Critical Issues to Address when Applying Item Response Theory (IRT) Models

Maria Orlando, Ph.D.

RAND Corp.

As indicated by its title, this paper outlines some critical issues to consider when applying IRT models to health outcomes and behavior data including model assumptions, model fit, and sample size requirements. The focus is on unidimensional parametric IRT models; however, multi-dimensional IRT models (Reckase, 1977) and non-parametric IRT models (Ramsay, 1991; 1995) also exist. Although these models have particular advantages that may make them appropriate in some applications, they are less commonly applied to health outcomes data, and thus are not discussed in this paper.

Evaluating assumptions of IRT models

One important assumption of unidimensional parametric IRT models is that the construct being measured is in fact *unidimensional*; that is, that the covariance among the items can be explained by a single underlying dimension. One way to check this assumption is by examining the relative sizes of the eigenvalues associated with a principal components analysis of the item set. A general rule of thumb in exploratory factor analyses of this type suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 in this analysis (Loehlin, 1987). However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis with IRT. A second way to determine the number of factors is with a plot of the successive eigenvalues, or a scree plot (Cattell, 1966; 1978; Loehlin, 1987). Using this, the decision about the number of factors is arrived at based on the point at which the curve of decreasing eigenvalues changes from a rapid, decelerating decline to a flat gradual slope. For example, in one IRT application, a principal components analysis of a 30-

item variant of the Center for Epidemiological Studies Depression Scale (CES-D) scale yielded five eigenvalues greater than 1. However, the first eigenvalue (13.37) was substantially greater than the next four (1.6, 1.5, 1.4, 1.1). In addition, 29 of the 30 items had standardized factor loadings greater than .35, ranging from .28 to .81 with an average of .65. Based on these results, the factor structure of the 30 items was deemed sufficiently unidimensional for application of IRT (Orlando, Sherbourne, & Thissen, 2001). As implied by this example, the relative size of the eigenvalues, as well as the extent and magnitude of the item loadings on the first principal component can be taken into consideration when evaluating the unidimensionality assumption for IRT applications. Similarly, sets of items that have a second order factor structure (i.e., several first order factors all loading onto one general higher order factor) may also meet requirements for unidimensionality. In cases where the assumption may be tentative, it is important to check the IRT model results for any anomalies that may arise due to violation of this assumption (e.g., one or more items with very low item slope parameters). Additionally, if it is not appropriate to assume that the item responses are continuous, it is preferable to conduct exploratory factor analysis for categorical data.

A second assumption of IRT models is that the items display *local independence.* This is technically subsumed under the unidimensionality assumption and requires that, given their relationship to the underlying construct being measured, there is no additional systematic covariance among the items. Local dependence (LD) can potentially arise among subsets of items that have a similar stem (e.g., a set of items that all refer to someone's experience of physical pain), items that have very similar content, or items that are presented sequentially. Software to identify LD in dichotomous items does exist (Chen & Thissen, 1997), but it is not appropriate for polytomous items. An alternative way to identify LD is with a confirmatory factor analysis. Excess covariation among items in the residual matrix of a single factor CFA model can be indicative of LD. Examining this matrix carefully, or looking at the modification indices associated with the one-factor solution, can reveal potential LD. It is also useful to examine the output from an IRT calibration. Often if there is LD in a pair of items they will have inflated slope estimates. This is especially true with short scales. Essentially the LD, if it is strong enough, becomes the "definition" of the latent variable, so the two items with the LD have very high slopes relative to the other scale items. If this occurs, one item from the pair should be removed and the scale should be recalibrated.

One of the most basic assumptions of the application of parametric IRT models is that the model is appropriate for the data. Evaluation of this assumption involves choosing the right model and evaluating model fit. These topics are detailed in the following two sections.

Choosing the right IRT model

There are several different IRT models to choose from (Thissen and Steinberg, 1986). The most commonly used models are listed in the Item Response Theory Models Table in the front of the conference notebook. The first consideration when choosing the right model involves the number of item response categories, as this obviously limits the choice of appropriate models. For dichotomous items, the 1, 2, and 3 parameter logistic models are available (1PLM, 2PLM, 3PLM). For polytomous items, variations of the Partial Credit Model (PCM, Masters, 1982; RSM, Andrich, 1978a, 1978b; GPCM, Muraki, 1992, 1997) as well as the Graded Response Model (GRM, Samejima, 1969, 1997) are available for ordered responses, and the Nominal Model (Bock, 1972) is appropriate for items with a non-specified response order.

Members of the Rasch (1960) family of models are indicated in the Item Response Theory Models Table with an asterisk. The distinguishing characteristic of this family of models is that all items are assumed to have an equal relationship to the underlying latent construct being measured, thus these models estimate a common discrimination parameter for all items. Whether your items are dichotomous or polytomous, an important consideration when choosing the right model is whether the item discrimination parameters, or slopes, should be free to vary across items, or whether a model from the Rasch family is more appropriate. Each class of models has advantages. The main benefit of the Rasch models is their parsimony, and the ensuing computational advantages (e.g., software with extensive interpretative output, straightforward assessment of item fit). However, it is often the case that a less constrained model that estimates separate slopes for each item is a more accurate reflection of the data.

Apart from the issue of varying versus constrained slopes, there is also the option with dichotomous items to estimate a non-zero lower asymptote (the 3PLM). This “guessing” parameter was introduced in models of educational test items to characterize respondents’ probability of getting a question correct simply by chance (e.g., guessing correctly on a multiple choice item). The utility of this parameter has been explored for non-educational items (e.g., Reise & Waller, 2003), but is not commonly estimated in this context, as its interpretation is somewhat unclear.

For polytomous items, the nominal model is appropriate if the item responses do not have a specified order, or if a researcher wants to confirm a response order. Usually in health outcomes research the item responses are polytomous and ordered, so either the GPCM (or Rasch-family constrained PCMs) or the GRM is the suitable model. The choice between these two models is somewhat arbitrary, as they generally produce nearly identical results, albeit with slightly different parameterizations. Choosing one of these models over the other tends to be primarily a result of personal preference and familiarity with software (PARSCALE is set up to estimate the PCMs more easily, whereas MULTILOG favors the GRM). Generating descriptive item plots, for example with TESTGRAF (Ramsay, 1995), can also be a useful tool in determining the appropriate model for your data.

Evaluating IRT model fit

All applications of IRT implicitly assume that the model is correct; the utility of the IRT model is dependent upon the extent to which the model accurately reflects the data. As part of the process of model fitting in IRT, it is therefore desirable to employ some diagnostic tool to evaluate the degree of model-data misfit. The fit of the model can be examined through the comparison of model predictions and the observed data in various ways.

The direct assessment of overall model fit poses challenges and is seldom directly evaluated. However, the relative IRT model-data fit can be assessed through the comparison of -2*log likelihood for nested models (Thissen, 1991). This statistic is distributed as χ^2 with the appropriate degrees of freedom. For example, to examine the relative fit of the 2PL and 3PL models to a set of items, one can evaluate the significance of the difference in the -2*log likelihood from each of these model formulations. This difference is distributed as χ^2 with the degrees of freedom equal to the difference in the number of parameters in the two models. A significant result would imply that the 3PLM provides superior fit to the data.

In addition to examining the overall fit of the model to the data, it is also possible to examine the fit for each item. Goodness of fit statistics for the 1PL or Rasch (1960) family of models are relatively straightforward to construct, as the observed score is a sufficient statistic for θ, and the model's predicted proportions can be directly compared to the observed data for each score group. Several indices for this family of models have been proposed (Andersen, 1973;

Glas, 1988; Rost and von Davier, 1994; Wright and Mead, 1977; Wright and Panchapakesan, 1969), and many are available as standard output in the Rasch-oriented software packages.

Goodness of fit statistics for the 2PLM and 3PLM have also been constructed. The construction of these indices is more complex because θ, the ability or proficiency that forms the basis of the model is a latent variable, and so the model's predicted proportions as a function of θ cannot usually be compared directly to the observed data. One class of dichotomous item fit indices gets around this problem by grouping respondents according to their model-based θ estimates, and calculating observed and expected responses for these groups. Indices of this class include Yen's Q_1(Yen, 1981), Bock's χ^2 (Bock, 1972), and McKinley and Mills (1985) likelihood ratio G^2 statistic based on computations parallel Yen's (1981) Q_1. The computer program BILOG generates item fit statistics of this class if the scale has 20 or more items, but these indices should be interpreted cautiously as their Type I error rates tend to be inflated.

An alternative approach to assessing item fit has been introduced recently (Orlando & Thissen, 2000; 2003). In this summed score likelihood-based approach, an item fit statistic is formulated based on the observed and expected frequencies correct and incorrect for each summed score. This summed score approach was used to form two new indices for dichotomous IRT models: $S\text{-}X^2$, a Pearson X^2 statistic and $S\text{-}G^2$, a likelihood ratio G^2 statistic. An extension of the likelihood-based item fit indices for use with polytomous items is currently underway (Bjorner et al, in progress). This extension will be especially useful for applications to health outcomes, which are often measured with polytomous items. Although these indices are not calculated as part of any commercial software, interest in them is increasing, and software to calculate $S\text{-}X^2$ is available as freeware.

Several graphical representations of item fit have also been proposed, to be used in conjunction with a fit statistic, or as an exploratory diagnostic for item fit. Hambleton and Swaminathan (1985) suggest a graphical comparison of the observed average item performance levels of various ability groups with the performance predicted by the model. Wainer and Mislevy (1990) display similar plots of data and a trace line, while Kingston and Dorans (1985) propose the analysis of item ability regressions as a diagnostic tool for item fit. Examples of this type of plot are shown in Figure 1. Data for these figures are based on 1,000 respondents in the PETS-A study (Stevens and Morral, 2003), which included a 10-item Substance Problem Index (SPI) as part of the Global Appraisal of Individual Needs (GAIN; Dennis, 1998). The top panel of Figure 1 displays an item from the SPI that was identified as misfitting with the $S\text{-}X^2$ index, and the lower panel shows an item that did not have a significant $S\text{-}X^2$ value. Plots constructed based on posterior probabilities have also been utilized (Drasgow et al, 1995; Mislevy and Bock, 1986).

Sample size requirements for instrument evaluation

Although there are no definitive answers regarding sample size requirements, there are some general statements and guidelines that can be outlined.

First, models with fewer parameters will require smaller samples, and more complex models will require larger sample sizes. Rasch models estimate the fewest parameters, and thus smaller sample sizes are adequate for stable parameter estimates – perhaps as few as 100 (Linacre, 1994, suggests 50 for the simplest Rasch model). For models with more parameters, sample size requirements are not entirely clear. Tsutakawa & Johnson (1990) recommend a sample size of approximately 500 for accurate parameter estimates. However, others have suggested that as little as 200 or fewer observations can be adequate (e.g., for DIF detection; Orlando & Marshall, 2002; Thissen, Steinberg, & Gerrard, 1986).

Second, IRT item parameter estimates and scores will be more accurate (have smaller standard errors) as sample size increases. This implies that the purposes of the calibration need to be considered, as different levels of precision may be acceptable given the nature of the question. For example, to evaluate questionnaire properties, one does not need large sample sizes for a clear picture of response behaviour, although it is important to have a heterogeneous sample that accurately reflects the range of population characteristics. But if the purpose is to generate accurate IRT scores for persons responding to a questionnaire, or to calibrate items in an item bank, larger samples are required.

Another related consideration is the sampling distribution of the respondents. A very large sample of homogeneous respondents that do not reflect the population of interest will result in highly precise parameter estimates, but only for a limited range of the underlying construct being measured. Ideally, respondents should be spread fairly uniformly over the range of interest. Items at extreme ends of the construct will have higher standard errors associated with their estimated parameters if fewer people are located there.

The better the item response data meet the IRT assumptions of unidimensionality, conditional independence, and hierarchical ordering by difficulty, the smaller the sample size need be. Also, the relationship between the items and the measured construct is important, as poorly related items may require larger sample sizes (Thissen, 2003). Increasing the number of response categories also increases the need for larger samples, as more item parameters must be estimated. The ideal is to have respondents in each cell of all possible response patterns for a set of items; however, this is rarely achieved. At the least, it is important to have some people respond to each of the categories for every item to allow the IRT model to be fully estimated.

Psychometrician's role in evaluating and revising instruments

IRT can be an extremely useful tool for scale development, refinement, and evaluation, as results from an IRT calibration provide valuable insights into the performance of items and scales. However, these insights are most useful when they are complemented by results from classical analyses, such as examination of item-total correlations and Cronbach's alpha, and communicated to content experts.

It is important to conduct classical analyses because these analyses can familiarize you with the data, and inferences from the IRT analyses should generally correspond to those from the classical analyses. Large discrepancies in results from these two types of analyses could indicate that the IRT model is inappropriate for the data. Discrepancies can also occur if the classical analyses assume that the items are measured on a continuous scale, this may be an unreasonable assumption.

Although it is challenging to effectively communicate the complex results from an IRT analysis to less technical researchers, sharing results with content experts and eliciting their feedback is extremely important. Often, results that appear anomalous to a non-expert are easily explained by an expert. Additionally, items that appear only marginally useful, and perhaps dispensable from an analytic standpoint may be substantively critical to the content validity of the scale. Input from content experts is essential if the implications of results are to be evaluated according to both their statistical as well as their clinical significance.

References

Anderson, E. (1973). A goodness of fit test for the rasch model. *Psychometrika, 38,* 123-140.

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

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

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

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate behavioral Research,* 1, 245-267.

Cattell, R.B. (1978). *The scientific use of factor analysis.* New York: Plenum.

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.

Dennis, M. L. (1998) G*lobal Appraisal of Individual Needs (GAIN) Manual: Administration, Scoring, and Interpretation*. Lighthouse Publications.

Drasgow, F., Levine, M.V., Tsien, S., Williams, B., & Mead, A. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19,* 143-165.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53(4),* 525-546.

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

Kingston, N., & Dorans, N. (1985). The analysis of item-ability regressions: an exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281-288.

Loehlin, J.C. (1987). *Latent variable models.* New Jersey: Lawrence Erlbaum Associates.

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

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 19,* 49-57.

Mislevy, R.J., & Bock, R.D. (1986). *Bilog: item analysis and test scoring with binary logistic models.* Mooresville, Indiana: Scientific Software.

Muraki, E. (1992). A generalized partial credit model: Application of the EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Orlando M. & Marshall, G.N. (2002) Differential item functioning in a Spanish translation of the PTSD checklist: detection and evaluation of impact. *Psychological Assessment, 14,1,* 50-9.

Orlando, M., Sherbourne, C.D. & Thissen, D. (2000) Summed-score linking using item response theory: Application to depression measurement, *Psychological Assessment, 12(3)*, 354-359.

Orlando, M., and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,1,* 50-64.

Orlando, M., & Thissen, D. (2003) Further examination of the performance of $S\text{-}X^2$, an item fit index for dichotomous item response theory models. *Applied Psychological Measurement, 27(4)*, 289-98.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611-630.

Ramsay, J. O. (1995). TestGraf - A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data [Computer software]. Montreal: McGill University.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Reckase, M.D. (1977). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden and Ronald K. Hambleton (Eds), Handbook of modern item response theory (pp. 271-286). New York: Springer-Verlag.

Reise, S.P. & Waller, N.G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8, 2,* 164-84.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for rasch models. *Applied Psychological Measurement, 18,* 171-182.

Samejima F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.

Samejima F. (1997). Graded response model. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Stevens, S.J. & Morral, A.R., Eds. (2003). *Adolescent Substance Abuse Treatment in the United States: Exemplary Models from a National Evaluation Study*. New York: Haworth Press.

Thissen, D. (2003). Estimation in Multilog, in M. du Toit (ed.) IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact, Lincolnwood, IL: Scientific Software International.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory.* Chicago: Scientific Software.

Thissen, D., & Steinberg, L. (1986). A Taxonomy of Item Response Models. *Psychometrika, 51*(4), 567-577.

Tsutakawa, R. K. & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55,* 371-390.

Wainer, H., & Mislevy, R.J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg, & D. Thissen, *Computerized adaptive testing: A primer* (65-101). Hillsdale NJ: Lawrence Earlbaum Associates.

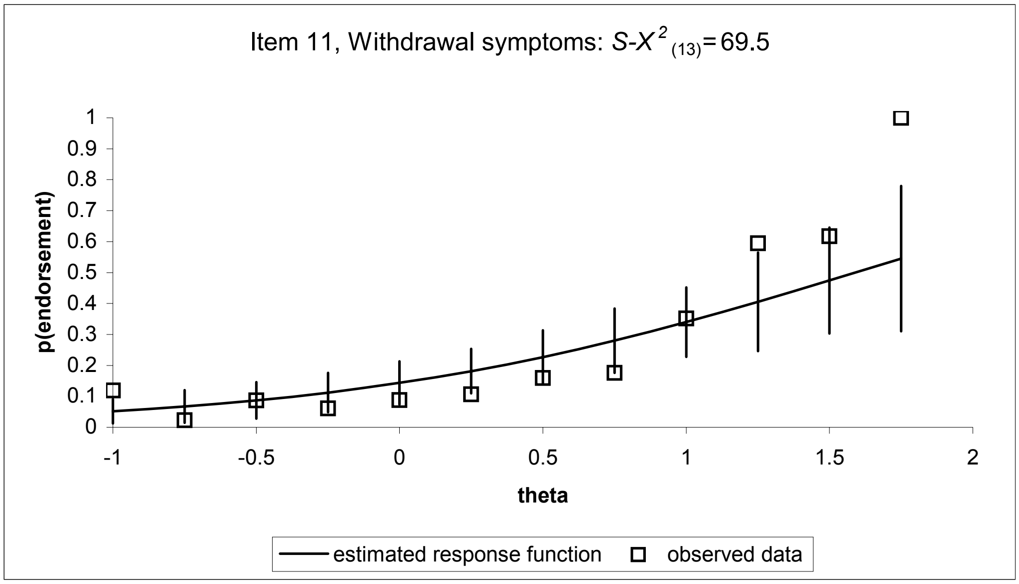
Wright, B., & Mead, R. (1977). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago IL: University of Chicago, Department of Education, Statistical Laboratory.

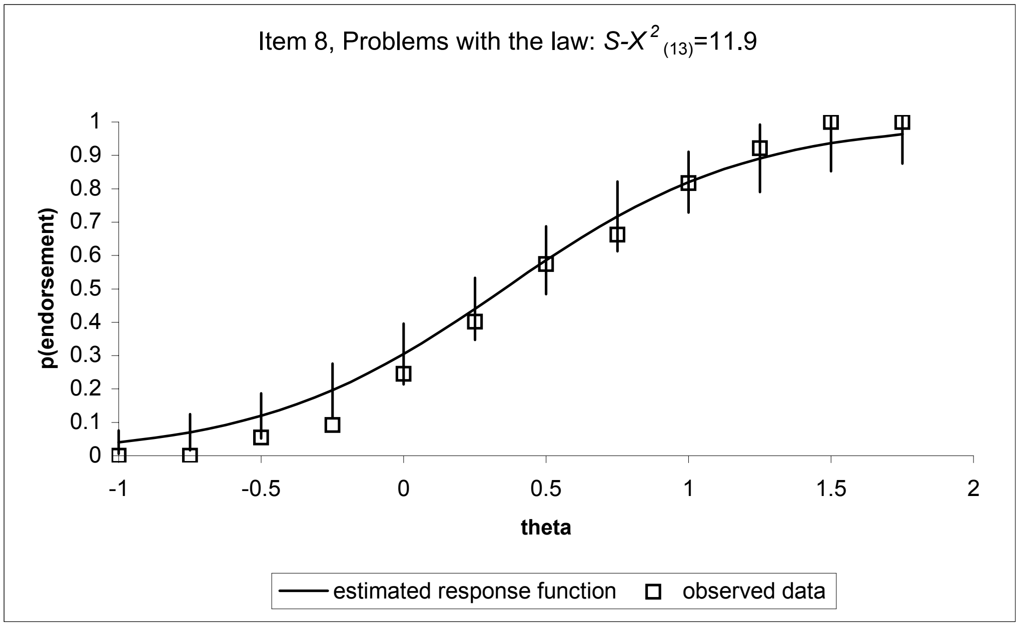
Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Figure Captions

Figure 1. Example graphical representations of item fit with two dichotomous items from the 10-item Substance Problems Index. The top panel shows a misfitting item, and the bottom panel, an item that fits according to the *S*-X^2 index.


Item 11, Withdrawal symptoms: $S\text{-}X^2_{(13)}$= 69.5
p(endorsement)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-1
-0.5
0
0.5
1
1.5
2
theta
estimated response function
□ observed data


Item 8, Problems with the law: $S\text{-}X^2_{(13)}$=11.9
p(endorsement)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-1
-0.5
0
0.5
1
1.5
2
theta
estimated response function
□ observed data