Social and Personality Psychology Compass 3/4 (2009): 507–529, 10.1111/j.1751-9004.2009.00194.x

An Introduction to Item Response Theory Using the Need for Cognition Scale

Michael C. Edwards*
The Ohio State University

Abstract

This paper provides an introduction to two commonly used item response theory (IRT) models (the two-parameter logistic model and the graded response model). Throughout the paper, the Need for Cognition Scale (NCS) is used to help illustrate different features of the IRT model. After introducing the IRT models, I explore the assumptions these models make as well as ways to assess the extent to which those assumptions are plausible. Next, I describe how adopting an IRT approach to measurement can change how one thinks about scoring, score precision, and scale construction. I briefly introduce the advanced topics of differential item functioning and computerized adaptive testing before concluding with a summary of what was learned about IRT generally, and the NCS specifically.

Many of the constructs psychologists are interested in studying are not directly observable. Examples include depression, intelligence, extroversion, and need for cognition. To study these constructs, researchers will often assemble a set of items which are believed to be related to the construct of interest. This set of items (also known as a scale, or questionnaire) can be viewed as one possible operationalization of the construct of interest. To the extent that responses to the items are related to the construct we intend to assess, the observed item responses can provide information about the unobserved construct. For this to be possible, we must have some model (often called a measurement model) to relate the observed response to the underlying construct.

There are many measurement models which could serve as a link between the observed and latent worlds. One long standing measurement model is true score theory, which is part of classical test theory (CTT). CTT is perhaps most strongly associated with reliability coefficients such as coefficient alpha (Cronbach, 1951) and has a long history in psychology. As useful as CTT can be, recent advances in latent variable modeling have provided access to measurement models which are more flexible and more powerful than CTT.

The remainder of this paper is dedicated to describing one such latent variable measurement model – item response theory (IRT). In the next section, I briefly describe the recurring example before reviewing two of the more widely used IRT models in psychological research. After this, I discuss assumptions of the IRT model and then describe how the IRT framework can change how we think about different aspects of measurement. Before concluding with a summary of what has been covered, I mention two advanced topics which may be of particular interest to psychological researchers.

Recurring Example

Throughout this paper, I will use the 18-item version of the Need for Cognition Scale (NCS; Cacioppo, Petty, & Kao, 1984) as a recurring example. The NCS was originally developed as a 34-item scale (Cacioppo & Petty, 1982) which has been widely used to measure need for cognition.[1] As described by Cacioppo and Petty (1982), the need for cognition is 'the tendency for an individual to engage in and enjoy thinking' (p. 116). Of the 18 items, Items 3, 4, 5, 7, 8, 9, 12, 16, and 17 are reverse scored so that higher summed scores indicate higher levels of need for cognition. The items have a 5-point response scale ranging from 'extremely uncharacteristic of me' to 'extremely characteristic of me'. The sample used for these analyses consists of 3364 subjects drawn from over 30 different studies conducted at a large midwestern university between 2002 and 2007. The subjects are primarily undergraduates and roughly 60% are female. All software syntax used in the analyses described here can be found at my website.[2] I felt it would be most useful if data were available as well, but it is not possible for me to post the data I used for these analyses. As a compromise, I simulated 3000 subject's worth of data from the IRT parameters provided in Table 1. This data file is also available on my website.

Two Popular IRT Models

Despite the name, IRT is not really a theory so much as a collection of models. There is tremendous variety contained within the term IRT, but the bulk of these models are non-linear latent variable models which attempt to explain the process by which individuals respond to items. The basic idea in IRT is that an observed item response (e.g., choosing the category 'strongly agree' on a 5-point Likert scale) is a function of person properties and item properties.

There are dozens of IRT models, many of which were developed for specialized circumstances and are not generally applicable. Even ignoring these sorts of models, there are still many potential models to choose from. In the following two sections I review two of the IRT models

Table 1 Graded response model parameter estimates for the Need for Cognition Scale.

Item	*a*	b_1	b_2	b_3	b_4
1	1.65	–2.04	–0.55	0.31	2.14
2	1.88	–2.34	–0.96	–0.25	1.61
3	1.67	–2.35	–0.91	–0.15	1.43
4	1.68	–2.49	–0.91	–0.17	1.55
5	1.43	–2.98	–1.18	–0.32	1.67
6	1.18	–1.81	–0.24	0.78	2.82
7	1.34	–2.47	–0.59	0.05	1.99
8	0.68	–3.11	0.08	1.45	4.16
9	1.07	–2.17	0.17	1.04	3.05
10	1.39	–3.26	–1.60	–0.51	1.59
11	1.63	–2.62	–1.17	–0.42	1.63
12	1.41	–3.06	–1.26	–0.51	1.41
13	1.17	–1.92	–0.34	0.87	3.17
14	1.35	–2.58	–1.22	–0.30	1.56
15	1.56	–2.69	–1.02	–0.16	1.62
16	0.78	–2.70	–0.26	0.73	3.41
17	1.10	–2.91	–0.80	–0.01	1.86
18	0.63	–4.45	–1.80	–0.30	3.59

Note. a represents the slope parameter estimates and *b* represents the threshold parameter estimates.

which are likely to be useful to psychological researchers. Both models deal with ordered response categories, which are very common in psychological research.

The two-parameter logistic model

One widely used model is the two-parameter logistic model (2PLM), which is appropriate for dichotomous observed responses. The 2PLM is written as[3]

$$P(x_j = 1 \mid \theta) = \frac{1}{1 + \exp[-a_j(\theta - b_j)]}, \qquad (1)$$

where x_j is the observed response to item j, a_j is the slope parameter for item j, b_j is the threshold parameter for item j, and θ is the construct being measured (which is typically assumed to follow a standard normal distribution, more on this later). Slopes, also known as discrimination parameters, contain information about how related a particular item is to the construct being assessed. The higher the slope, the more variability in item responses is attributable to differences in the latent construct. IRT slopes are mathematically related to factor loadings and in general,

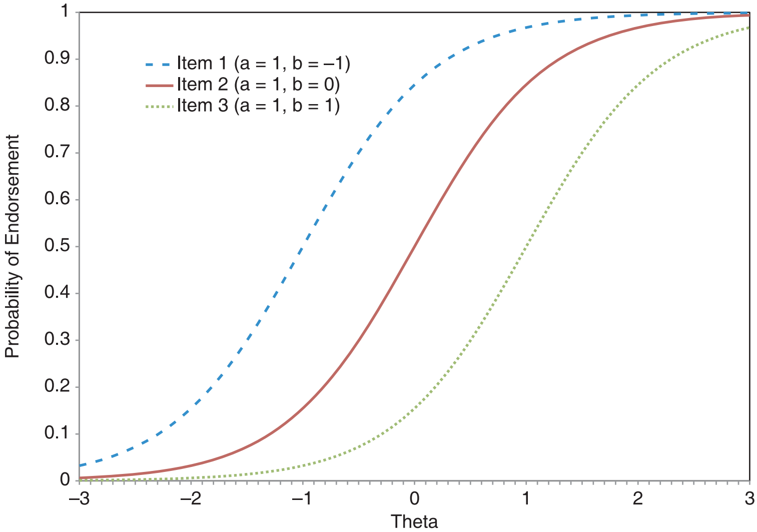

Figure 1 2PLM trace lines for three dichotomous items. The three items all have a slope of 1, but vary in their threshold parameter. Item 1 has a threshold of –1, Item 2 has a threshold of 0, and Item 3 has a threshold of 1.

intuitions about factor loadings will also hold for IRT slopes (although the metrics are different). The threshold parameter (also called difficulty or severity depending on the context) indicates the point along the latent continuum where an individual would have a 50% chance of endorsing a particular item. The higher the threshold, the higher an individual must be on the latent trait to have a 50% chance of endorsing that item.

IRT models have a long tradition of being represented graphically. Trace lines, or item characteristic curves, are the visual manifestation of Equation 1. Six such trace lines are presented in Figures 1 and 2 to help illustrate the impact of the slope and threshold parameters. The three items portrayed in Figure 1 all have the same slope, but different thresholds. The dashed (blue) item has a threshold of −1, the solid (red) item has a threshold of 0, and the dotted (green) item has a threshold of 1. To have a 50% chance of IRT endorsing these three items, individuals would have to be one standard deviation below the average, at the average, and one standard deviation above the average, respectively. Changes in threshold parameters shift the trace lines left and right horizontally. The further to the right an item is, the less often it will be endorsed. In the realm of education, such an item is said to be very difficult. In psychopathology, an item with a high threshold is said to be severe. In the context of personality research, an item with a high threshold would be an extreme statement which would require a high level of the construct to be endorsed. For example, if one was measuring extraversion, an item like ‘I

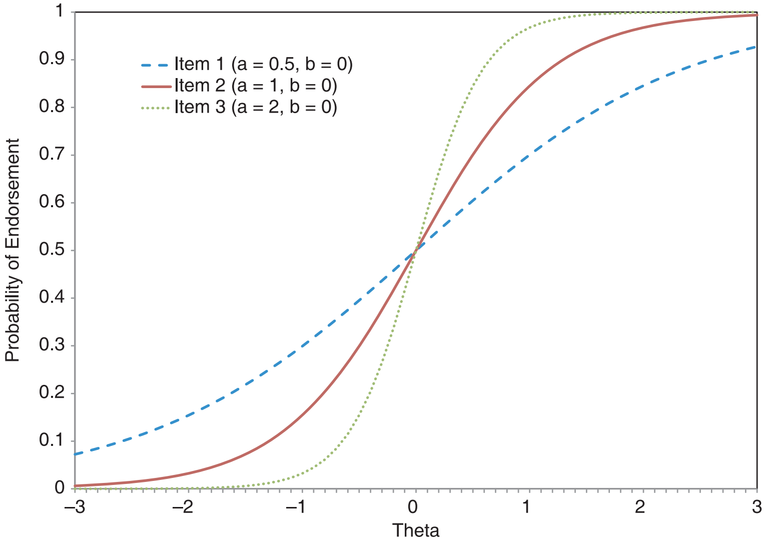

Figure 2 2PLM trace lines for three dichotomous items. The three items all have a threshold of 0, but vary in their slope parameter. Item 1 has a slope of 0.5, Item 2 has a slope of 1, and Item 3 has a slope of 2.

enjoy hanging out with friends' would require a lower level of extraversion to endorse than 'I need constant social contact to survive'. This would result in the latter item having a higher threshold.

Figure 2 contains trace lines for three items which all have the same threshold value (0), but different slopes. The dashed (blue) item has a slope of 0.5, the solid (red) item has a slope of 1, and the dotted (green) item has a slope of 2. The higher the slope, the more quickly the probability of endorsement changes in the region of the threshold. The slope parameter is also known as the discrimination parameter as it contains information about how well a response to a particular item discriminates between individuals above or below the threshold value for that item. The impact of the slope can be seen by examining the change in probability of endorsement that occurs on two items (with different slopes) for a fixed change on the latent construct. Let's look at a one standard deviation increase from –0.5 to 0.5. For the item with the slope of 0.5, an individual who is half a standard deviation below the mean has a 40% chance of endorsing the item. An individual who is half a standard deviation above the mean has a 60% chance of endorsing the item. In this case, a shift of one standard deviation (from –0.5 to 0.5) resulted in a 20% increase to the probability of endorsement. If we repeat the same exercise for the item with a slope of 2 we see that the probability of endorsement goes from 15% to 85%. The same one standard deviation increase now corresponds to a 70% increase in the probability of endorsement. Put another way, individuals who are above and below the threshold value for a high slope item are more likely to behave differently than is the case for a low slope item.

The graded response model

Many items use a response set with more than two options. In psychological research, the 5-point Likert response set is ubiquitous. The graded response model (GRM; Samejima, 1969) is an extension of the 2PLM which can accommodate more than two categories. The GRM could theoretically be used with any number of categories, but restrictions in current software (e.g., Multilog) limit the number of possible categories to 36. Typical choices seen in the literature range from three to ten response categories. The GRM is.

$$P(x_j = c \mid \theta) = \frac{1}{1 + \exp[-a_j(\theta - b_{cj})]} - \frac{1}{1 + \exp[-a_j(\theta - b_{(c+1)j})]}, \tag{2}$$

where c indexes response category and all other parameters are as previously defined. The GRM has $C - 1$ estimated threshold parameters, where C is the number of response alternatives.[4]

As can be seen in Equation 2, the probability of choosing any particular category c involves calculating a difference between two different 2PLMs. The standard 2PLM represents the probability of endorsing an item as a function of the latent construct. For an item with five response categories, there are four thresholds, corresponding to four different 2PLMs:

1. b_1 is the threshold for the trace line describing the probability of choosing category 2, 3, 4, or 5.
2. b_2 is the threshold for the trace line describing the probability of choosing category 3, 4, or 5.
3. b_3 is the threshold for the trace line describing the probability of choosing category 4 or 5.
4. b_4 is the threshold for the trace line describing the probability of choosing category 5.

To determine the probability that someone will choose category 2, we subtract the probabilities dictated by the trace line defined by b_2 from those dictated by the trace line defined by b_1.

Although it is possible to plot the 2PLM trace lines described above, it is more common to plot the response probabilities for the individual categories. As with Figures 1 and 2 for the 2PLM, it is possible to see the impact of different threshold or slope values in GRM trace lines. Figure 3 displays the impact in the GRM of thresholds shifting in Items 10 and 13 from the NCS. The top panel contains the trace line for Item 10 (The idea of relying on thought to make my way to the top appeals to me) and the

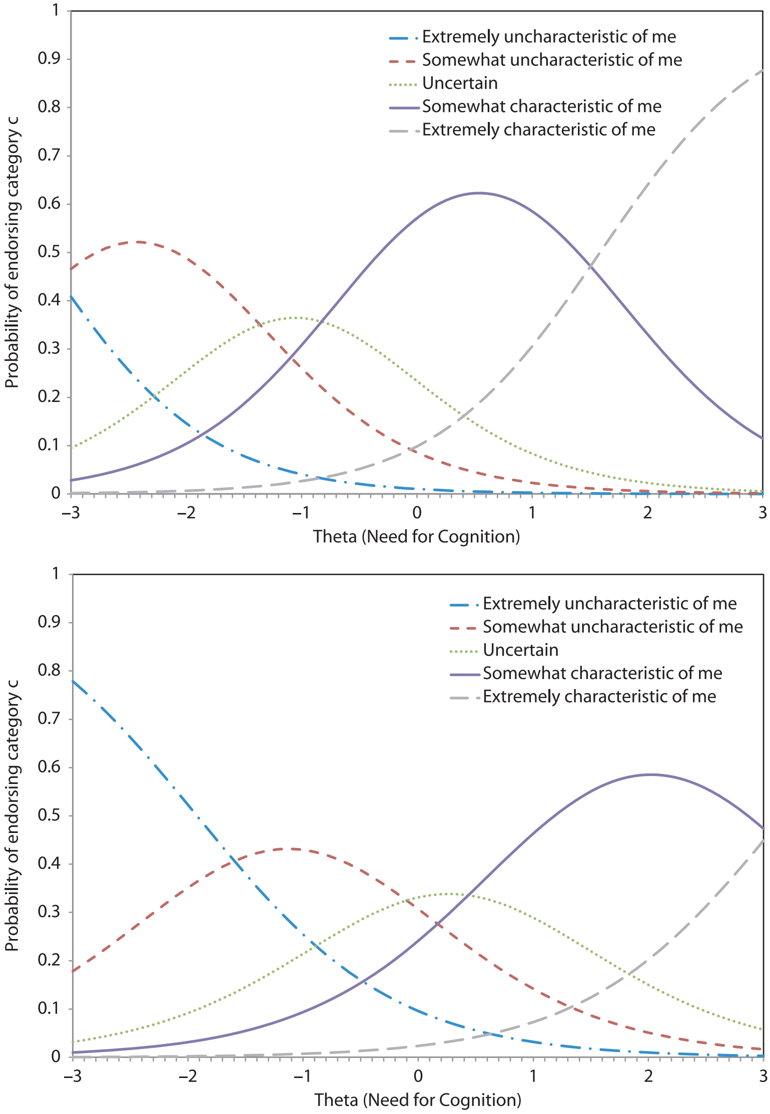

Figure 3 GRM trace lines for Items 10 and 13 from the NCS. The item parameters for these two items are given in Table 1. The two panels illustrate the impact of thresholds in the GRM. Item 10 (in the top panel) has lower thresholds, which is reflected in the horizontal shift in the response functions to the left.

bottom panel contains the trace line for Item 13 (I prefer my life to be filled with puzzles I must solve). An examination of the threshold parameters in table 1 reveals that the set of threshold parameters for Item 13 are higher than those for Item 10. This can be seen in the horizontal shift to the right evident in the trace line for Item 13. An individual with an average level of need for cognition is most likely (about 57%) to choose category four ('somewhat characteristic of me') for Item 10. To have a 57% chance of choosing category four for Item 13, an individual would have to be 1.7 standard deviations above the average level of need for cognition in the population.

The impact of the slope parameter can be seen in Figure 4. Item 2 (I like to have the responsibility of handling a situation that requires a lot of

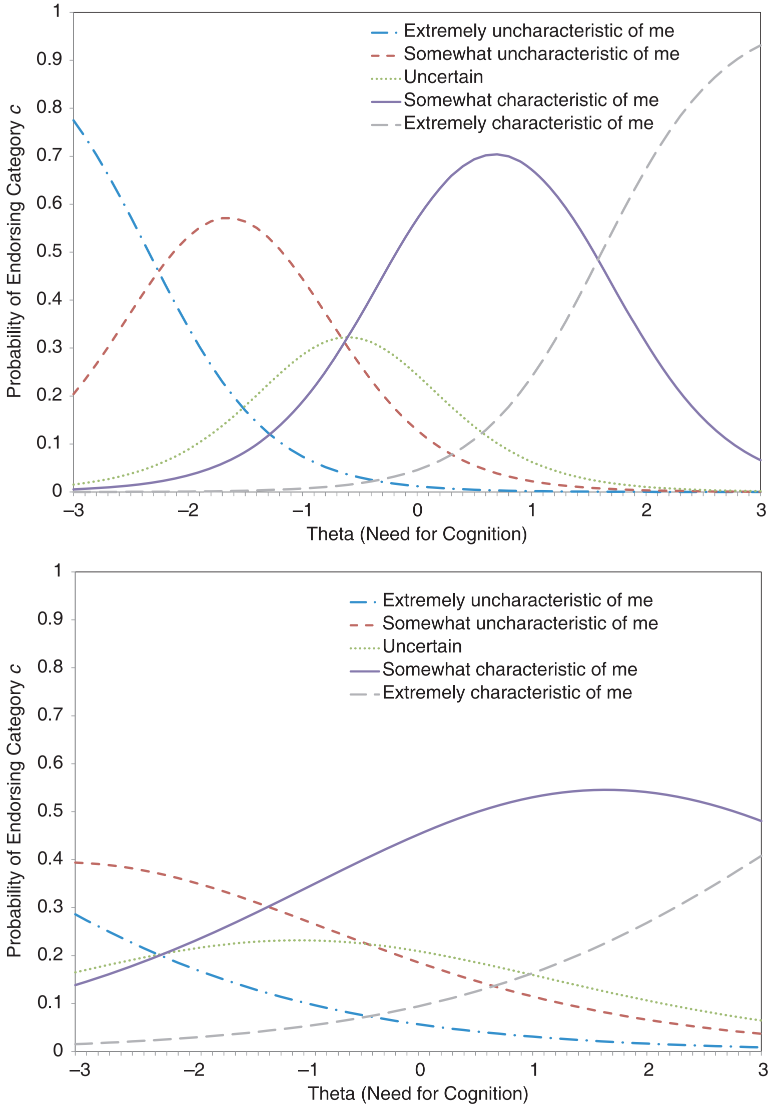

Figure 4 GRM trace lines for Items 2 and 18 from the NCS. The item parameters for these two items are given in Table 1. The two panels illustrate the impact of slopes in the GRM. Item 2 (in the top panel) has a higher slope, which is reflected in the peakedness of the response functions.

thinking) is presented in the top panel and Item 18 (I usually end up deliberating about issues even when they do not affect me personally) in the bottom. Item 2 has a slope of 1.88 and Item 18 has a slope of 0.63. The higher the slope parameters, the more peaked the individual response curves become. A more peaked response curve indicates that the probability of choosing a particular category is changing more quickly as you move higher or lower on the latent construct. At any given level of need for cognition, the response curve with the highest probability is the most likely to be chosen. In this manner, a respondent's choice among the various response categories tells us something about their level of need for cognition. Choosing 'somewhat characteristic of me' for Item 2 tells us that the respondent is most likely between −0.6 and 1.5 on the need for cognition latent dimension. Choosing that same category for Item 18 suggests that the respondent is somewhere between −1.3 and 3.3 (which is off the right side of the figure). The smaller probable range based on the response to Item 2 is an indication that this item provides more information (both colloquially and, as we shall see in a later section, statistically) than Item 18.

As with other statistical models, the 2PLM and GRM make certain assumptions about the nature of the data. These assumptions are reviewed in the next section, along with discussions of how to assess the extent to which they are met.

Assumptions

The 2PLM and GRM make four major assumptions: Unidimensionality, local independence, monotonicity, and a normally distributed latent trait. Unidimensionality and local independence are closely related and for the 2PLM and GRM, the presence of one implies the presence of the other. Unidimensionality means that one and only one common latent variable is being measured. Local independence means that, conditional on the latent variable, item responses are independent. If a set of items is unidimensional, then the only reason item responses should be related to one another is the latent variable. Once the variance accounted for by the latent variable is removed, the residuals should be uncorrelated. In practice, no data (barring simulations) will ever be truly unidimensional. The focus of assumption check regarding dimensionality is whether or not a single factor provides a reasonably approximation to the observed data. There are a number of methods available to assess the extent to which a unidimensional model is plausible. One popular approach is to use a combination of exploratory and confirmatory factor analysis (EFA and CFA, respectively) to assess dimensionality. Although some estimation difficulties exist due to the categorical nature of the data (see Wirth & Edwards, 2007), many of the major software packages that are used for EFA and CFA can correctly estimate model parameters in the presence

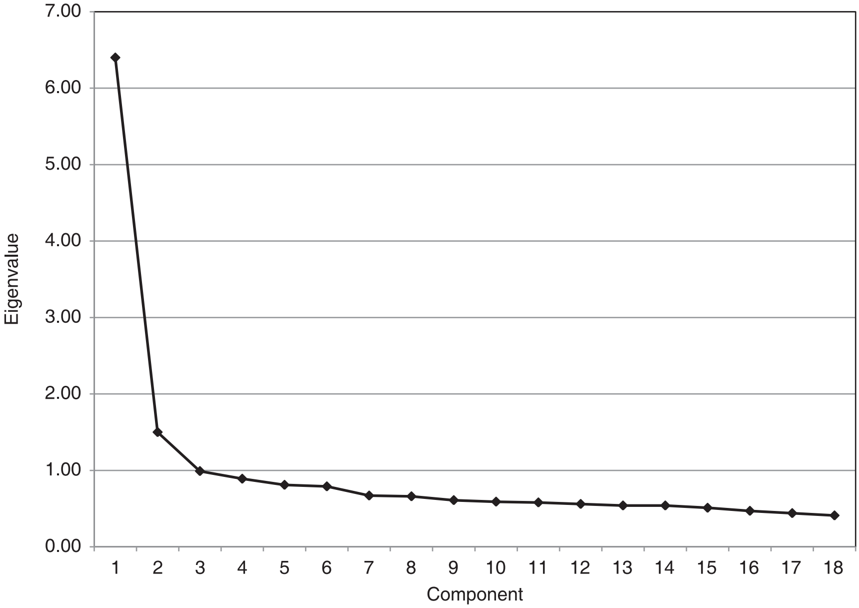

Figure 5 Scree plot from an EFA of the 18-item NCS.

of categorical data. The vector of Eigenvalues from the EFA (and any resulting factor loading matrices) and measures of fit from CFA are often considered sufficient evidence of unidimensionality.

An EFA and CFA were performed on the NCS prior to conducting the IRT analysis reported in this paper. The EFA was conducted using the CEFA software package (Browne, Cudeck, Tateneni, & Mels, 2004), which can properly handle categorical data via polychoric correlations and ordinary least squares estimation. The resulting scree plot[5] is shown in Figure 5 and suggests that there is one primary dimension. There is a slightly elevated second component, which may or may not cause a problem. In cases such as this I try to extract a second factor in EFA to see if it is interpretable prior to moving on to the CFA. A two-factor model was estimated and the loadings were rotated using an oblique quartimax rotation. The second factor consisted entirely of reverse coded items and was correlated 0.72 with the first factor. A correlation of this magnitude suggests that the constructs are closely related, but not completely overlapping. Thus, the EFA suggests that one- or two-factor models may be plausible.

I then estimated a one-factor CFA using LISREL (Jöreskog & Sörbom, 2003). Polychoric correlations were analyzed and the diagonally weighted least squares (DWLS) estimator was used. This combination corrects for the categorical nature of the data and provides valid indices of fit. There are a wide variety of fit indices provided by structural equation modeling-based CFA software, but I have come to rely on the root mean square error of approximation (RMSEA), the comparative fit index (CFI), the goodness of fit index (GFI), and the root mean square residual (RMSR). There is no 'silver bullet' for gauging fit – the goal should be to present evidence to support any claims that a model sufficiently accounts for the observed data. With the NCS, the one-factor CFA appears to fit the data reasonably well (RMSEA = 0.06, CFI = 0.96, GFI = 0.98, RMSR = 0.05). The debate about what constitutes 'good fit' is ongoing, but based on the findings of Hu and Bentler (1999), the NCS results suggest a one-factor model is plausible. Considering the one-factor model showed good fit, I did not fit any two-factor models.

In the event that a one-factor model was not plausible, there are several possible remedies. If the scale is intended to measure one construct, then the evidence from any dimensionality analyses can be used to alter the scale accordingly. It is often possible to remove a few items from the analysis and achieve plausible unidimensionality. Note that the items do not necessarily have to be removed from the scale, but just not included in the IRT analysis. Alternately, one could use a multidimensional IRT (MIRT) model, which also goes by the name confirmatory item factor analysis (Wirth & Edwards, 2007). MIRT models are more complex to estimate than unidimensional models, but efforts are underway which should make user-friendly software available in the next few years.

It is worth noting that violations of unidimensionality can occur in forms both big and small. In some instances, two major constructs are being assessed rather than one. In other instances, two items are so similar that respondents treat them as the same question (called surface local dependence, Chen & Thissen, 1997). Both cases result in departures from unidimensionality, but the former is often much easier to detect than the latter.

Monotonicity is another important assumption for many IRT models. In the 2PLM, for example, monotonicity implies that (for items with positive slopes), the probability of endorsement never decreases as you move to higher levels of the latent construct. The 2PLM is part of the parametric family of IRT models, which implies that the response functions follow some specific functional form (in this case, a logistic curve). It is possible that this logistic function may not be appropriate for some items and/or responses. There are several methods available for assessing monotonicity (Liang, 2007; Molenaar, 1997; Ramsay, 2000), which all involve some form of non-parametric IRT models. Non-parametric models, as opposed to parametric models, do not assume any particular functional form. The non-parametric models can be used to assess the extent to which monotonicity is a reasonable assumption. Several of the approaches have only been implemented for dichotomous data, although the MPS5 software package (Molenaar & Sijtsma, 2000) can accommodate polytomous data.

The last major assumption I'll talk about is that the latent trait being measured is normally distributed. This is an assumption of statistical convenience, which makes the parameter estimation simpler. Since IRT is a latent variable model, we must also set scaling constraints, similar to those imposed in CFA. In IRT, it is common to set the scale by fixing the mean and variance of the latent trait to 0 and 1, respectively.

New work by Woods (2006) provides a method to examine the extent to which this is a plausible assumption as well as correct for departures from normality if they are detected. This work also suggests that estimation of parameters from standard IRT models (e.g., the 2PLM and GRM) is reasonably robust to departures from normality in the latent variable. Although it is likely that extreme departures from normality in the latent trait will be more common in areas of psychology dealing with psychopathology, it is worth bearing in mind when assessing any construct.

Thinking Differently about Scales

After conducting the dimensionality assessment detailed above, I performed an IRT analysis using the GRM as implemented in Multilog (Thissen, 1991). The resulting parameter estimates are provided in Table 1. Once these estimates are obtained, several additional facets of IRT become accessible. In the next sections, I cover three such facets: scoring, score precision, and scale construction.

Scoring

The item parameters in Table 1 tell us quite a bit about how the items are related to need for cognition as well as the levels of need for cognition required to choose any particular response category. The trace lines presented in Figures 3 and 4 are merely graphical representations of the item parameters. Any information conveyed in the trace lines is present in the item parameters. If you take the item parameters in Table 1 as reflections of what is going on in the data, then using a summed score or proportion score to provide an individual's score begins to make less sense. For instance, the item parameters suggest that the fourth threshold (b_4) is higher for Item 13 than Item 10 (which was shown in Figure 3). This means that choosing 'extremely characteristic of me' for Item 13 indicates a higher level of need for cognition than choosing that same response category for Item 10. In a summed score or proportion score, both responses would be weighted the same. Slope parameters play an important role in the response pattern weighting as well. Item 2 has a slope which is about three times as high as Item 18. There are many different ways to interpret a difference between slopes, but one useful interpretation is that the response to Item 2 is more likely due to the a subject's level of need for cognition than the response to Item 18. That is to say Item 18 has a weaker relationship to need for cognition, suggesting that more of the observed variability is due to things other than need for cognition (e.g., error) than in an item with a higher slope. While Item 18 still contributes some information to our knowledge about an individual's need for cognition, it does not contribute as much as Item 2.

In the IRT framework, it is possible to create scores for individuals which reflect the estimated item parameters. IRT scale scores weight the observed response patterns using the item parameters. These scores have a number of desirable properties. First, if the construct is assumed to follow a standard normal distribution, the IRT scale scores are in a standard normal metric. This means that all our knowledge about the standard normal distribution can be brought to bare. For instance, someone with an IRT scale score of one is one standard deviation above average and we would expect roughly 84% of the sample to have lower scores and 16% to have higher scores. A 0.2 difference between two individuals (or the means of two groups) can be directly interpreted a difference of two tenths of a standard deviation.

Second, IRT scale scores are more variable than summed scores. Any particular summed score (or proportion score) can be arrived at a variety of ways. For instance, a summed score of two on a scale with eight dichotomous items can be arrived at 28 different ways. In the IRT framework, it is possible that each of those 28 different response patterns could receive different scores. An example of the potential for IRT scores to increase variability is provided in Figure 6. In the sample used for these

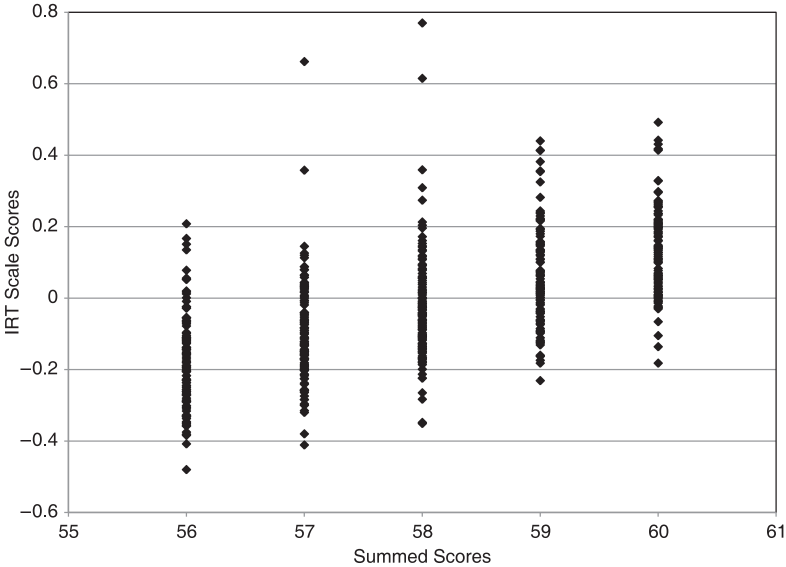

Figure 6 IRT scale scores corresponding to NCS summed scores from 56 to 60.

analyses, the average summed score on the NCS was approximately 58. Figure 6 shows the IRT scales scores which correspond to scores within two points of (and including) the average. Within the average score in this sample we see that there is a full standard deviation's worth of variability in IRT scores. For the 130 individuals in this sample who received a score of 58, we are completely unable to distinguish them from one another with respect to their need for cognition. This is not to say that IRT scales scores will differentiate everyone. If two individuals have the same response pattern they will receive the same IRT scale score. However, for the 130 individuals with a summed score of 58, there are only nine instances where two individuals have the same response pattern (and hence the same IRT scale score) and two instances where three individuals have the same response pattern. The remaining 106 individuals received unique IRT scale scores, which provides a much greater level of differentiation.

Lastly, if the proper steps are taken during item parameter estimation (commonly called *calibration*), IRT scale scores from different subsets of items are still directly comparable. This process, called equating (see, e.g., Kolen & Brennan, 2004; pp. 201–208), allows for tremendous flexibility in designing and administering scales. For example, if we had item parameters for the original 34 items of the NCS (assuming the assumptions described above still held), any subset of that scale could be used without loss of comparability. This is not to say that the different versions would have identical measurement properties (these would depend on the items making up the subset), but that they are both still estimating the same quantity (an individual's latent need for cognition score) on the same metric. Non-IRT methods exist for equating, but these are total-score focused methods that require the existence of two scales (and the data) which one wants to equate. For a more detailed discussion of IRT-based equating, and the advantages of such an approach, see Cook and Eignor (1991) and Edwards and Wirth (2009).

Imagine if we created two 17-item NCS scales from the original 34 items. Also imagine that we have performed an IRT calibration on the original 34 items (and that all assumptions were satisfied). Now suppose that by accident there are more items indicative of higher levels of need for cognition in the second set of 17 items than the first. In terms of summed scores, this would result in lower summed scores on the 17-item subset including the higher need for cognition items. However, if we use IRT scoring, this information is contained in the item parameters and can be used to correct for this imbalance. Another application of equating is the ability to choose scale lengths based on purpose (which I will say more about in the next section). Using equating, it would be possible to have small (nine items), medium (18 items), and large (34 items) versions of the NCS yet still maintain comparability of scores across forms.

Score precision

Traditional definitions of reliability focus on one-number summaries meant to convey information about how reliable scores from a particular scale are. Coefficient alpha (Cronbach, 1951) and the standard error of measurement are two such indicators of score reliability. Both indicators suggest that all scores produced by a scale are equally reliable or equally precise. While this might make sense if all items were equally reliable and equally severe, it is my experience that this is generally not the case. If we acknowledge that items do differ in these respects, then it is reasonable to suppose that not all scores are equally reliable. For instance, if a scale consists of items which assess low levels of a construct, then our estimate for someone who is high on that construct will be less reliable than our estimate for someone who is low on that construct.

As previously mentioned, we gain information about individuals based on their responses to items and the properties of those items. The statistical concept of information (i.e., Fisher information), helps us understand where a given scale provides more or less information about individuals. Each item has an information function, calculated based on that item's parameters. Three item information functions are presented in Figure 7, each corresponding to one of the sample 2PLM items in Figure 1 or Figure 2. The long-dash (red) line and dotted (green) line are test information functions for two items with the same slope, but different thresholds. Item information functions for the 2PLM will have

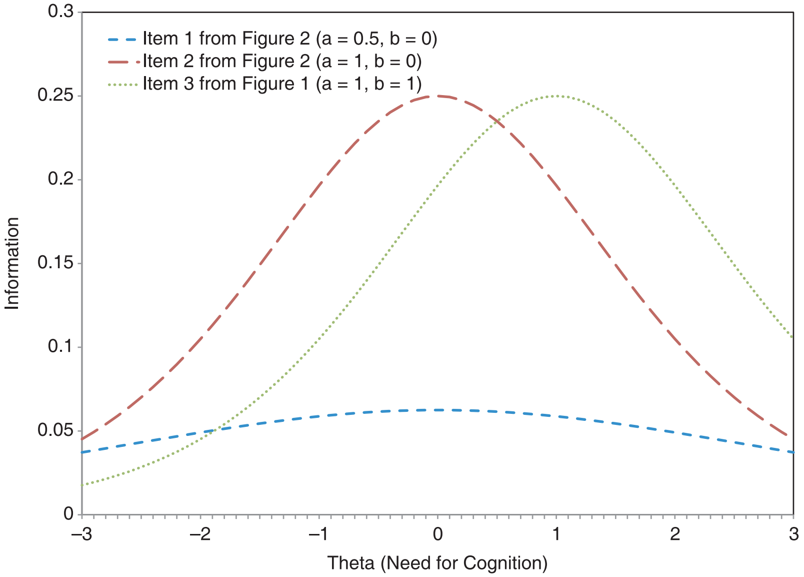

Figure 7 Item information functions for three of the six items displayed in Figures 1 and 2.

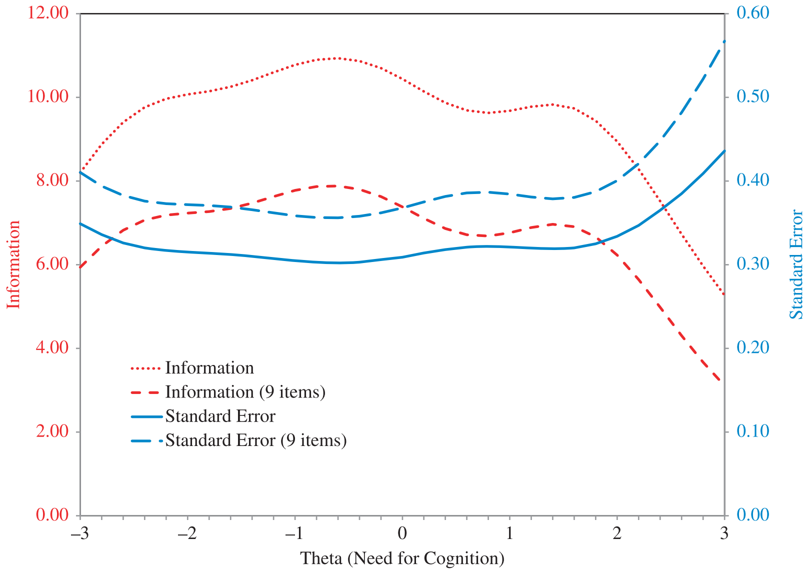

Figure 8 Test information function and standard error curves for the 18-item NCS and a shortened 9-item version.

their maximum value at the value of the threshold. Moving the threshold up or down would simply move the item information function right or left on the *x*-axis in Figure 7. The long-dash line and the dashed (blue) line are test information functions for two items with the same threshold, but different slopes. Slopes control how peaked the item information function is. The higher the slope value, the more information that item provides around the threshold.

To understand how the test is functioning as a whole, we can combine the item level information functions (which are additive) to create a test information function (TIF). This is a nice feature of IRT which supports a 'bottom-up' view of reliability. A scale can provide reliable scores to the extent that the items comprising the scale provide useful information about individuals in the region of that score. The TIF for the 18-item NCS is the dotted (red) line in Figure 8. The metric of information is not particularly interpretable, but the curve allows us to make relative statements about where a scale provides more or less information. For the NCS, we can see that it provides relatively uniform information about individuals between −2 and 2 standard deviations, with a reasonably sharp decline past those points in either direction.

The square root of the inverse of information ($\sqrt{1/INF}$) at a given level of the latent construct provides the standard error which would be attached to that particular score. This is helpful for two reasons. First, the concept of standard error is more widely known than that of information, so researchers tend to be more comfortable talking about things in terms of standard errors. Second, since the IRT scale scores are in a standard normal metric, so too are the standard errors. The solid (blue) line in Figure 8 is the standard error curve (SEC) for the 18-item NCS. Given the relationship to the TIF, the general story will be the same. However, we can now say that an average score on the NCS will have a standard error of 0.31, which is just under a third of a standard deviation.

Scale construction

The possibility of equating, combined with the notion of variable precision just discussed, provides new ideas about scale construction. These ideas can be tremendously valuable, but they do require us to to think more carefully about what we're trying to accomplish by administering a particular scale. Questions to be considered include:

- Who are we trying to measure?
- How reliable do the scores need to be?
- How small a change are we interested in detecting?
- What is the primary purpose of these scores?

The answers to these questions can be converted to a target TIF, which can serve as a blueprint for the subsequent scale construction process. For example, if you were constructing a scale meant to assess high levels of need for cognition then you would want the TIF to be high for high values of need for cognition, but you might not care much about the level of the TIF for those with below-average need for cognition. A test that was meant to classify individuals as being above or below some cut point would look very different from one meant to broadly assess a construct. The former would consist of items with threshold values near the cut point (resulting in a very peaked TIF near the cut point) while the latter would result in a nearly uniform distribution of item thresholds (resulting in a flat TIF across the range of interest). For example, if a scale was intended to identify clinically depressed individuals, there would be no reason to include items that are frequently endorsed (i.e., have a low threshold) by individuals who are not clinically depressed.

The topics covered in this section represent standard parts of the IRT models described in this paper. IRT scale scores, TIFs, and SECs are easily obtainable from standard IRT software such as Multilog. In the next section I review two advanced topics which build on the properties of item response models covered in this section.

More Advanced Topics

The topics included in this section are covered in more detail in Edwards and Edelen (2009), but are presented here to give the reader a sense for some of the possibilities once inside the IRT framework. Even the coverage in Edwards and Edelen (2009) is sparse, as there are entire books on both topics (Holland & Wainer, 1993; Wainer, 2000, for example). My hope is that if a particular topic resonates with a reader they can refer to the chapter, which makes a more concerted effort to point the reader towards relevant literature.

Differential item functioning

Differential item functioning (DIF), is the IRT equivalent of measurement invariance in the factor analytic and structural equation modeling literatures. The essential idea is that in some instances, the same item response may mean different things in different groups. What classifies as 'groups' varies widely in DIF studies, but examples include male versus female, native English versus native Spanish speakers, and young versus old. An item is said to exhibit DIF if, after conditioning on any latent differences between groups, the item behaves differently in the groups. This different behavior is manifested in the item parameters, so in essence an item exhibiting DIF has different item parameters depending on which group a respondent is in.

One classic example of gender DIF involves depression scales and items about crying. Several studies (Cole, Kawachi, Mailer, & Berkman, 2000; Schaeffer, 1988) have found that the crying item on the CES-D (Radloff, 1977) is a more severe indicator of depression for men than it is for women. This is equivalent to saying that the crying item has a higher threshold when a male is answering then when a female is answering. This is important to know because if this difference isn't somehow modeled, then there could be bias in the resulting scores (i.e., men who endorse this question will appear less depressed then they are and women who endorse it will appear more depressed then they are). Aside from the possibility that the DIF itself can be of interest, once it is detected it is possible to pursue a course of action to correct for the presence of DIF. There are a wide variety of methods used to assess DIF (both CTT- and IRT-based) and for an excellent overview of these methods see Teresi (2006).

Computerized adaptive testing

A second advanced topic that warrants mentioning is computerized adaptive testing. A computerized adaptive test (CAT) is one where the items are administered adaptively based on the recorded responses. If you have taken the Graduate Record Exam (GRE) some time in the past decade then you have first-hand experience with a CAT. A CAT will administer an item, observe the response, and then based on that response update the current estimate of proficiency (which in many cases is the

IRT scale score described above). It is then possible to choose an item which is, in some sense, most useful for learning more about an individual with that level of the construct. The CAT will then continue asking questions until a pre-determined level of reliability is achieved for the estimate of the person's latent trait or a pre-determined number of questions has been asked.

For a CAT to be possible there must be a set of items (called an item bank) for which there are already IRT parameters. By virtue of the equating mechanism mentioned previously, it is possible to selectively administer items to individuals but still obtain scores which are comparable across individuals. As long as an item bank is available, it is even possible to provide comparable scores when two individuals being assessed have no items in common. This is not possible with CTT-based equating methods. The technology of adaptive testing provides the most efficient way to assess a construct. It can also enable uniformly reliable scores without drastically increasing the length of the scale. By using adaptive testing, researchers could assess more constructs in the same time period or achieve their original goals in less time.

Conclusion

In this final section I'll review what we've learned about the NCS, about IRT in general, and suggest some resources for those interested in learning more about IRT.

What we've learned about the Need for Cognition Scale

In the course of using the NCS as a recurring example throughout this paper we've learned quite a bit about it. First, there is evidence to suggest that the 18-item NCS is reasonably well explained by one factor. This allowed us to perform a unidimensional IRT analysis, which offers additional information about the NCS. All 18 items contribute useful information to the total score, but some contribute more than others. The items cover a broad range of need for cognition and as seen in Figure 8 they do so relatively uniformly between two standard deviations below the mean and two standard deviations above the mean. IRT scale scores in this region will have a standard error near 0.31, which corresponds roughly to a reliability of 0.9. Information drops fairly quickly above 2 (less quickly below −2), suggesting that if there was improvement to be made in the NCS it would involve trying to add items indicative of higher levels of need for cognition (this assumes that a fairly uniform TIF is desired). In contrast, CTT would tell us that coefficient alpha for the NCS is 0.87 and the standard error of measurement is 4.01. While these numbers convey an average sense of what is happening with the NCS, they lose much of the information contained in Figure 8.

The dashed line (red) and long-dash line (blue) show what would happen to the TIF and SEC (respectively) if a new 9-item NCS was created by selected the nine items with the highest slopes. The shapes of the TIF and SEC are similar between the 18-item NCS and the 9-item 'high slope' NCS, but the TIF drops (there is less information from nine items) and the SEC rises (there scores are less precise when estimated from nine items). The standard errors for scores between −2 and 2 are near 0.37 for the 9-item NCS, which corresponds to a reliability of about 0.86. So, if we were willing to live with a reliability of 0.86 for most scores versus a reliability of 0.9, we could shorten the NCS to nine items.

What we've learned about IRT

Beyond the NCS, we've also explored several important aspects of IRT. IRT can provide item-level weights which may more accurately reflect researchers' operationalizations of a construct. This item-level weighting leads to IRT scale scores which are more variable than summed scores, allowing for finer distinctions to be made between individuals. IRT scale scores can also be directly compared even when different subsets of items are used as long as some equating work can be done. This ranges from creating short forms which still retain comparability to their longer parent form to CATs, where some individuals may have no items in common. The notion of reliability is somewhat different in the IRT framework, focusing more on information provided and how precise scores at various levels of the construct would be. We also saw that by thinking about scale development in the IRT context it is possible, through a target TIF, to construct a scale to meet a set of measurement goals. Once inside the IRT framework, there are a number of extensions including DIF and adaptive testing. These two extensions (among many others) build on the core features of IRT to address more complex questions and handle more complex situations.

Although the sample used here was quite large (over 3,000) it is not necessary to have a sample of this magnitude to use IRT. IRT is very often used in educational settings, where sample sizes tend to be larger. This has helped propagate a belief that IRT can only be used with samples in the thousands. Two points are worth addressing here. First, the available literature suggests that IRT parameters can be recovered very well with 500 subjects. My own experience suggests that in some instances adequate parameter recovery is possible with data from as few as 200 subjects. As with most estimation procedures though, the bigger the sample size the better the estimates. This leads to my second point: IRT calibrations only have to happen once for a particular scale. Once IRT parameters are available for particular scale they will apply to any new sample that comes from the same population as the calibration sample. While a new calibration would be in order if the population of interest was different from the population sampled for calibration (e.g., children versus adults), many scales are used in similar populations quite often. In these instances, a concerted effort to obtain stable IRT parameters would mean that future researchers wouldn't need to do IRT analyses – they would just use the published item parameters and produce scores based on their observed data.

Where to go next?

There are hundreds of books and papers about IRT and its various extensions. In my opinion, Thissen and Orlando (2001, pp. 73–98) is one of the best chapter length introductions to IRT models for dichotomous data. The companion piece is Thissen, Nelson, Rosa, and McLeod (2001, pp. 141–150), which covers IRT models for polytomous data. The back end of both chapters covers IRT scale scores and provides excellent resources for those wanting to learn more about scoring in IRT. An excellent and accessible book-length treatment is Hambleton, Swaminathan, and Rogers (1991), which covers an astonishing amount of material for its 174 pages. For some IRT writings embedded firmly in the social and personality literature, see Fraley, Waller, and Brennan (2000) and Steinberg and Thissen (1995).

Acknowledgements

I would like to thank Richard Petty and Kenneth DeMarree for providing the data used throughout this paper. Thanks are also due to Li Cai, Kentaro Fujita, Lisa Libby, and R. J. Wirth, who provided helpful feedback on earlier drafts.

Short Biography

Michael C. Edwards received his PhD in December of 2005 from the L. L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill, where he was an advisee of Dr. David Thissen. Since January of 2006 he has been an assistant professor in the quantitative area of the psychology department at The Ohio State University. He is interested in measurement issues in the social sciences, with a focus on item response theory, factor analysis, and computerized adaptive testing.

Footnotes

[*] Correspondence address: 1827 Neil Ave, Columbus, OH 43210, USA. Email: edwards.134@osu.edu

[1] According to a brief search conducted on ISI Web of Knowledge, Cacioppo and Petty (1982) and Cacioppo et al. (1984) have been cited over 1250 times.

[2] http://faculty.psy.ohio-state.edu/edwards/

[3] The 2PLM is often presented with a scaling constant, D, which puts the parameters in a normal, rather than logistic, metric. It is a holdover from very early work in IRT and serves no real purpose, so I omit it here.
[4] For the model to function, we must also assume that there is a b_0 equal to $-\infty$ and a b_C equal to ∞ and that c ranges from 0 to C.
[5] A scree plot is a plot of the Eigenvalues from a particular set of data. To identify a plausible number of factors, one looks for the bend (or elbow) in the plot. The idea being that beyond the bend, the remaining values are so similar that to argue for taking one would be to argue for taking them all.

References

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2004). *CEFA: Comprehensive Exploratory Factor Analysis, Version 2.00* [Computer software]. Retrieved from http://faculty.psy.ohio-state.edu/browne/.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, **42**, 116–131.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, **48**, 306–307.

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

Cole, S. R., Kawachi, I., Mailer, S. J., & Berkman, L. F. (2000). Test of item response bias in teh CES-D scale: Experience from the New Haven EPESE study. *Journal of Clinical Epidemiology*, **53**, 285–289.

Cook, L. L., & Eignor, D. R. (1991). Irt equating methods. *Educational Measurement: Issues and Practice*, **10**, 37–45.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Edwards, M. C., & Edelen, M. O. (2009). Special topics in item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of Quantitative Methods In Psychology* (pp. 178–198). New York, NY: Sage.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the Study of Change. *Research in Human Development*, **6**, 74–96.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item reponse theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, **78**, 350–365.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, **6**, 1–55.

Jöreskog, K. G., & Sörbom, D. (2003). *LISREL 8.54* [Computer software]. Chicago, IL: Scientific Software International, Inc.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking*. New York, NY: Springer.

Liang, L. (2007). *A semi-parametric Appraoch to Estimating Item Response Functions*. Unpublished doctoral dissertation, The Ohio State University.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 369–380). New York, NY: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). *MPS5 for Windows: A Program for Mokken Scale Analysis for Polytomous Items* [Computer software]. Groningen, Germany: iec ProGAMMA.

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

Ramsay, J. O. (2000). TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data [Computer software]. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

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

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. *Sociological Methodology*, **18**, 271–307.

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). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, **11**, S39–S49.

Thissen, D. (1991). *MULTILOG: Multiple Cateogry Item Analysis and Test Scoring Using Item Reponse Theory* [Computer software]. Chicago, IL: Scientific Software International, Inc.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 141–186). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H. (2000). Introduction and history. In H. Wainer et al. (Eds.), *Computerized Adaptive Testing: A Primer* (2nd edn, pp. 1–20). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, **12**, 58–79.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods*, **11**, 253–270.