characteristics hold true regardless of the group being tested; a group of respondents low on the trait will produce the same ICCs as a group high on the trait: "the item parameters are not dependent upon the ability level of the examinees responding to the item" (Baker, 1985/2001, p. 51).

Goodness of Fit for Very Short Tests

Goodness of fit can be tested via chi-square if one can count the frequencies in the test sample of all possible response patterns. For example, the MULTILOG program allows one to enter data representing counts of response patterns as one of the options; doing so produces output that includes a chi-square value for the test that can be evaluated for goodness of fit – a **non**-significant chi-square value indicates a fit with the model. A 4-item test, dichotomously scored, produces 2^4 (or 16) possible patterns:

Table 2: Possible Patterns of Responses (16) to a 4-item Test with Items a, b, c, and d

abcd	abcd
0000	1000
0001	1001
0010	1010
0011	1011
0100	1100
0101	1101
0110	1110
0111	1111

When the four items are listed from left to right in increasing order of difficulty, then we would expect to see five of these 16 patterns strongly represented in the respondents' data: 0000, 1000, 1100, 1110, and 1111.

Table 3 provides an example in which over 78% of the possible patterns produced by respondents were consistent with an increasing order of item difficulty from left to right (Thorpe, Owings, McMillan, Burrows, & Orooji, 2011).

Table 3: Example of Frequencies of Possible Patterns of Responses. Frequencies of the 16 possible response patterns in 416 students taking a 4-item legal knowledge test when the items a, b, c, and d are ranged from left to right in order of increasing difficulty

abcd		abcd	
0000	**4**	**1000**	**13**
0001	0	1001	11
0010	0	1010	30
0011	1	1011	8
0100	1	**1100**	**55**
0101	1	1101	37
0110	1	**1110**	**99**
0111	1	**1111**	**154**

The patterns in bold print (shown by 325 out of 416 respondents) indicate responding consistent with expectation, if indeed the 4 items are accurately scaled for difficulty.

The degrees of freedom for tests of overall fit when the number of items is 10 or fewer are given by:

$$df = 2^n - kn - 1 \tag{4}$$

where n is the number of binary item scores (4 in the example above) and k is the number of parameters in the model (2 in the case of the 2PL model; du Toit, 2003, p. 30).

Comparing Models for Fit

In MULTILOG, the output files from analyses using the 1PL or 2PL models produce a *negative twice the loglikelihood* statistic (chi-square for several times more examinees than cells); higher values indicate a poorer fit of the data to the model. Comparing the values from different models can indicate which model represents a better fit. For example, a set of 14 items from the legal knowledge test noted earlier (Thorpe et al., 2011) was analyzed using the MULTILOG programs for the 1PL (estimating item difficulty only) and 2PL (estimating item difficulty and discrimination) models. The 2PL model produced a lower value than the 1PL model. Subtracting the former value from the latter gave $\chi^2(14) = 25.1$, $p < .05$, indicating that the 2PL model fit the data significantly better than the 1PL model (Kline, 2005). The degrees of freedom are determined by the number of additional parameters in the 2PL model versus the 1PL model, 14 in this case (one additional parameter, discrimination, for each of the 14 items).

ASSUMPTIONS IN USING IRT METHODOLOGY

"Generally, the process proceeds from data collection, to evaluating assumptions, selecting and fitting a model, and determining the fit, and then finishes with applications" (Morizot, Ainsworth, & Reise, 2007, p. 411). The following sections draw heavily from the Morizot et al. (2007) chapter.

Sample Size

Concerning sample size in IRT analyses, "there is no gold standard or magic number that can be proposed" (Morizot et al., 2007, p. 411).

1. *Dichotomous response formats*
   Some authorities suggest that 100 respondents will suffice for the 1PL or Rasch model with a dichotomously-scored test, and some suggest that as few as 200 will suffice for the more complex 2PL model, but others recommend at least 500. Much larger sample sizes are recommended for polytomously-scored tests (Bond & Fox, 2007; Morizot et al., 2007).

2. *Polytomous response formats*
   "(I)t can be shown that the Graded Response IRT model can be estimated with 250 respondents, but around 500 are recommended for accurate parameter estimates" (Reeve & Fayers, 2005, p. 71). Reeve and Fayers refer to Embretson and Reise (2000) to support this statement. The authors are referring to polytomous data with a 5-point

Likert-scale format.

Monotonicity

Item endorsements should increase as trait levels increase (Morizot et al., 2007). "Monotonicity of items is established by high positive point-biserial correlation between the item score and the test score" (Nandakumar & Ackerman, 2004, p. 94).

Another way to quantify monotonicity is by calculating a coefficient of reproducibility, a form of Guttman scaling. Thorpe et al. (2011) describe it as follows:

> A coefficient of reproducibility (CR) can be calculated from examinees' patterns of scaled test responses (Jobling & Snell, 1961; Kline, 2005) from the formula: $C_R = 1-$ (total errors / total responses). Reproducibility coefficients below 0.85 are viewed as low, and indicate that a number of unexpected responses were made – in turn suggesting a "degree of non-unidimensionality in the set of items" (Kline, 2005, p. 44). Conversely, "a value of 0.90 or more is usually taken to indicate the existence of a scale" (Jobling & Snell, 1961, p. 110).

Some disagree with Kline's suggestion that the C_R addresses unidimensionality (next section). However, it can serve as an index of monotonicity (M. Linacre, personal communication, September 21, 2011).

Unidimensionality

Unidimensionality is defined through the property of local independence; "a scale is unidimensional when a single latent trait accounts for all the common variance among item responses" (Morizot et al., 2007, p. 413). The issue is "unidimensional enough" (M. Linacre, personal communication, June 4, 2011). There is no one accepted method for determining unidimensionality. Suitable methods include conducting an exploratory factor analysis (e.g., Funk & Rogge, 2007) and inspecting factors' eigenvalues, the ratio of the eigenvalue of the first factor to the second and subsequent factors, and the "knee" or bend in scree slopes (Ruscio & Roche, 2012). Figure 3 provides an example. Bifactor modeling, provided for example in the TESTFACT program (du Toit, 2003), can be helpful (Morizot et al., 2007). Turk, Dworkin, Burke, Gershon, Rothman, Scott et al. (2006) suggest "the use of factor analysis to examine patterns of covariation among responses, and if multidimensionality is found, then each factor can be used as a unique scale if doing so would be consistent with the overall theoretical approach" (p. 214).

Sample size for principal components analysis
"Many sources suggest that a certain minimum ratio of sample size to number of variables (e.g., 5 to 1, 10 to 1, etc.) is needed to ensure the stability of factor analytic results. These suggestions are entirely misguided. . . . The factor analysis of a large variable set does not require an especially large sample size" (Lee & Ashton, 2007, p. 430).

ANALYZING POLYTOMOUS DATA WITH IRT PROGRAMS

Reasons for Using Polytomous Models

In measuring personality and social variables, “dichotomous distinctions are often less clear in this context than in ability measurement settings” (Ostini & Nering, 2006, p. 7). Polytomous items provide “more information over a wider range of the trait continuum than . . . dichotomous items” (p. 8). But “response functions for middle categories are difficult to model mathematically” (p. 9).

Test and Item Information in Polytomous Models

- Samejima (2004): “In general, the amount of test information is larger for graded response items than for dichotomous response items, and more important, it is substantially larger for a wider range of θ” (p. 82).
- Hays, Morales, and Reise (2000/2007): “Generally speaking, items with higher slope parameters provide more item information. The spread of the item information and where on the trait continuum information is peaked are determined by the between-category threshold parameters” (p. 4).
- Embretson and Reise (2000): “[I]n several of the polytomous models the amount of information a particular item provides depends on both the size of the slope parameter and the spread of the category thresholds or intersection parameters” (p. 185).

Two Types of Polytomous Model

Rasch models use a dichotomous ICC to define category boundaries, and deal with the “two categories immediately adjacent to the category boundary” (Ostini & Nering, 2006, p. 13). Thurstone/Samejima polytomous models use the 2PL model ICC (p. 11), and deal with all “possible response categories for an item above and below the category boundary” (p. 13).

MULTILOG Graded Response Model

A *graded response model* (GRM) is suitable for Likert-type ratings. “To fit the GRM within a measure the items need not have the same number of response categories; no complications arise in item parameter estimation or the subsequent parameter interpretation as a result of a measure having items with different response formats” (Embretson & Reise, 2000, pp. 97-98). However, the MULTILOG GRM requires that the number of response options per item remain constant within a test, whereas the PARSCALE GRM allows different numbers of response options among the test items. This is achieved by assigning groups of items with the same number of response options to the same block.

The MULTILOG program uses the Samejima (1969) model for polytomous data. The following example uses data from the General Attitude and Belief Scale (GABS; Bernard, 1998), a 55-item survey of irrational beliefs of the kind that are disputed in rational emotive behavior therapy (David, Lynn, & Ellis, 2010). The GABS uses a 5- point Likert scale (1, strongly disagree; 2, disagree; 3, neutral; 4, agree; 5, strongly agree) for each item. The 55 GABS items form one "rational" and six "irrational" subscales.

The Self-Downing subscale of the GABS consists of 9 items. An exploratory principal components analysis of that subscale with 514 student respondents revealed two factors with eigenvalues greater than 1, and in a graphical depiction (Figure 3) the scree slope revealed a sharp bend between those two and the remaining factors (Owings, Thorpe, McMillan, Burrows, Sigmon, & Alley, 2012). Selecting as factors those with eigenvalues greater than 1 (the Kaiser criterion) and those above the bend in a scree slope is recommended by Lee and Ashton (2007). The first factor comprised 7 and the second two of the subscale's 9 items. The 7 items of Factor 1 were analyzed via IRT methodology.

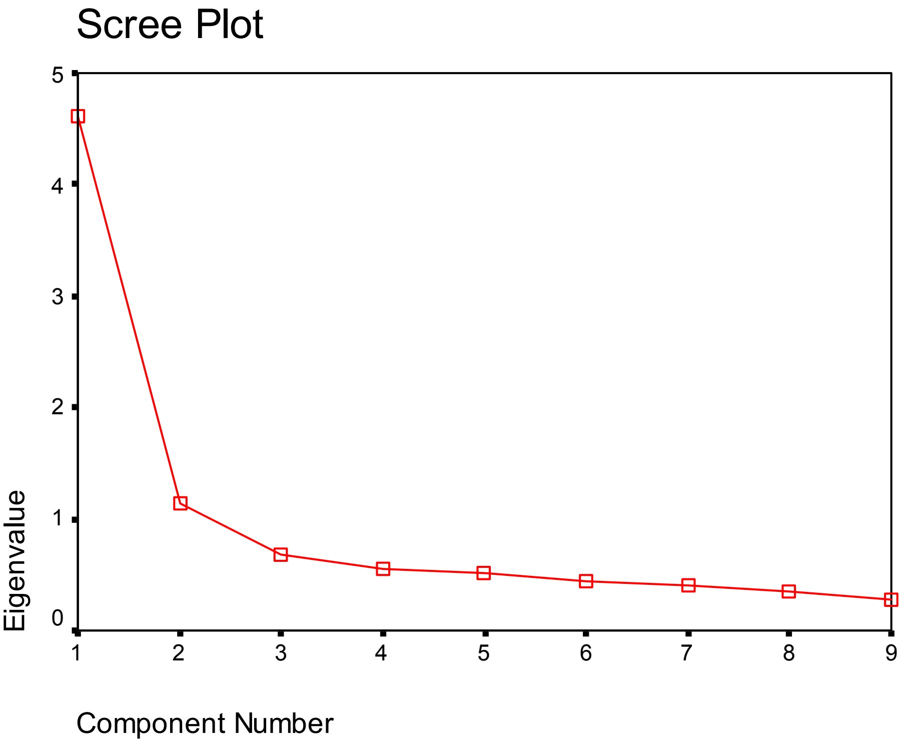


Figure 3: Scree slope from an exploratory Principal Components Analysis of the 9 Self-Downing subscale items in the GABS

The fifth of the 7 items from Factor 1 of the Self-Downing scale is GABS Item 32: ***If important people dislike me, it goes to show what a worthless person I am.*** The MULTILOG GRM printed output for this item reads as follows:

Table 4: Part of the Printed Output from a MULTILOG GRM Analysis

```
ITEM   5:       5 GRADED CATEGORIES
        P(#) ESTIMATE (S.E.)
 A        21    4.03  (0.37)
 B(1)     22   -0.09  (0.05)
 B(2)     23    1.23  (0.08)
 B(3)     24    1.79  (0.12)
 B(4)     25    2.40  (0.20)
```

A is the slope; B(1), B(2), etc. are the ability values at the thresholds between the response-option categories for the item. In the above example there are 5 graded categories or response options, and 4 B values. If there are k categories, there are $k-1$ thresholds.

- "The value of b_{k-1} is the point on the θ–axis at which the probability passed 50% that the response is in category k or higher" (du Toit, 2003, p. 567).

- "Threshold parameters represent the trait level necessary to respond above threshold with 0.50 probability" (Hays, Morales, & Reise, 2000/2007, p. 4).

- "Note that the a and b value parameters are not interpreted as difficulty and discrimination parameters as they were in the 1PL, 2PL, and 3PL models. Instead, the b value for each alternative for each item represents the ability level needed to respond above a specific threshold with 50% probability. While the a values indicate slopes, they are not interpreted as discriminations" (Kline, 2005, p. 132).

There are 4 thresholds between the 5 response-option categories in the Likert scale used in the GABS. "(W)hat occurs in the GRM is that the item is treated as a series of $m_i = K - 1$ dichotomies (i.e., 0 vs. 1, 2, 3, 4; 0, 1 vs. 2, 3, 4; 0, 1, 2 vs. 3, 4; 0, 1, 2, 3 vs. 4) and 2PL models are estimated for each dichotomy with the constraint that the slopes of the operating characteristic curves are equal within an item" (Embretson & Reise, 2000, p. 99).

PARSCALE Graded Response Model

In Muraki's *Modified Graded Response Model* (M-GRM; Embretson & Reise, 2000, pp. 102 ff.; Kline, 2005, p. 143) all items have the same number of rated categories: item-slope parameters (a) vary across items; but the category threshold parameters are divided into a location parameter (b) for each item, and a set of category threshold parameters (c) for the entire scale. *Note that this "c" is not the same as the "c" used to designate the guessing pseudo-parameter in the 3PL model.* "One advantage of the M-GRM, relative to the GRM, is that it allows separation of the estimation of item location and category threshold parameters. However, note that there is an arbitrariness to the scale of the category threshold parameters" (Embretson & Reise, 2000, p. 103).

In PARSCALE, when there are different response formats (e.g., some items have 4 response-option categories and some have 3), "then items with similar formats would have to be treated as a 'block' and category and item parameters estimated within blocks"

(Embretson & Reise, 2000, p. 103). Item parameters within a block cannot be directly compared with parameters in other blocks. Chi-square goodness-of-fit values are provided for each item (Embretson & Reise, 2000, p. 104).

Back to MULTILOG briefly: "(I)n Samejima's model, category thresholds are generated for each item in the scale" (Kline, 2005, p. 135).

POLYTOMOUS DATA: SAMPLE ANALYSES

Example in PARSCALE

This example uses data from all 9 of the GABS Self-Downing subscale items. The raw data had been recorded in an Excel file.

The MULTILOG GRM analysis had shown that the 5-item Likert scale was not functioning well in all 9 items. Accordingly, 4 of the items were recoded. The instructions for the procedure are:

1. Import the Excel data file into SPSS for Windows.

2. In SPSS, go to **data view** and recode the data using the pull-down menu (Transform, Recode, Into Same Variables).

3. Save the file as an Excel file.

4. In Excel, adjust the column widths as appropriate to the number of digits in each cell. In my example, the ID column has 4 digits and thus a width of 4, and the 9 columns of data each have one digit and thus a column width of 1. **Save the file as a .prn file.**

5. Save the PARSCALE command file (.psl file extension) in PARSCALE under All Programs. In PARSCALE, open the Excel data file with the .prn extension and save that in the PARSCALE folder. Then run all 4 phases of the program (phase 0, phase 1, phase 2, and – if you need to – phase 3, which scores the test for all respondents). The plots (figures) can be opened from the Run pull-down menu. If you do not run phase 3 the plots may not run.

Here is the PARSCALE command file for the recoded Self-Downing subscale analysis with 513 respondents. Under >COMMENT are 9 more lines that the program does not pick up. These show that items 2, 4, 5, 6, and 8 have the original 5-point Likert-scale coding. For item 1, response options 1, 2, and 3 were recoded as 1, and options 4 and 5 were recoded as 2, leaving that item with 2 response options or ***categories***. A similar notation is used for the remaining recoded items; categories separated by commas were not recoded.

Exhibit 1: Sample PARSCALE command program

```
SD513recoded.PSL
>COMMENT 03/02/2010; GABS; the 9 ‘‘self-downing" items, recoded:
        1        123/45          2
        2                        5
        3        1,2,3/45        4
        4                        5
        5                        5
        6                        5
        7        123/45          2
        8                        5
        9        1,2/345         3
>FILE    DFNAME='SD513recoded.prn',SAVE;
>SAVE    PARM='SD513recoded.PAR',SCORE='SD513recoded.SCO';
>INPUT   NIDCHAR=4,NTOTAL=9,NTEST=1,LENGTH=9;
(4A1,9A1)
>TEST    ITEM=(1(1)9),NBLOCK=9;
>BLOCK1  NITEMS=1,NCAT=2,ORIGINAL=(1,2);
>BLOCK2  NITEMS=1,NCAT=5,ORIGINAL=(1,2,3,4,5);
>BLOCK3  NITEMS=1,NCAT=4,ORIGINAL=(1,2,3,4);
>BLOCK4  NITEMS=1,NCAT=5,ORIGINAL=(1,2,3,4,5);
>BLOCK5  NITEMS=1,NCAT=5,ORIGINAL=(1,2,3,4,5);
>BLOCK6  NITEMS=1,NCAT=5,ORIGINAL=(1,2,3,4,5);
>BLOCK7  NITEMS=1,NCAT=2,ORIGINAL=(1,2);
>BLOCK8  NITEMS=1,NCAT=5,ORIGINAL=(1,2,3,4,5);
>BLOCK9  NITEMS=1,NCAT=3,ORIGINAL=(1,2,3);
>CALIB   GRADED, LOGISTIC, SCALE=1.7, NQPTS=30, CYCLES=(100,1,1,1,1),
         NEWTON=5, CRIT=0.005, ITEMFIT=10;
>SCORE   MLE, SMEAN=0.0, SSD=1.0, NAME=MLE, PFQ=5;
```

This program runs successfully when the data file is properly formatted and has been given the right address. The figures below represent the Plot output for GABS Item 32 on the Self-Downing subscale: ***"If important people dislike me, it goes to show what a worthless person I am."*** This is a useful item that is informative within a narrow range of irrationality between approximately $z = +1.00$ to $z = +2.00$. Most respondents endorse category 1, "strongly disagree," on this irrational item. Only as trait levels for irrationality approach z = 0.00 do people just "disagree." People with trait levels around z = +2.00 start using categories 3 and 4. Above $z = +2.50$ respondents are most likely to choose category 5, "strongly agree."

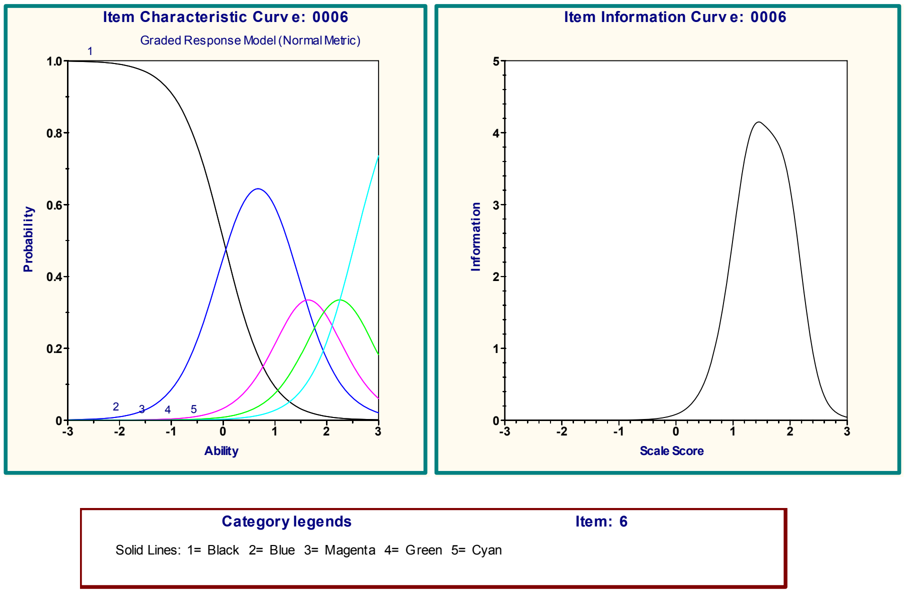


Figure 4: Item Characteristic Curves and Item Information Curve from a survey item with 5 Likert-scale response options

Exhibit 2: Sample MULTILOG command program

In the example that follows the MULTILOG GRM model is used for the Likert scale (1 – 5) responses of 544 students to the 46 "irrational" items of the GABS. Unidimensionality and monotonicity have not been addressed in this example, which is provided only to illustrate the procedures and computations. There were 4 digits (and therefore columns) in the ID column, and there was one column for each of the 46 item responses.

```
MULTILOG for Windows 7.00.2327.2
Created on: 22 September 2008, 10:30:19
>PROBLEM RANDOM,
         INDIVIDUAL,
         DATA = 'F:GABS46.prn',
         NITEMS = 46,
         NGROUPS = 1,
         NEXAMINEES = 544,
         NCHARS = 4;
>TEST ALL,
      GRADED,
      NC = (5(0)46);
>END ;
5
12345
1111111111111111111111111111111111111111111111
2222222222222222222222222222222222222222222222
3333333333333333333333333333333333333333333333
4444444444444444444444444444444444444444444444
5555555555555555555555555555555555555555555555
(4a1,46a1)
```

In this example the data file had been saved on a thumb drive ("F:"). This program (F:GABS46.MLG) ran successfully when the Excel data file was saved as a .prn file and given the correct address in the .mlg program above.

Exhibit 3: Sample PARSCALE command program illustrating use of NOCADJUST and CSLOPE options for the BLOCK command, and including three items in the same BLOCK

```
MS181913141639.PSL
>COMMENT mental status ratings of 198 criminal defendants referred for
competency examinations (Owings, 2010);
18. appropriateness of affect 1=ok, 2=intermediate, 3=inappropriate
19. labile affect 1=ok, 2=intermediate, 3=labile
13. disorganized speech 1=ok, 2=somewhat ok, 3=somewhat disorganized,
4=disorganized
14. pressured speech 1=ok, 2=somewhat ok, 3=somewhat pressured,
4=pressured
16. delusions 1=ok, 2=somewhat ok, 3=somewhat delusional,
4=delusional
39. examiner's rating of incompetence 1=competent, 2=maybe competent,
3=maybe incompetent, 4=incompetent
>FILE    DFNAME='F:MS181913141639.PRN',SAVE;
>SAVE    PARM='F:MS181913141639.PAR',SCORE='F:MS181913141639.SCO';
>INPUT   NIDCHAR=4,NTOTAL=6,NTEST=1,LENGTH=6;
(4A1,6A1)
>TEST    ITEM=(1(1)6),NBLOCK=4;
>BLOCK1  NITEMS=1,NCAT=3,ORIGINAL=(1,2,3),NOCADJUST;
>BLOCK2  NITEMS=1,NCAT=3,ORIGINAL=(1,2,3),NOCADJUST;
>BLOCK3  NITEMS=1,NCAT=4,ORIGINAL=(1,2,3,4),NOCADJUST;
>BLOCK4  NITEMS=3,NCAT=4,ORIGINAL=(1,2,3,4),CSLOPE,NOCADJUST;
>CALIB   GRADED, LOGISTIC, SCALE=1.7, NQPTS=30, CYCLES=(100,1,1,1,1),
         CRIT=0.005, ITEMFIT=10;
>SCORE   MLE, SMEAN=0.0, SSD=1.0, NAME=MLE, PFQ=5;
```

In the BLOCK command, the CADJUST keyword (the default) controls the location adjustment for the category parameters; the default setting is 0 (the category threshold values are centered on zero). The NOCADJUST option omits that adjustment. In the BLOCK command, the CSLOPE option requests the estimation of a single common slope parameter for all items in the block. The option is unnecessary if there is only one item in the block.

IRT in Perspective: Comments

IRT shows some commonality with classical test theory. For example, in examining how each test item functions in the context of the test as a whole, IRT is functioning similarly to a reliability or internal consistency model in classical test theory.

For example, the MULTILOG 2PL analysis of the 46 irrational items of the GABS indicated that the most informative items were Item 48 ($a = 2.41$) and Item 45 ($a = 2.39$). An internal consistency analysis of the 46 items revealed that, among all the items, Item 48 showed the highest correlation with the remaining items ($r = .6818$), and the value of alpha would be reduced from .9526 to .9509 if the item were deleted. Similarly, Item 45 showed the second highest correlation with the remaining items ($r = .6793$), and – if deleted – would also reduce the value of alpha to .9509.

Table 5: The 46 "irrational" GABS items: The most informative items as assessed by IRT (all items with $a > 2.00$) and internal consistency (all items with item-remainder $r > .6400$) Analyses

Item	Slope (a)	Item-Remainder Correlation	Alpha if Deleted
48.	2.41	.6818	.9509
45.	2.39	.6793	.9509
54.	2.19	.6646	.9509
42.	2.16	.6464	.9511
21.	2.08	.6401	.9511
39.	2.07	.6777	.9508

Graphics in Scientific Software International IRT Programs

When writing up IRT studies for publication or in a dissertation it can be helpful to edit the ICCs, Item Information Curves, and Total Information Curves produced by MULTILOG and PARSCALE before copying them into Word files. The trace lines can be edited by double-clicking on a line, clicking on Line Attributes, and using the menus for Color and Width. For example, Color can be changed to black, and line widths can be changed from 1 to 3. Legends on the Figure, such as Ability, Probability, Item Characteristic Curve n, a= *** and b= *** can each be edited by double-clicking on the legend and selecting and changing the Color, Font, and Size as desired. Then go to the top right tool bar and select Edit, Copy Graph. The graph can then be pasted in a Word file.

If the graph is copied into Word in Rich Text Format, then one can right click in the graph, click on Format Picture, go to Color, change Automatic to Grayscale and click OK, and the graph will have been converted to monochrome as in Figure 5.

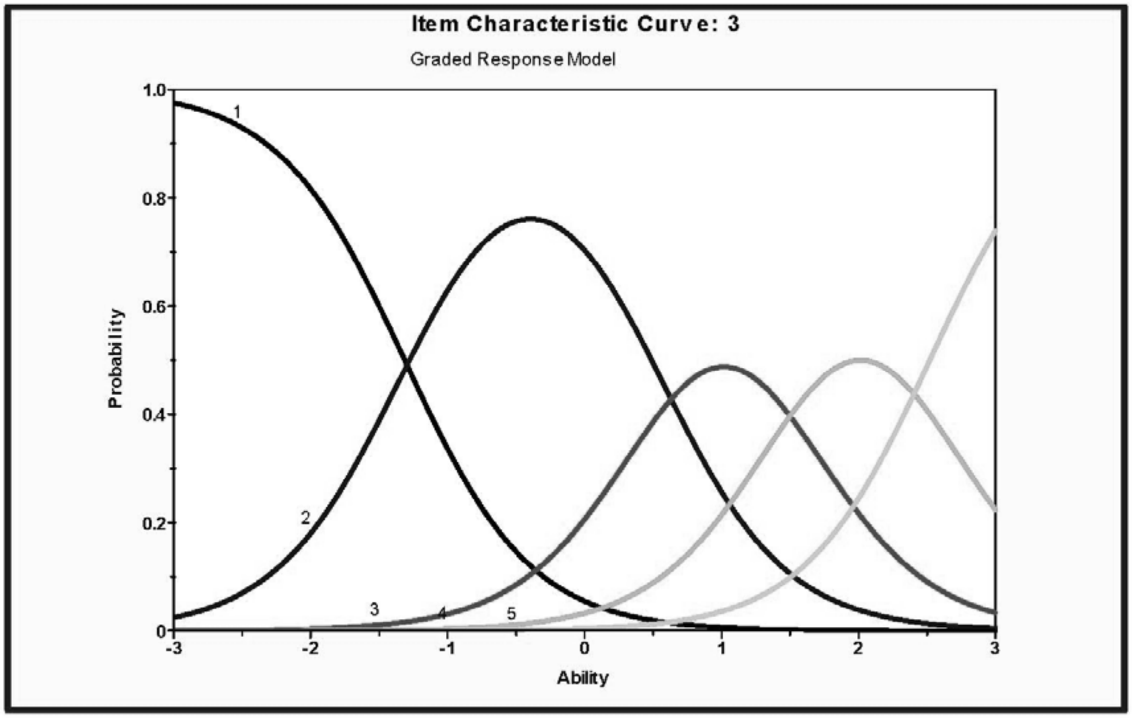


Figure 5: Item characteristic curve for a survey item with five Likert-scale response options.

Graphics from WINSTEPS

The WINSTEPS Rasch-model program (Linacre, 2006) provides figures with useful features such as the following. ICCs give the model curve, the empirical curve, and the limits of the 95% confidence interval. Furthermore, all ICCs from a test can be generated in a single figure.

Figure 6 portrays the ICC for the following item from a brief test of elementary legal knowledge (Thorpe et al., 2011):

Item 48.
A defendant goes to trial and is found not guilty. Another name for this is:
a. acquittal
b. dropping the charges
c. nolo contendere
d. actus reus

This is, of course, a very easy item that includes obviously spurious response options that could be developed as possible indicators of malingering. Figure 6 presents data from 423 undergraduate psychology students. Very similar patterns were produced by 55 individuals from the wider college campus community and by 32 criminal defendants who had been referred for court-ordered competency examinations.

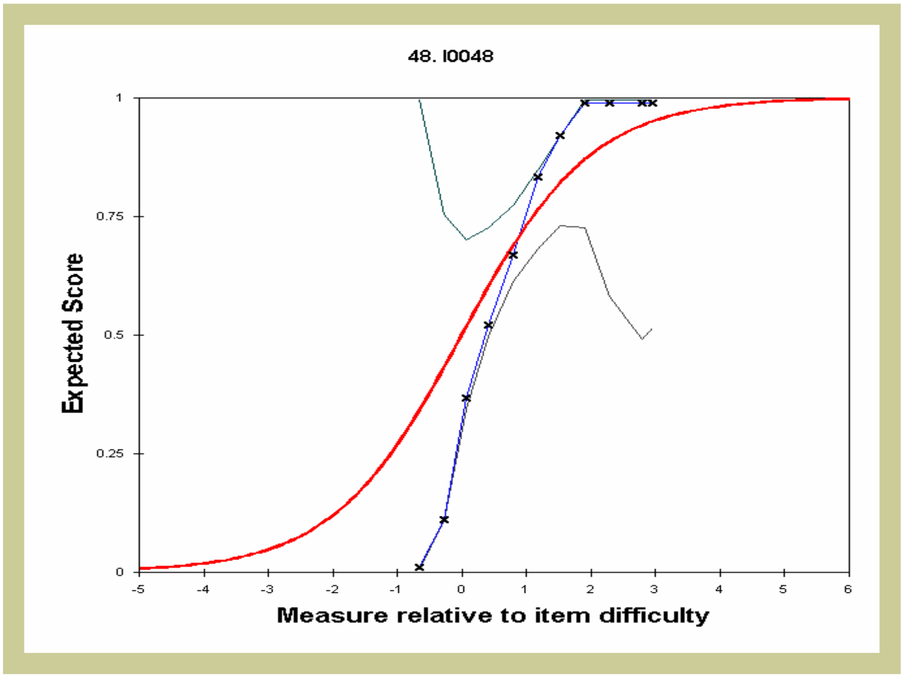


Figure 6: Item characteristic curve for a dichotomously-scored (correct/incorrect) legal knowledge test item (from the Rasch-model WINSTEPS program). The red line is the model curve, the blue line with data points is the empirical curve, and the other lines delineate the limits of the 95% confidence interval. The metric for the x-axis is in log-odds units or logits

THE MATHEMATICAL FOUNDATIONS OF IRT

As noted by Wallace & Bailey (2010) and Morizot, Ainsworth, & Reise (2007), the common practice in psychometric analysis is to use classical test theory (CTT), which compares the difference in the observed vs. true participant score, or the observed score variation vs. the true score variation. The reliability of CTT thus relies on parameters that strongly depend on the sample. An alternative is to use item response theory (IRT), in which item difficulty is established *independent of participant abilities*, and so item and scale information eliminate the dependence on statistical reliability. IRT is also capable of clarifying the extent of discrimination between two participant groups, that is, “to differentiate between individuals at different trait levels” (Morizot, Ainsworth, & Reise 2007). As Kline (2005) and Funk & Rogge (2007) note, given a large sample size for a group of well-correlated items, the standard error of the mean per item converges more rapidly in IRT than in CTT, indicating that IRT is generally more precise than CTT in determining the psychometric trait.

Item response theory (IRT) is a statistical theory that distinguishes the *latent trait* (designated “ability”) of a participant from the difficulty of a set of items with well-correlated response patterns. IRT methodology assumes *unidimensionality* among the items in the set of items, that is, a group of items are assumed to have well-correlated response patterns such that the difficulty of each item can be reasonably compared. As noted by McDonald (1999) and Morizot, Ainsworth, & Reise (2007), unidimensionality is the property that the items in the set are *locally independent*, meaning, vary only in difficulty. They further note that the items themselves should have enough common variance to give reasonably unbiased estimates of item difficulty. They state that in general, an unbiased analysis for dichotomously-scored items (those with two possible response codes, e.g., 0 or 1) may have as few as 100 participants, whereas 5-point response formats require a sample size of at least 500 participants.

The 1 and 2 parameter logistic models

We begin our technical overview of IRT with a dichotomous model, in which responses are tabulated based on the frequency of responding in favor of some concept being probed in the study. We can construct a logistic curve model for the probability distribution of the responses (Wallace & Bailey 2010, Morizot, Ainsworth, & Reise 2007, McDonald 1999, and references therein). Logistic curves indicate the probability of obtaining at least a certain score in respondents with different trait levels. The idea is that the logistic curves are related to an individual item, so they are often called item characteristic curves.

In the one-parameter logistic (1PL) model, or Rasch model, the probability for participant p whose ability level is θ to obtain a score X_{pj} (typically taken to be 1 for a correct answer and 0 for an incorrect answer) for a dichotomously scored item j of difficulty b_j is given by

$$P(X_{pj} = 1|\theta, b_j) = (1 + e^{-(\theta - b_j)})^{-1}. \tag{5}$$

The plot of $P(X_{pj} = 1|\theta, b_j)$ vs. θ is called the item characteristic curve (ICC). The ICC in Figure 7 has the parameter $b_j = 0.8$, such that at the ability level $\theta = 0.8$, participant p has a 50% chance of responding with the correct answer. This axis is scaled on a z-score metric such that its mean and standard deviation are respectively 0 and 1 (Morizot, Ainsworth, & Reise 2007).

We should note that term “ability” in IRT literature is somewhat arbitrary (Morizot, Ainsworth, & Reise 2007). The 1PL is used to determine the probability of responding with the correct answer, as a function of some latent trait of the participants. An example of a latent trait used by Thorpe et al. (2007) in their study of common irrational beliefs is to graph the 1PL as a function of the participants’ irrationality, in which case “irrationality” would be the label for the θ axis.

In the 1PL model, the shape of the probability distribution is the same for each item in a group. That is to say, the *spread* of the probability of right vs. wrong answers (responding in favor of vs. not endorsing the desired characteristic) is assumed to be the same for each item. Some items will more sharply discriminate higher and lower-achieving participants than other items. In Figure 8, the steepness, or discrimination, at $b_j = 0.8$ of the second ICC (in red) is 3 times higher than that of the previous item (in blue).

In order to incorporate changes in the probability distribution shape among the grouped items, we use a two parameter logistic (2PL) model, as suggested by Wallace & Bailey (2010), with a discrimination parameter a_j that measures the steepness of the ICC. In the 2PL model, the probability of responding in favor of the probed trait is given by

$$P(X_{pj} = 1|\theta, a_j, b_j) = (1 + e^{-a_j(\theta - b_j)})^{-1}. \tag{6}$$

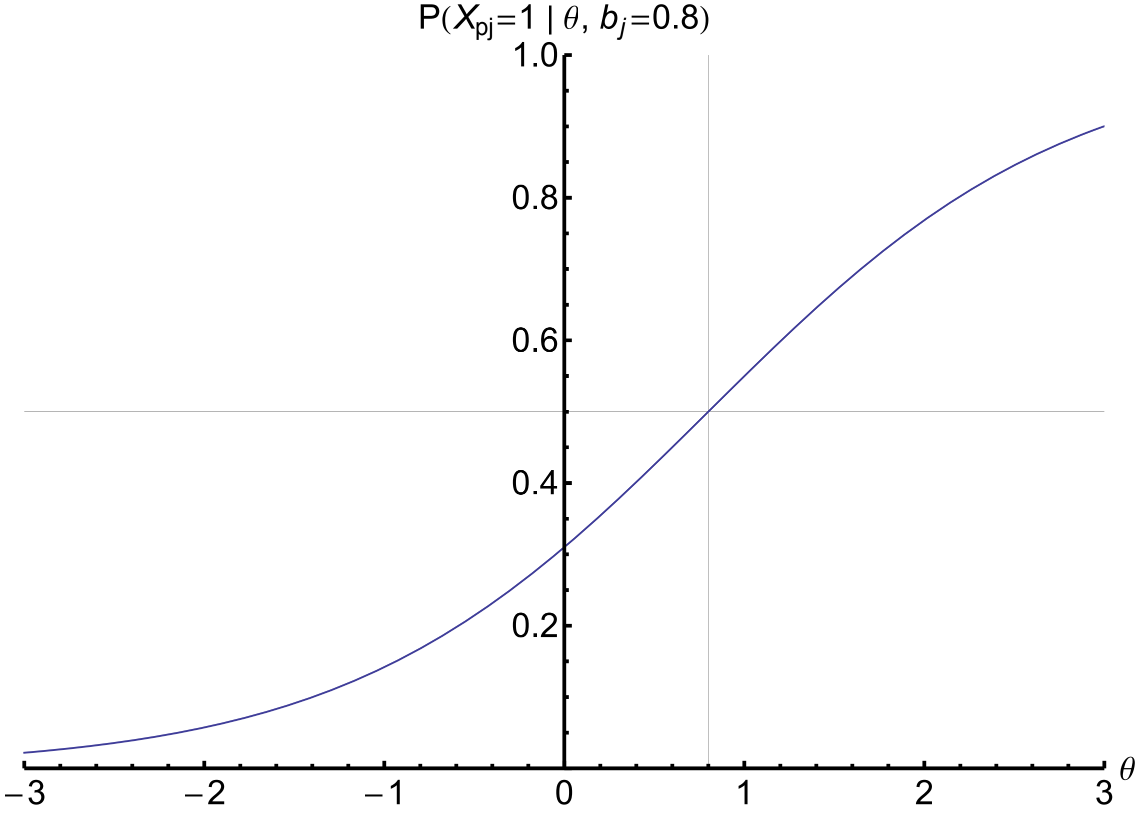


Figure 7: Graph of an ICC with the parameter $b_j = 0.8$. ***Participants with a higher ability level than*** b_j ***have more than a 50% chance of responding with the correct answer***

A high value for a_j corresponds to high item discrimination, which indicates a strong division between higher and lower-achieving participants for the item. Values of $a_j \to 0$ correspond to a broader mixing of participant trait levels. There further exists a 3-parameter logistic (3PL) model, which incorporates a guessing parameter c_j for right or wrong answers. The 3PL model applicable on, e.g., multiple choice tests, in which for n options $c_j = 1/n$. In the 3PL model, the probability of responding in favor of the probed trait is given by

$$P(X_{pj} = 1|\theta, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta - b_j)}}. \tag{7}$$

The graded response model

While the 1PL and 2PL models are used in IRT for dichotomous scoring, a polytomous model allows for more than 2 responses. Hence, there will be more than one "characteristic curve" to plot the probability of responding with a particular score. The more proper term to use to describe each curve, as noted by Reeve and Fayers (2005), is *category response curve* (CRC). We will thus hereafter refer to the curves as CRCs.

To perform an IRT analysis on polytomous scoring, one may choose from a number of models (e.g. Edelen & Reeve 2007). Of these models, the graded response model (GRM), first introduced by Samejima (1969), is relevant if we intend to secure the order of participant

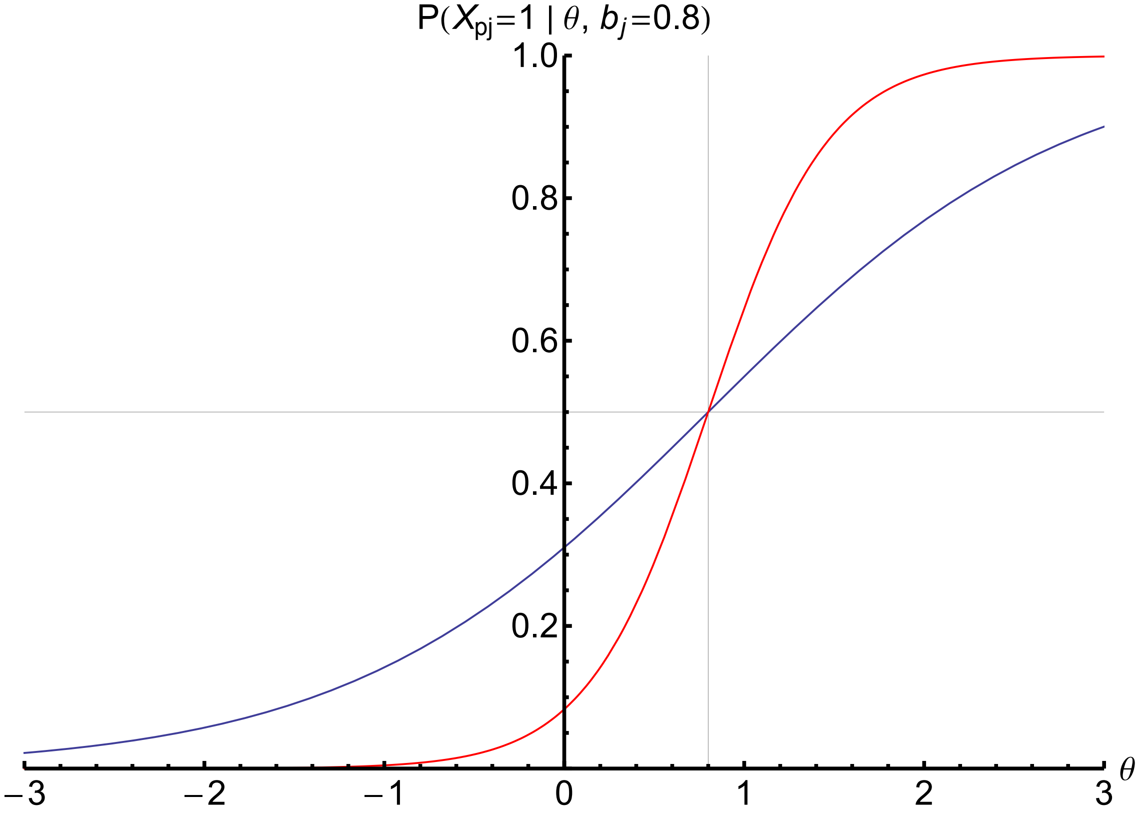


Figure 8: ***Graph of two ICCs, each with the parameter*** $b_j = 0.8$. ***The red ICC discriminates between higher and lower-ability participants 3 times more than the blue ICC***

responses throughout our analysis. In the GRM, the total probability of any response is normalized to 1, that is,

$$\sum_{k=1}^{K} P_{jk}(\theta) = 1 \tag{8}$$

for K scores. One can thus plot all of the CRCs on the same graph, as in Figure 9.

Because of probability conservation, $K - 1$ CRCs have a defined parameter, and we follow the convention introduced by Samejima to chose the locations of all but the first CRCs. Hence, in Figure 9, the second through fifth CRCs, indicated by purple, blue, green, and yellow, have the respective parameters $b_{jk} = -2.0, 0.0, 1.5, 2.5$, whereas the first CRC, in red, has no parameter. We specifically note that instead of just b_j, we now require that b have two subscripts, one for item j, and one for the particular score k for $K - 1$ unique parameters.

Following the construction by Samejima (1969), in which k goes from 1 to K, the 2PL function for the highest score, P_{jK}, is given by

$$P_{jK}(\theta) = 1 + e^{-a_j(\theta - b_{jK})})^{-1}. \tag{9}$$

The 2PL for the $K - 1$th score is then

$$P_{j,K-1}(\theta) = (1 + e^{-a_j(\theta - b_{j,K-1})})^{-1} - P_{jK}(\theta). \tag{10}$$

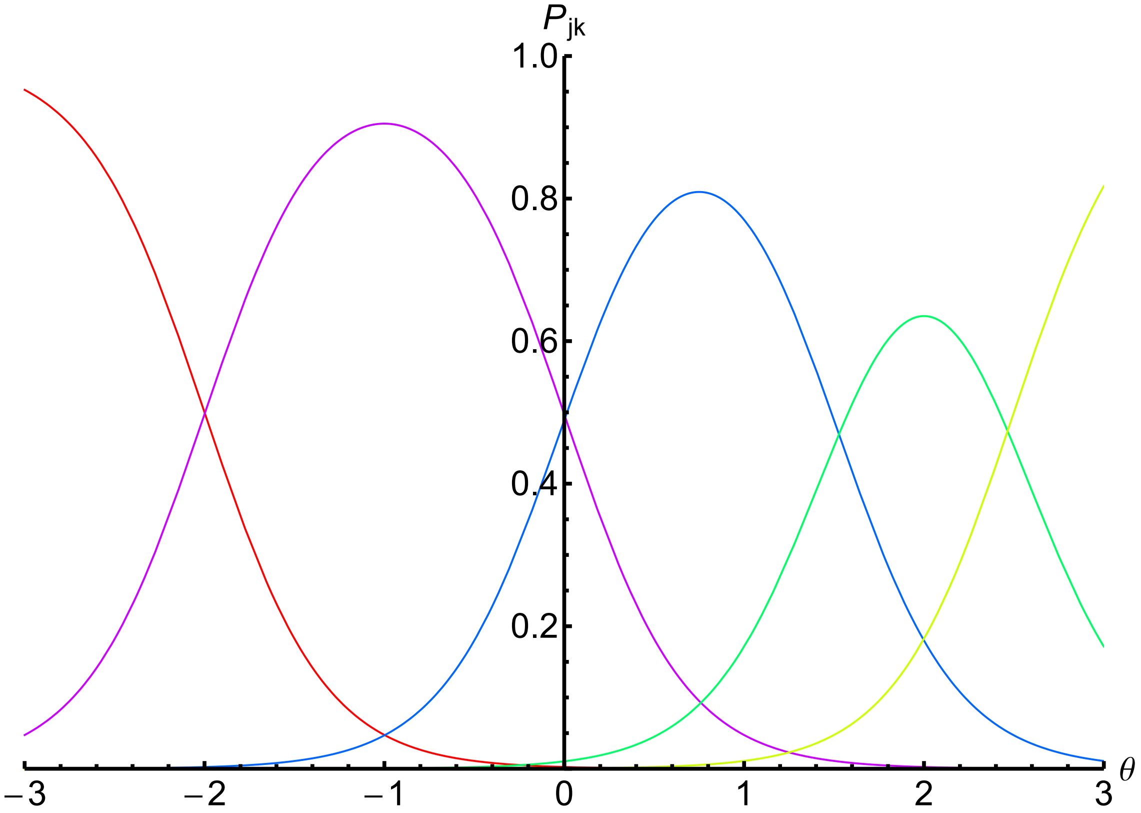


Figure 9: Graph of five 2PL CRCs according to Samejima's graded response model, with the second through fifth CRC having the parameters $b_{jk} = -2.0, 0.0, 1.5, 2.5$. ***In Samejima's graded response model, the parameter of the first CRC is determined by the other CRCs in the plot***

To conserve probability, the 2PL for the $K-2$th score is then

$$P_{j,K-2}(\theta) = (1 + e^{-a_j(\theta - b_{j,K-1})})^{-1} - P_{j,K-1}(\theta) - P_{jK}(\theta). \tag{11}$$

This pattern continues until we get to the lowest score, which is

$$P_{j1}(\theta) = 1 - P_{j2}(\theta) - \ldots - P_{jK}(\theta). \tag{12}$$

Samejima introduces the notation

$$P^{+}_{jk}(\theta) = P_{j,k+1} + P_{j,k+2} + \ldots + P_{jK}, \tag{13}$$

that is, $P^{+}_{jk}(\theta)$ is the sum of all CRCs from $P_{j,k+1}$ up to P_{jK}. Comprehensively, using the graded response model, the probability of responding with score k for each item is

$$P_{jk}(\theta) = \begin{cases} 1 - P^{+}_{jk}(\theta) & \text{for} \quad k = 1, \\ (1 + e^{-a_j(\theta - b_{jk})})^{-1} - P^{+}_{jk}(\theta) & \text{for} \quad 2 \leq k \leq K-1, \\ (1 + e^{-a_j(\theta - b_{jK})})^{-1} & \text{for} \quad k = K. \end{cases} \tag{14}$$

Item information

In Fisher information theory, the item information is statistically the variance of the score (Lehmann & Casella 1998); more conceptually, that information is a relative measure of how reliable the value of an CRC is at θ, or how well each score is being estimated at θ. Hence, in IRT literature, one typically refers to an item as being "most informative" (or "best estimating" each score) where the item information curve peaks. Away from these peaks, where $I(\theta)$ is lower, the scores are not estimated very well, and so the reliability of the CRC values for the score decreases with information (Baker 2001). We thus seek a relationship between the CRCs and the information curves for a group of items. Extending Fisher information theory to polytomously-scored items, the item information curve for item j is given (Chajewski & Lewis 2009) by

$$I(\theta) = \sum_{k=1}^{K} \frac{1}{P_{jk}(\theta)} \left(\frac{dP_{jk}(\theta)}{d\theta} \right)^2 . \tag{15}$$

Figure 10 shows the total information $I(\theta)$ for the example in Figure 9.

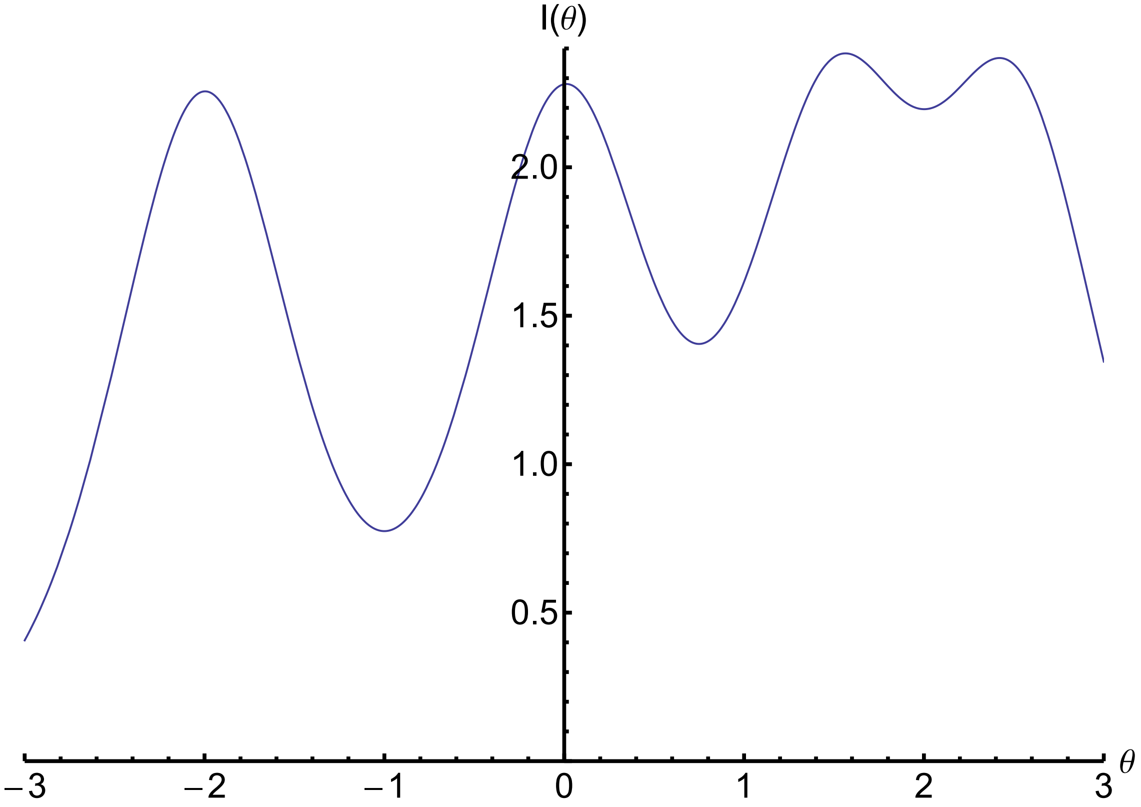


Figure 10: Graph of the total information for the five CRCs in Figure 9

Local information maxima are obtained by finding the abilities θ_s that satisfy

$$\left.\frac{dI(\theta)}{d\theta}\right|_{\theta=\theta_s} = \sum_{k=1}^{K} \frac{d}{d\theta} \left(\frac{1}{P_{jk}(\theta)} \left(\frac{dP_{jk}(\theta)}{d\theta} \right) \right) \Bigg|_{\theta=\theta_s} = 0, \tag{16}$$

where θ_s marks the trait location of a transition from one score to another. A quick approximation for θ_s between two neighboring CRCs with scores k and $k+1$ can be found by estimating the coordinates of the information peaks. Our estimation assumes that the contribution to $I(\theta)$ from each of the other CRCs is negligible, in other words, for another score n, $P_{jn} \approx 0$. Our approximation thus applies to both the red-purple and purple-blue CRC intersections of Figure 9, which represent the first two information peaks in Figure 10.

If, in Eqn. (16), we assume that two neighboring CRCs overlap with $P_{jn} \approx 0$, then

$$\frac{d}{d\theta}\left(\frac{1}{P_{jk}(\theta)}\left(\frac{dP_{jk}(\theta)}{d\theta}\right)^2\right) = -\frac{d}{d\theta}\left(\frac{1}{P_{j,k+1}(\theta)}\left(\frac{dP_{j,k+1}(\theta)}{d\theta}\right)^2\right). \tag{17}$$

After taking the derivative and setting $\theta = \theta_s$, one has, on the left-hand side,

$$\frac{dP_{jk}(\theta_s)}{d\theta_s}\left(\frac{2}{P_{jk}(\theta_s)}\frac{d^2P_{jk}(\theta_s)}{d\theta_s^2} - \frac{1}{P_{jk}^2(\theta_s)}\left(\frac{dP_{jk}(\theta_s)}{d\theta_s}\right)^2\right)$$

and, on the right-hand side,

$$-\frac{dP_{j,k+1}(\theta_s)}{d\theta_s}\left(\frac{2}{P_{j,k+1}(\theta_s)}\frac{d^2P_{j,k+1}(\theta_s)}{d\theta_s^2} - \frac{1}{P_{j,k+1}^2(\theta_s)}\left(\frac{dP_{j,k+1}(\theta_s)}{d\theta_s}\right)^2\right).$$

The only way for the left and right sides of the equation to be equal is if

$$P_{jk}(\theta_s) = P_{j,k+1}(\theta_s), \quad \text{and} \quad \frac{dP_{jk}(\theta_s)}{d\theta_s} = -\frac{dP_{j,k+1}(\theta_s)}{d\theta_s}. \tag{18}$$

But the identification $P_{jk}(\theta_s) = P_{j,k+1}(\theta_s)$ is also where two CRCs intersect. Therefore, if the contribution to $I(\theta)$ from each of the other CRCs is negligible, then *maximum information is obtained at the intersection of CRCs.*

Note that for polytomous scoring, up to $K-1$ peaks at the solutions θ_s may be clearly identified from the shape of $I(\theta)$. When the parameters b_{jk} for two neighboring CRCs match more closely, their information peaks can either form a plateau or converge into a single peak in the plot of $I(\theta)$ vs. θ. For example, if, instead of our example parameters being $b_{jk} = -2.0, 0.0, 1.5, 2.5$, we had $b_{jk} = -2.0, 0.0, 1.5, 2.3$, the total information curve would look like that in Figure 11, in which the right-most information peaks appear to merge into a plateau, whereas the first two peaks corresponding to $b_{jk} = -2.0, 0.0$ are not affected. Or, if we had $b_{jk} = -2.0, 0.0, 1.5, 1.9$, the two right-most information peaks would merge into one, as in Figure 12, while, again, the first two peaks would remain unaffected.

The significance of the "plateau" and merged peaks is as follows: near the ability level θ where this occurs, there exist two almost-simultaneous transitions between neighboring scores, i.e., from $k-1 \to k$ and $k \to k+1$. This suggests that around this θ, there exists a net transition from $k-1 \to k+1$, where participants are not very likely to respond with score k. Such a merger suggests overlapping responses between two audience groups at that ability level.

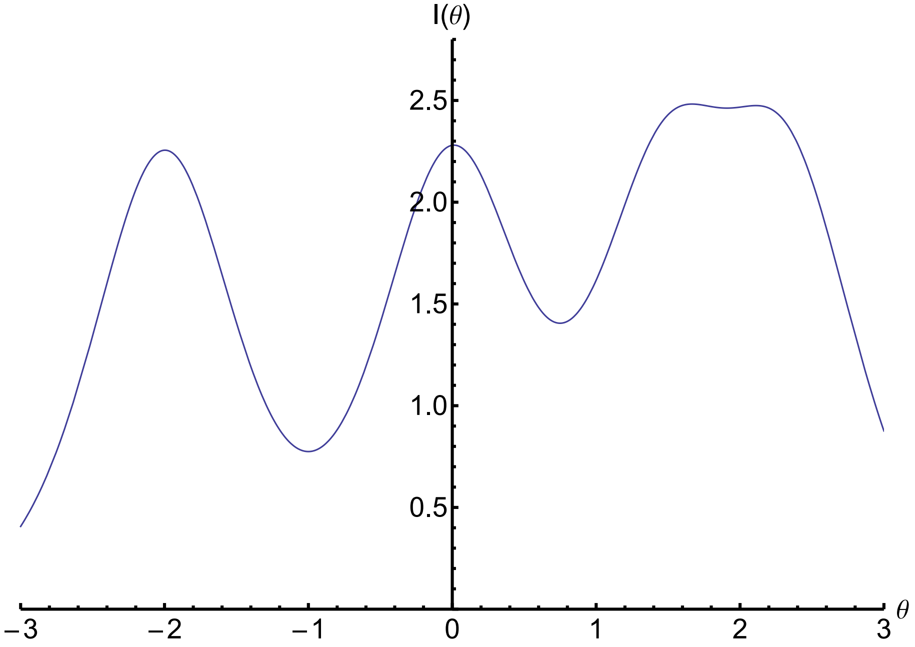


Figure 11: Graph of the total information for a variation of the five CRCs in Figure 9, except with $b_{jk} = -2.0, 0.0, 1.5, 2.3$

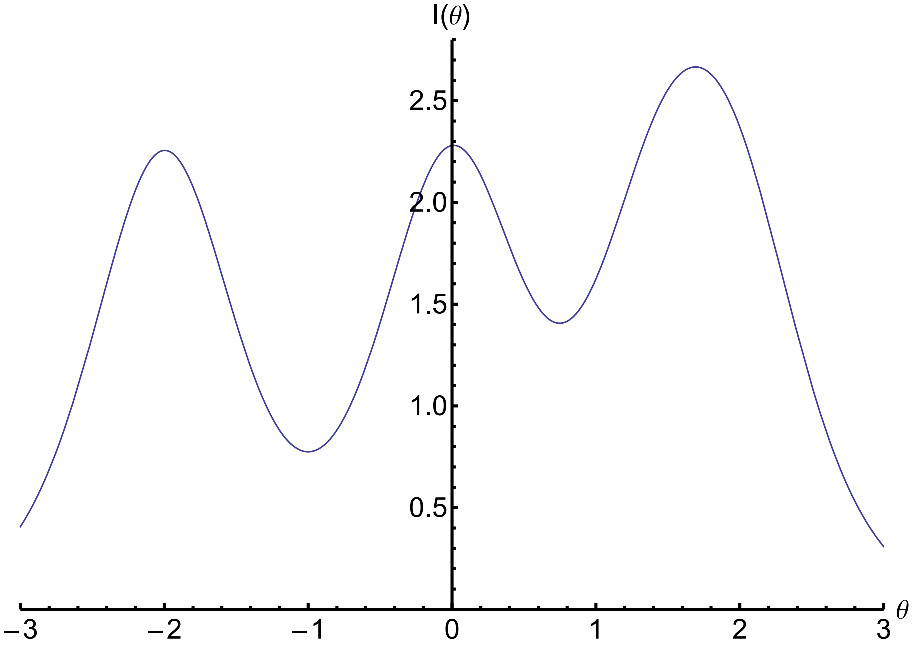
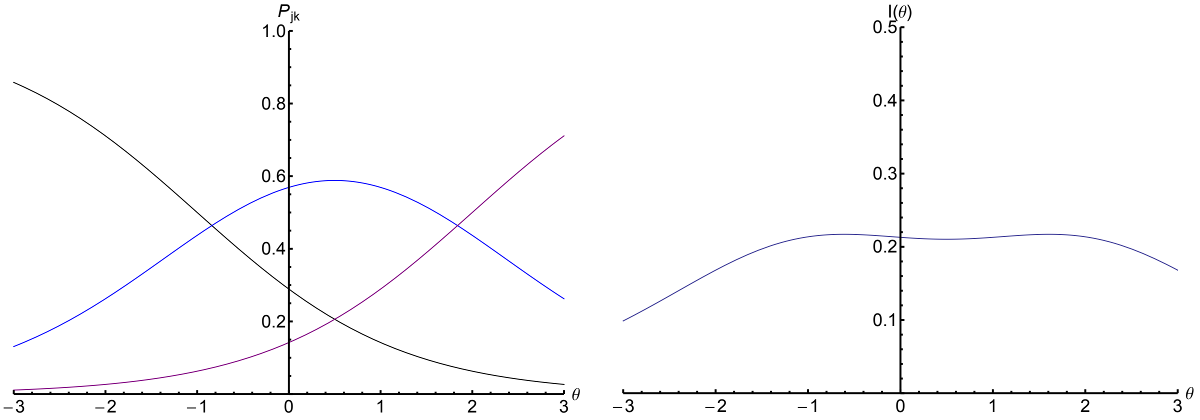


Figure 12: Graph of the total information for a second variation of the five CRCs in Figure 9, except with $b_{jk} = -2.0, 0.0, 1.5, 1.9$

An interesting case may arise when the discrimination parameter $a_j < 1$. Consider a three-score polytomous model, where $a_j = 0.9$, $b_{jk} = -1.0, 2.0$. The left panel of Figure 13 shows the CRCs for these parameterized logistic curves, while the right panel shows the representative total information curve. Because the discrimination parameter a_j is associated with the steepness of the CRCs, an item with a lower discrimination parameter implies higher mixing of individuals who respond with scores "1" vs. "2" or "2" vs. "3." Hence, we say that items with a low a_j (typically < 1) are generally less informative. But because information is inversely related to the standard error of the mean, items that are not very informative are also poor fits to the GRM. Hence, items that have low values for their information curves are not reliable estimates of the CRC intersection coordinates.



(a)

(b)

***Figure 13:* *Graph of three 2PL CRCs and the total information curve for a graded response model with the parameters* $a_j = 0.9,\ b_{jk} = -1.0, 2.0$**

SUMMARY AND COMMENTS

This paper has provided an outline of some of the essentials of IRT, with illustrative examples of data analysis with selected software programs. It is recommended that readers consult the references and other sources for more advanced coverage. This is a work in progress. Please contact the authors with your suggestions and corrections, which will be acknowledged in any future iterations of this document.

Author Notes and Acknowledgements

Geoffrey L. Thorpe, Ph.D., Professor, Department of Psychology Room 301, 5742 Little Hall, University of Maine, Orono, ME 04469; phone (207) 581-2743; e-mail: geoffrey.thorpe@umit.maine.edu

Andrej Favia, B.A., doctoral student, Department of Physics and Astronomy, Bennett Hall, University of Maine, Orono, ME 04469; email: andrej.favia@umit.maine.edu

The authors gratefully acknowledge the constructive comments of Steve Balsis, Ph.D., Neil F. Comins, Ph.D., Michael A. Robbins, Ph.D., and Richard M. Ryckman, Ph.D. However, this manuscript has not been formally peer-reviewed.

REFERENCES

Baker, F. B. (2001). *The basics of item response theory.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Original work published in 1985. Retrieved from `http://echo.edres.org:8080/irt/baker/`

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory – Second Edition Manual.* San Antonio, TX: The Psychological Corporation.

Bernard, M. E. (1998). Validation of the General Attitude and Belief Scale. *Journal of Rational-Emotive and Cognitive-Behavior Therapy*, *16*, 183 – 196.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd. ed.). Mahwah, NJ: Erlbaum.

Chajewski, M. & Lewis, C. (2009). Optimizing item exposure control algorithms for polytomous computerized adaptive tests with restricted item banks. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*

David, D., Lynn, S. J., & Ellis, A. (Eds.) (2010). *Rational and irrational beliefs: Research, theory, and clinical practice.* New York: Oxford.

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

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Edelen, M. O. & Reeve, B. R. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res*, *16*, 5.

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

Frumkin, I. B. (2010). Evaluations of competency to waive Miranda rights and coerced or false confessions: Common pitfalls in expert testimony. In G. D. Lassiter & C. A. Meissner (Eds.), *Police interrogations and false confessions: Current research, practice, and policy recommendations* (pp. 191-209). Washington, DC: American Psychological Association.

Funk, J. L., & Rogge, R. D. (2007). Testing the ruler with item response theory: Increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of Family Psychology*, *21*, 572-583.

Grisso, T. (1998). *Instruments for assessing understanding and appreciation of Miranda rights.* Sarasota, FL: Professional Resources.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000/2007). *Item response theory and health outcomes measurement in the 21st century.* NIH Public Access Author Manuscript. Retrieved 04/30/2010 from `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1815384/`

pdf/nihms14476.pdf

Jobling, D., & Snell, E. J. (1961). The use of the coefficient of reproducibility in attitude scaling. *The Incorporated Statistician*, *11*, 110-118.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation*, Thousand Oaks, CA: Sage.

Lee, K, & Ashton, M. C. (2007). Factor analysis in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 424-443). New York: Guilford.

Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation*, New York: Springer-Verlag.

Liao, T. F. (2006). Series editor's introduction. In R. Ostini & M. L. Nering, *Polytomous item response theory models* (pp. v-vi). Thousand Oaks, CA: Sage.

Linacre, J. M. (2005). Correlation coefficients: Describing relationships. *Rasch Measurement Transactions*, *19:3*, 1028-1029. Retrieved from http://www.rasch.org/rmt/rmt193c.htm

Linacre, J. M. (2006). *WINSTEPS* (Version 3.61.2) [Computer Software]. Chicago: Winsteps.com

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

McDonald, R. P. 1999, *Test Theory: A Unified Approach*, Mahwah, N.J.: Lawrence Earlbaum.

Miranda v. Arizona, 384 U.S. 336 (1966).

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 407-423). New York: Guilford.

Mueller, W. (nd). *Exploring precalculus.* Retrieved from http://www.wmueller.com/precalculus/index.html

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Nandakumar, R., & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology in the Social Sciences* (pp. 93-105). Thousand Oaks, CA: Sage.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* Thousand Oaks, CA: Sage.

Owings, L. R. (2010). *Factors related to high ratings on the implied latent trait of incompetence in Maine archival competency evaluations.* Unpublished doctoral dissertation, University of Maine.

Owings, L. R., Thorpe, G. L., McMillan, E. S., Burrows, R. D., Sigmon, S. T., & Alley, D. C. (2012). *Scaling irrational beliefs: Using modern test theory methodology to evaluate the General Attitude and Belief Scale.* Manuscript submitted for publication.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In P. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods of practice* (2nd ed). New York: Oxford University Press, 2005. p. 55–73.

Rogers, R. (2011). Getting it wrong about Miranda rights: False beliefs, impaired reasoning, and professional neglect. *American Psychologist*, *66*, 728-736.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*, 282-292.

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

Samejima, F. (2004). Graded response model. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 77-82). New York: Academic Press.

Sansone, R. A., Wiederman, M. W., & Sansone, L. (1998). The Self-Harm Inventory (SHI): Development of a scale for identifying self-destructive behaviors and borderline personality disorder. *Journal of Clinical Psychology*, *54*, 973-983.

Thissen, D., Chen, W-H, & Bock, R. D. (2003). *MULTILOG 7 for Windows: Multiple category item analysis and test scoring using item response theory* [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Thorpe, G. L., McMillan, E., Sigmon, S. T., Owings, L. R., Dawson, R., & Bouman, P. (2007). Latent trait modeling with the Common Beliefs Survey: Using item response theory to evaluate an irrational beliefs inventory. *Journal of Rational- Emotive & Cognitive-Behavior Therapy*, *25*, 175-189. doi: 10.1007/s10942-006-0039-9

Thorpe, G. L., Owings, L. R., McMillan, E. S., Burrows, R. D., & Orooji, B. A. (2011). *A brief test of elementary legal knowledge for pretrial competency evaluations: Item selection and preliminary development of the LAWTEST14.* Manuscript in preparation.

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

Turk, D. C., Dworkin, R. H., Burke, L. B., Gershon, R., Rothman, M., Scott, J., et al.

(2006). Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations. *Pain*, *125*, 208-215. doi:10.1016/j.pain.2006.09.028

Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, *9*, 010116.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside.

REFERENCES OF RELATED INTEREST

Hoyle, R. H. (2007). Applications of structural equation modeling in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 444-460). New York: Guilford.

Krane, W. R., & Slaney, K. L. (2005). A general introduction to the common factor model. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 125-151). Mahwah, NJ: Erlbaum.

Miller, L. A., McIntire, S. A., & Lovler, R. L. (2011). *Foundations of psychological testing: A practical approach* (3rd. ed.). Thousand Oaks, CA: Sage.

Roesch, R., Zapf, P. A., & Hart, S. D. (2010). *Forensic psychology and law.* Hoboken, NJ: Wiley.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353.