Qual Life Res (2007) 16:5–18
DOI 10.1007/s11136-007-9198-0

# Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement

**Maria Orlando Edelen · Bryce B. Reeve**

Received: 25 August 2006 / Accepted: 13 February 2007 / Published online: 21 March 2007


**Abstract**
*Background* Health outcomes researchers are increasingly applying Item Response Theory (IRT) methods to questionnaire development, evaluation, and refinement efforts.
*Objective* To provide a brief overview of IRT, to review some of the critical issues associated with IRT applications, and to demonstrate the basic features of IRT with an example.
*Methods* Example data come from 6,504 adolescent respondents in the National Longitudinal Study of Adolescent Health public use data set who completed to the 19-item Feelings Scale for depression. The sample was split into a development and validation sample. Scale items were calibrated in the development sample with the Graded Response Model and the results were used to construct a 10-item short form. The short form was evaluated in the validation sample by examining the correspondence between IRT scores from the short form and the original, and by comparing the proportion of respondents identified as depressed according to the original and short form observed cut scores.
*Results* The 19 items varied in their discrimination (slope parameter range: .86–2.66), and item location parameters reflected a considerable range of depression (–.72–3.39). However, the item set is most discriminating at higher levels of depression. In the validation sample IRT scores generated from the short and long forms were correlated at .96 and the average difference in these scores was –.01. In addition, nearly 90% of the sample was classified identically as at risk or not at risk for depression using observed score cut points from the short and long forms.
*Conclusions* When used appropriately, IRT can be a powerful tool for questionnaire development, evaluation, and refinement, resulting in precise, valid, and relatively brief instruments that minimize response burden.

**Keywords** IRT · Health outcomes · Adolescent depression · Short form

A new generation of health outcomes instruments is being developed based on the principles of item response theory (IRT) [1]. IRT comprises a collection of modeling techniques for the analysis of item level data obtained to measure interindividual variation (e.g., in health status). This collection of techniques generates rich item level information and offers many advantages over classical test theory (CTT) [2–5]. IRT can be used to evaluate the psychometric properties of an existing scale and its items, to optimally shorten the scale when necessary, and to evaluate the performance of the reduced scale. When used appropriately, IRT modeling can produce precise, valid, and relatively brief instruments resulting in minimal response burden.

The goals of this paper are to provide a brief overview of IRT, to review some of the critical issues associated with IRT applications, and to demonstrate the basic features of IRT with an example. In this paper, we focus solely on unidimensional parametric IRT models. However, non-parametric and multidimensional IRT models also exist [6–8].

M. O. Edelen (✉)
Department of Psychiatry & Human Behavior, Brown Medical School, Box G-BH, Providence, RI 02912, USA
e-mail: edelen@brown.edu

B. B. Reeve
National Cancer Institute, Bethesda, MD, USA

## Brief overview of IRT

The item characteristic curve (ICC), or trace line, is the basis of IRT, and is most commonly defined as a logistic function that models the relationship between a person's response to an item and his/her level on the construct measured by the scale. For items with dichotomous response options, the two parameter logistic (2PL) model is often applied. This model yields a trace line that is described by the location ($b$) and slope ($a$) parameters. The $b$ parameter (also called the threshold parameter) is the point along the ICC at which the probability of a positive response for a dichotomous item is 50%. The larger the location parameter, the more of the measured construct (often denoted as $\theta$) a respondent must have to endorse that item. The $a$ parameter (also called the discrimination parameter) represents the slope of the ICC at the value of the location parameter and indicates the extent to which the item is related to the underlying construct. A steeper slope indicates a closer relationship to the construct and therefore a more discriminating item.

Several models have been developed to estimate responses to items with more than two response categories [9–11]. These polytomous response models differ slightly in their parameterization, but all models essentially include the specification of a location and a slope parameter, and thus a corresponding ICC, for each response category [12].

Regardless of the number of item responses, the ICCs from an IRT calibration provide a visual representation of item properties that can be useful in scale development and refinement. In addition, the ICCs of several items can be combined to create scale characteristic curves, which indicate the relationship between an individual's expected score on the set of items and his/her level on the construct measured by the scale (i.e., $\theta$).

## Reliability

An important characteristic of IRT models is that reliability, or measurement precision, is described as a continuous function conditional on values of $\theta$, the measured construct. Precision is often depicted by item information curves (or functions), which indicate the range over $\theta$ where an item is best at discriminating among individuals. These curves are a function of the item parameters and can be calculated for individual items, sets of items, or entire scales. This feature of IRT is used to evaluate the performance of items and sets of items, and is very useful in constructing short forms or tailored assessments, ensuring that the selected subset of items provide adequate precision across the entire range of interest as well as maximizing precision along critical segments of the construct continuum. The inverse of the square root of the information function is equivalent to the standard error of measurement with respect to $\theta$.

## Linking

An important distinction between IRT and CTT is that IRT defines a scale for the underlying latent variable that is being measured by a set of items, and items are calibrated with respect to this same scale. This is why IRT is said to have a ''built-in'' linking mechanism. Linking is a general term that can be used to refer to both equating and calibration. Whereas the requirements for equating are stringent, calibrating two assessments of different lengths is less so and can easily be achieved using an IRT approach [13, 14]. The development of computerized adaptive testing (CAT) is based on this principle [5].

The linking aspect of IRT means that once items are calibrated for a population (i.e., item parameters are known), comparable scores on a given construct may be calculated for respondents from that population who answered only a subset of the items without intermediate equating steps. In applications where the item parameters are not known prior to administration, the linking of two or more test forms is still fairly straightforward provided both forms measure the same construct and there are some overlapping items on the forms. Item parameters can be estimated based on responses to the items from both forms, unique and overlapping, as if they comprised a single scale, and IRT scores can be derived based on these item parameters for the response patterns in the data for each test [15].

## Critical issues for IRT application

IRT application requires a number of assumptions, and the usefulness of the IRT model is contingent on the extent to which these assumptions are met. 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. Examination of output from an exploratory factor analysis, including eigenvalues, scree plots, and the magnitude of item loadings on the first factor can help in evaluating this assumption [16–18]. In addition, the fit of a one-factor confirmatory factor analysis (CFA) model can also be examined. Whether an exploratory or confirmatory approach is taken, the factor analysis estimation method should appropriately model the ordinal nature of the item responses. In cases where the unidimensionality 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).

A lack of unidimensionality may also occur in cases where item properties differ according to some grouping variable (e.g., gender). When this occurs, response patterns are a function of both the underlying dimension that is being measured as well as group membership. This type of violation is difficult to discern with single group factor analytic methods but can be examined with methods that evaluate differential item functioning (DIF) [19]. This topic is discussed in detail in another paper in this issue [20].

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 is available [21], but it is not appropriate for polytomous items. An alternative way to identify LD is with a CFA. Excess covariation among items in the residual matrix of a single-factor CFA model can be indicative of LD. Examining this matrix and determining whether the modification indices associated with the one-factor solution suggest adding item covariance parameters 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. In this case, the two items with the LD have very high slopes (e.g., >4) relative to the other scale items, which have more typical slopes (i.e., ranging between .5 and 2.5). If this occurs, a calibration of the item set omitting one of the LD items is most appropriate.

## Model choice and model fit

One of the most basic assumptions of the application of parametric IRT models is that the model is appropriate for the data. This assumption involves choosing the right model and evaluating model fit.

### Choosing the right IRT model

There are several different parametric unidimensional IRT models to choose from [12]. The first consideration when choosing the right model involves the number of item response categories. For dichotomous items, the 1, 2, and 3 parameter logistic models are most common (1PLM, 2PLM, 3PLM), and models including an upper asymptote parameter (e.g., 4PLM) are also possible. For polytomous items, variations of the Partial Credit Model (PCM [9]; Rating Scale Model, RSM [22, 23]; Generalized Partial Credit Model, GPCM [24, 25]) as well as the Graded Response Model (GRM) [10, 11] are available for ordered responses, and the Nominal Model [26] is appropriate for items with a non-specified response order.

A second 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 [27] family is more appropriate. The distinguishing characteristic of the Rasch family of models is that they estimate a common discrimination parameter for all items. 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 [15].

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. The utility of this parameter has been explored for non-educational items [28], but is not commonly estimated in this context, as its interpretation is somewhat unclear. A non-zero upper asymptote is also possible. Its use has been explored both in lieu of and in conjunction with the non-zero lower asymptote [28].

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) [29].

Generating descriptive item plots can also be a useful tool in determining the appropriate model for a particular set of data [30]. The software program TESTGRAF [31] generates non-parametric ICC plots describing the

relationship between the probability of an item response and the construct being measured. Examination of these plots can provide some insight into the suitability of various parametric models.

### 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. Generally speaking, parametric models attempt to characterize data in a parsimonious fashion, and some extent of misfit is inherent in every unsaturated model. As with any parametric model, however, considerable misfit in an IRT model implies model misspecification, and the usefulness of inferences based on parameters from a misspecified model is negligible. 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 in non-Rasch model applications. However, the relative IRT model-data fit can be assessed through the comparison of nested models (e.g., the 2PLM vs. the 3PLM) [32]. In addition to examining the overall fit of the model to the data, it is also possible to examine the fit for each item. This is useful because an IRT model may be an accurate representation of the data in general, but a poor reflection of the observed responses to a particular item (i.e., the item's responses do not conform to the specified logistic function).

Item goodness of fit statistics for the Rasch family of models are relatively straightforward to construct. Several indices for this family of models have been proposed and are available as standard output in the Rasch-oriented software packages [33–37].

Item goodness of fit statistics for non-Rasch models are also available, although their construction is rather complex. Most work in this area has been restricted to dichotomous items and many of the more traditional indices do not perform optimally [26, 38, 39]. A new class of item fit statistics based on an alternative approach tends to perform better [40, 41], and has recently been extended for use with polytomous items [42]. 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 [3, 43–46].

It is also possible to examine model-data fit at the individual level with person fit indices. Person fit indices evaluate the consistency of individual response patterns with a proposed model of valid responding based on the IRT model [15]. There are a number of different types of person fit indices that vary in their applicability [47]. They are often used to detect guessing and have also been applied to personality inventories [48–50].

## Sample size requirements

Although there are no definitive answers regarding sample size requirements, there are some general statements and guidelines that can be outlined. First, sample size needs increase with the complexity of the model. Sample sizes as small as 100 are often adequate for estimating stable Rasch-model parameters [51]. For models with more parameters, sample size requirements are not entirely clear. Tsutakawa and Johnson [52] 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) [53, 54]. Complexity aside, the better the item response data meet the IRT assumptions, the smaller the sample size need be. For example, precise parameter estimates for poorly related items may require larger sample sizes [55].

Second, IRT item parameter estimates and scores will have smaller standard errors as sample size increases. This implies that the purpose of the calibration needs to be considered, as different levels of precision may be acceptable given the nature of the question. For example, if the items are being calibrated as a basis for producing accurate individual IRT scores or to generate parameters for an item bank, large samples are required. However, smaller samples may be adequate to evaluate questionnaire properties.

Another related consideration is the sampling distribution of the respondents. In all calibrations, item properties only generalize to the population represented by the sample of respondents. A very large sample of homogeneous respondents that do not reflect the population of interest may result in highly precise parameter estimates, but only for a limited range of the construct being measured, and possibly relative to a mean that is not a good estimate of the population mean. Thus the sample should reflect the population of interest and contain enough respondents so that items even at extreme ends of the construct continuum will have reasonable standard errors associated with their estimated parameters. 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 necessary to have responses in each of the categories of every item for the IRT model to be estimated.

## Example

## Method

### Data and sample

Data for this example come from the National Longitudinal Study of Adolescent Health (AddHealth). AddHealth is a school-based study of the health-related behaviors of adolescents in grades 7–12 [56]. Self-administered questionnaires were given at school to more than 90,000 students in grades 7–12 between September 1994 and April 1995. Interviewer-administered questionnaires were also given at respondents' homes to a core sample of 12,105 adolescents between April and December 1995. Using systematic sampling methods and implicit stratification in the AddHealth study design ensured that this sample is representative of US schools with respect to region of country, degree of urbanicity, school type, ethnicity, and school size.

This example will utilize data from a total of 6,504 respondents in the AddHealth public use dataset. This sample consists of 50% of the core sample ($n$ = 6072) that completed the in-home survey and an oversample of black adolescents with a parent with a college degree ($n$ = 520). The total for the core sample and the high education black sample do not equal the total N due to some respondents being in both samples.

### Measure

The 19-item Feelings Scale, administered as part of the at-home survey, asked respondents to indicate on a 4-point scale (0 = *never*, 1 = *sometimes*, 2 = *a lot of the time*, 3 = *most or all of the time*) how often each of the 19 statements were true during the past week. Items expressing positive content were reverse-scored so that for all items, higher scores indicated greater levels of depression. The 19-item Feelings Scale is very similar in content to the widely used 20-item Center for Epidemiologic Studies Depression Scale (CES-D) [57]. In fact, researchers using the AddHealth data have used the CES-D scoring as a guideline for identifying cut-scores indicative of depression on the Feelings Scale [58].

As can be seen in Table 1, which displays abbreviated item content and item response frequencies, very few respondents in this sample endorsed the response category associated with the most depression (*most or all of the time* for the negative affect items, and *never or rarely* for the positive affect items). To determine whether it was

**Table 1** Abbreviated item content and response frequencies for 19-item feelings scale ($N$ = 6,504)

| | Item content | Content type | Scale value | | | |
|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 |
| 1. | Bothered by things | A | 60.37 | 31.90 | 5.94 | 1.79 |
| 2. | Had poor appetite | V | 64.62 | 26.88 | 6.32 | 2.17 |
| 3. | Had the blues | A | 72.04 | 20.00 | 5.74 | 2.22 |
| 4. | Felt just as good as other people (r) | C | 36.17 | 31.93 | 20.87 | 11.03 |
| 5. | Had trouble keeping mind focused | C | 40.46 | 42.68 | 12.58 | 4.27 |
| 6. | Felt depressed | A | 61.60 | 28.58 | 6.85 | 2.98 |
| 7. | Too tired to do things | V | 42.47 | 45.23 | 9.71 | 2.59 |
| 8. | Hopeful about the future (r) | C | 30.93 | 33.75 | 24.20 | 11.12 |
| 9. | Felt life had been a failure | C | 84.16 | 12.07 | 2.53 | 1.24 |
| 10. | Felt fearful | A | 72.67 | 23.82 | 2.51 | 1.00 |
| 11. | Felt happy (r) | A | 36.94 | 41.45 | 18.96 | 2.65 |
| 12. | Talked less than usual | A | 56.16 | 34.02 | 7.34 | 2.48 |
| 13. | Felt lonely | A | 64.10 | 27.56 | 6.18 | 2.16 |
| 14. | People unfriendly to you | C | 66.37 | 28.34 | 3.95 | 1.34 |
| 15. | Enjoyed life (r) | A | 48.43 | 31.56 | 16.08 | 3.93 |
| 16. | Felt sad | A | 52.47 | 40.51 | 5.18 | 1.85 |
| 17. | Felt people dislike you | C | 65.46 | 28.66 | 4.26 | 1.62 |
| 18. | Hard to start doing things | V | 48.18 | 43.40 | 7.13 | 1.30 |
| 19. | Felt life not worth living | A | 88.33 | 8.40 | 2.30 | 0.97 |

*Note*: Content type refers to affective (A), vegetative (V), and cognitive (C) sub-domains. Scale values from 0 to 3 for negative affect items, and from 3 to 0 for positive affect items correspond to response options ''never or rarely'' ''sometimes'' ''a lot of the time'' ''most or all of the time''

appropriate to analyze a 3-category version of the scale that combines the two highest response scale values, we compared the descriptive item statistics as well as the overall alphas and the Fisher's $z$-transformed average interitem correlations for the original (4-category) and 3-category versions. Table 2 shows the mean scores, standard deviations, and item total correlations for the 4- and 3-category item scoring. The descriptive item statistics, the overall alphas ($\alpha$ = .864 and .868 for 4- and 3-category versions, respectively), and the Fisher $z$-transformed average interitem correlations (Fishers $z$ = .520 and .524 for 4- and 3-category versions, respectively) for the two scale versions are comparable. Therefore, we elected to analyze the 3-category version of the scale for parsimony.

## Analytic approach

The descriptive information in Tables 1 and 2 was generated for the entire sample. However, for the subsequent analyses described below, we divided the data randomly into a development ($n$ = 3252) and validation ($n$ = 3252) sample.

*Dimensionality.* For parametric IRT analysis, essential unidimensionality [59, 60] holds when the dominant factor is strong enough that estimation of examinee trait levels is not affected by the presence of minor factors [15]. To check the assumptions of unidimensionality and local independence, we conducted an exploratory factor analysis (EFA) of the 19-item set using the developmental sample. This was followed by a confirmatory factor analysis (CFA) with the validation sample. All factor analyses were conducted using Mplus [61] and treating all items as ordinal. From the EFA, we examined the scree plot, eigenvalues, and the magnitude of item loadings on the single-factor solution to evaluate dimensionality. Multifactor solutions were also examined with as many factors as eigenvalues >1. These alternative solutions were evaluated based on the content and interpretability of the additional factors as well as the additional variance accounted for.

Following the EFA, a single-factor CFA model was estimated using the validation sample. Fit was evaluated based on three indices: the root mean square error of approximation (RMSEA) [62] the Nonnormed Fit index (NNFI) [63], and the comparative fit index (CFI) [64]. For the RMSEA, a smaller value indicates a closer fit; an RMSEA ≤.06 is considered to reflect good fit, values ≤.08 are fair, and values above .10 are generally considered to reflect poor fit. Values of the NNFI and CFI above .90 and .95 are generally accepted as reflecting adequate and good fit [65, 66]. Alternatives to the basic one factor model were considered to improve fit.

*IRT Calibration.* Using the random half of the respondents in the development sample, we calibrated the items

**Table 2** Item mean (SD) scores and correlations with total for 3- and 4-category response versions of the 19-item feelings scale ($N$ = 6,504)

| Item | 4-category | | | 3-category | | |
|---|---|---|---|---|---|---|
| | Mean | SD | Item total correlation | Mean | SD | Item total correlation |
| 1 | 0.492 | 0.690 | 0.504 | 0.474 | 0.636 | 0.498 |
| 2 | 0.460 | 0.711 | 0.417 | 0.439 | 0.645 | 0.410 |
| 3 | 0.381 | 0.696 | 0.598 | 0.359 | 0.624 | 0.590 |
| 4 | 1.068 | 1.004 | 0.373 | 0.957 | 0.824 | 0.389 |
| 5 | 0.807 | 0.815 | 0.471 | 0.764 | 0.719 | 0.480 |
| 6 | 0.512 | 0.752 | 0.673 | 0.482 | 0.668 | 0.664 |
| 7 | 0.724 | 0.741 | 0.416 | 0.698 | 0.676 | 0.414 |
| 8 | 1.155 | 0.986 | 0.327 | 1.044 | 0.813 | 0.349 |
| 9 | 0.208 | 0.538 | 0.538 | 0.196 | 0.483 | 0.535 |
| 10 | 0.318 | 0.572 | 0.432 | 0.308 | 0.533 | 0.434 |
| 11 | 0.873 | 0.806 | 0.513 | 0.847 | 0.750 | 0.514 |
| 12 | 0.561 | 0.736 | 0.341 | 0.537 | 0.667 | 0.353 |
| 13 | 0.464 | 0.709 | 0.579 | 0.442 | 0.643 | 0.584 |
| 14 | 0.403 | 0.632 | 0.374 | 0.389 | 0.586 | 0.378 |
| 15 | 0.755 | 0.862 | 0.523 | 0.716 | 0.777 | 0.536 |
| 16 | 0.564 | 0.679 | 0.611 | 0.546 | 0.623 | 0.615 |
| 17 | 0.420 | 0.653 | 0.492 | 0.404 | 0.599 | 0.495 |
| 18 | 0.615 | 0.676 | 0.398 | 0.602 | 0.639 | 0.404 |
| 19 | 0.159 | 0.488 | 0.490 | 0.149 | 0.439 | 0.486 |

using the GRM [11], which estimates a slope (*a*) parameter and two location (*b*) parameters for each 3-category item. We estimated two IRT models: (1) a parsimonious GRM that specified a single slope for all the items (similar to a Rasch model); and (2) a full GRM that specified unique slopes for each of the 19 items. We compared the suitability of these two nested models by evaluating the change in fit using –2*loglikelihood which is distributed as Chi-square with degrees of freedom equal to the difference in the number of parameters for the two models. For both solutions, we also evaluated fit at the item level.

As a final check on the unidimensionality assumption, we first examined the item properties of the final model, focusing on items that showed excess dependence based on the factor analyses. If the excess dependence is problematic, we would expect these items to have high slopes relative to values for the other items. Next, we conducted sensitivity analyses, examining results from a calibration that excluded items showing excess dependency. If this excess item covariance was a threat to the unidimensionality assumption we would expect to observe meaningful differences in the parameter estimates under these conditions. For example, the range of parameter values may differ, or the rank ordering of the items according to their slope or location parameters may change.

*Item properties and short form selection.* We used information from the IRT calibration to identify a shortened 10-item instrument that maintained adequate content coverage with maximum precision.

*Preliminary validation of short form.* We used data from the validation sample to compare the original and short forms as follows. First, we generated individual IRT scores based on the original and short forms and calculated the correlation between the two IRT scores as well as the average difference in the scores across individuals. We also examined the screening performance of the original and short forms by comparing the proportion of respondents who are identified as depressed according to the original and short form observed cut scores.

## Results

### Assessing dimensionality

The scree plot of eigenvalues from the EFA in the developmental sample was strongly suggestive of a single factor, with the first value substantially larger than the others (8.03, 1.71, 1.11, .955, .822, etc.). However, solutions for up to three factors were examined following the Keiser–Guttman criterion. In the single-factor solution, item factor loadings were all positive, ranging from .375 to .835. The promax-rotated two-factor solution extracted the 15 negative affect items in the first factor and the four positive affect items into a second factor; the three factor solution also distinguished the positive affect items as the second factor, and extracted an item pair (14 and 17) with similar content (''people unfriendly to you,'' ''people dislike you'') as the third factor.

The basic 1-factor CFA model did not fit well to the validation sample data ($\chi^2_{(107)} = 3{,}726$, CFI = .78, NNFI = .92, RMSEA = .10). However, the addition of correlations among the residuals of the four positively worded items as well as between the residuals of the item pair (14 and 17) resulted in a dramatic decrease in the $\chi^2$ and improved the practical fit measures to within acceptable ranges ($\chi^2_{(106)} = 1101$, CFI = .94, NNFI = .98, RMSEA = .05).[1] In this solution, all item loadings were positive and significant, with standardized values ranging from .352 to .853. We also examined alternative solutions that posited additional factors (e.g., a model that allowed positive content items to load on a ''methods factor'' in addition to the overall factor). These solutions did not yield superior fit relative to the single-factor model with correlated errors.

Based on these results, we determined that the 19-item scale was sufficiently unidimensional for IRT analysis. However, the excess item covariation among the positively worded items and the item pair in the single-factor solution indicated a violation of the local independence assumption. Although it is reasonable to elect to remove one or more items at this stage to alleviate the excess item covariation, IRT applications are often robust to violations of local dependence, especially when the scale consists of 10 or more items. In general, it is preferable to retain the full set of items, especially if they comprise a commonly used existing scale. Removing suspected LD items before calibrating the item set precludes gaining information about all items' performance in the context of the entire scale. Instead, the impact of the LD will be directly evaluated when examining the results of the IRT calibration and items will be removed at that stage if necessary.

### IRT Calibration

Using data from the developmental sample, the calibration of the 19-item scale with the reduced GRM (single slope for all items) resulted in a –2*LL value of 43,309.8, whereas a fully specified GRM with unique slopes for all items yielded a value of 42,090.0. The difference in these two values (1,219.8) is distributed as chi-square with

[1] In these analyses, items were treated as ordinal and the WLSMV estimator was used resulting in approximated $\chi^2$ and df values; thus the difference in the df for these two models (1) does not directly correspond with the difference in the number of estimated parameters (6).

**Table 3** GRM item parameter estimates, standard errors, and fit statistics for the 19-item feelings scale ($n$ = 3,252)

| | Item content | $a$ | $b_1$ | $b_2$ | $S - X^2$ | $p$ |
|---|---|---|---|---|---|---|
| 1. | Bothered by things | 1.43 (0.07) | 0.36 (0.04) | 2.27 (0.10) | 49.71 | 0.56 |
| 2. | Had poor appetite | 1.08 (0.06) | 0.68 (0.06) | 2.61 (0.14) | 53.38 | 0.54 |
| 3. | Had the blues | 2.14 (0.10) | 0.75 (0.03) | 1.80 (0.06) | 45.80 | 0.56 |
| 4. | Felt just as good as other people (r) | 0.87 (0.05) | –0.72 (0.07) | 0.99 (0.08) | 77.61 | 0.01 |
| 5. | Had trouble keeping mind focused | 1.22 (0.06) | –0.46 (0.05) | 1.62 (0.08) | 47.44 | 0.62 |
| 6. | Felt depressed | 2.66 (0.11) | 0.32 (0.03) | 1.54 (0.05) | 62.01 | 0.04 |
| 7. | Too tired to do things | 0.97 (0.06) | –0.43 (0.06) | 2.31 (0.13) | 57.07 | 0.33 |
| 8. | Hopeful about the future (r) | 0.83 (0.05) | –1.17 (0.09) | 0.80 (0.08) | 65.18 | 0.10 |
| 9. | Felt life had been a failure | 2.09 (0.11) | 1.26 (0.05) | 2.30 (0.09) | 47.68 | 0.33 |
| 10. | Felt fearful | 1.22 (0.07) | 0.99 (0.06) | 3.17 (0.17) | 75.20 | 0.01 |
| 11. | Felt happy (r) | 1.30 (0.06) | –0.61 (0.05) | 1.26 (0.06) | 68.63 | 0.04 |
| 12. | Talked less than usual | 0.86 (0.05) | 0.31 (0.06) | 2.89 (0.18) | 50.98 | 0.63 |
| 13. | Felt lonely | 1.89 (0.08) | 0.49 (0.03) | 1.86 (0.07) | 54.71 | 0.27 |
| 14. | People unfriendly to you | 0.95 (0.06) | 0.80 (0.07) | 3.39 (0.20) | 66.21 | 0.11 |
| 15. | Enjoyed life (r) | 1.36 (0.06) | –0.07 (0.04) | 1.29 (0.06) | 73.74 | 0.02 |
| 16. | Felt sad | 2.36 (0.09) | 0.05 (0.03) | 1.83 (0.06) | 72.04 | 0.01 |
| 17. | Felt people dislike you | 1.37 (0.07) | 0.56 (0.04) | 2.49 (0.11) | 57.90 | 0.21 |
| 18. | Hard to start doing things | 0.93 (0.06) | –0.14 (0.06) | 2.87 (0.17) | 71.92 | 0.05 |
| 19. | Felt life not worth living | 2.08 (0.12) | 1.51 (0.05) | 2.44 (0.10) | 67.21 | 0.01 |

18 degrees of freedom and is highly significant, indicating that the exclusion of the 18 unique item slope parameters in the more restricted GRM significantly detracts from the fit of the model. We also examined fit at the item level for the two models using the polytomous extension of $S - X^2$ [40]. For the simplified model, 15 of the 19 items were identified as misfitting at $p < .05$ after controlling for Type I error rates with the Benjamini–Hochberg adjustment [67].[2] In contrast, none of the items from the fully specified GRM were identified as misfitting using the same criteria (see Table 3). Consequently the fully specified GRM was adopted as the more appropriate model for this item set.

To evaluate whether the violation of the local independence assumption was problematic, we first examined the slope values from this final calibration (see Table 3), focusing on those items that showed some excess dependence in the CFA (the positive affect items [4, 8, 11, 15], as well as items 14 and 17). The slope values for these six items were not exceedingly high given the full range of item slope values from .83 to 2.66. In fact, these item slopes tended to fall in the low end of the range.

Next, we conducted sensitivity analyses, examining results from a calibration that excluded item 14, as well as a calibration that excluded the four positive affect items. The range of item slope values for the two additional calibrations was very similar to the 19-item results (.83–2.66 for all 19 items; .85–2.72 without item 14; .84–2.88 without positive affect items). In addition, the rank ordering of the items according to slope values was nearly identical in the three solutions. Only one item in each of the sensitivity calibrations did not correspond to the 19-item rank order. Finally, the changes in the location parameters were very small in magnitude for the two sensitivity calibrations relative to the 19-item, with differences for the majority of location parameters <.01. Based on these sensitivity analysis results, we concluded the 19-item GRM calibration was sufficiently unidimensional and robust to the excess item covariation observed in the data.

## Item properties and short form selection

The parameter estimates and their standard errors from the fully specified GRM calibration are listed in Table 3. The slope estimates ranged from .83 to 2.66, indicating considerable variation in item discrimination (these values correspond to factor loadings of .45–.84). The location parameters for the 19 items reflect a sizeable range of underlying depression (–.72–3.39), but the majority of item response categories are only endorsed by respondents who have higher than average levels of depression (i.e., $\theta > 0$),

[2] This is similar to the Bonferroni adjustment in that it considers the total number of evaluations, but uses less stringent comparison values for obtaining significance depending on the rank order of the observed $p$-values. The largest observed $p$-value has a comparison value of .05, the smallest observed $p$-value has a comparison value of .05 divided by the number of comparisons, and all other comparison values lie within this range, adjusted according to the rank-order of the magnitude of the observed $p$-values.

implying that the item set as a whole is most useful in discriminating among individuals at the high end of the depression continuum.

Figure 1 displays ICCs for three of the 19 Feelings Scale items. These items were selected to show how ICCs vary depending on the slope parameter (item 3 has a relatively high slope, $a = 2.14$; item 1's is moderate, $a = 1.43$; and item 4's is low, $a = 0.87$), as well as the location parameter (item 4 is endorsed at relatively low levels of depression whereas response categories for items 1 and 3 are endorsed at higher levels of depression). The item information functions in Fig. 2 demonstrate how these variations affect measurement precision across the continuum. Item 3 has the highest slope, thus reaches maximum information levels among the 3 items. However, items 1 and 4, although providing generally lower information, have more precision than item 3 at lower levels of depression because of their lower location parameters.

As an example of how one might use IRT results to construct a short form, this section demonstrates the

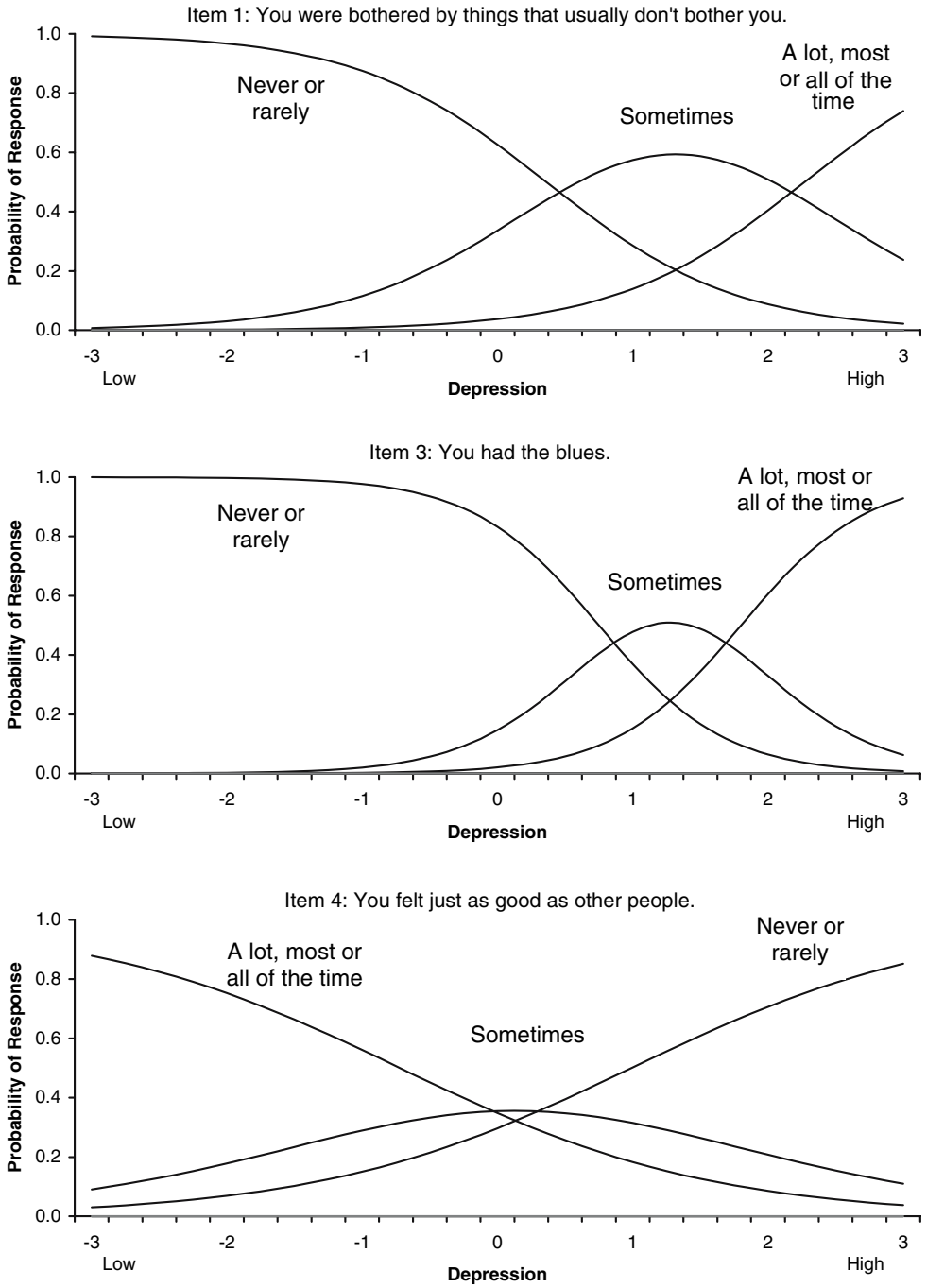


**Fig. 1** Three items from the AddHealth Feelings Scale with a variety of slope and location parameter values

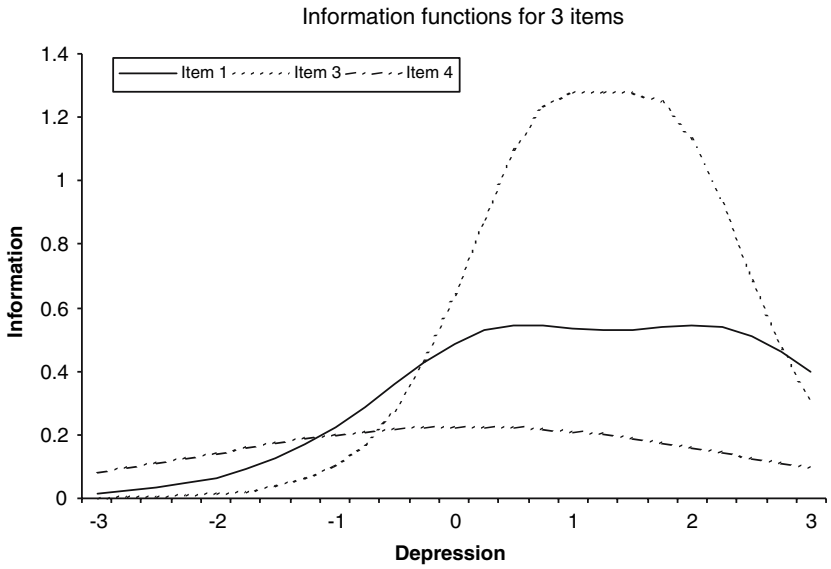


**Fig. 2** Information functions corresponding to the three example items in Fig. 1

selection of a 10-item form that maintains a reasonable balance of content from three sub-domains of depression (affective, cognitive, and vegetative), which were designated by two clinical experts (see Table 1). To maintain content coverage, five affective items, three cognitive items, and two vegetative items were retained. Although this yields a slightly larger proportion of vegetative items than in the long form, we chose to retain two vegetative items to adequately represent that sub-domain.

To guide item selection, we examined the item information functions in groups according to these content designations (Fig. 3). From the affective item set, which has the largest amount of information, we selected items 6, 11, 13, 15, and 19 (indicated in bold print and solid lines in Fig. 3). Although these five items do not combine to produce the maximum amount of information (e.g., items 3 and 16 have high information levels but are not in the selected set), we chose these five to maximize information with minimal content overlap. For example, we may have selected item 16, but judged that ''felt sad'' was too similar in content to ''felt depressed.'' Instead we chose item 13, ''felt lonely,'' which has relatively high information across a wide range of depression, and has unique content. We also selected items to maximize information across the entire continuum. For example, although item 11 does not reach high levels of information, it is the highest at the lower end of the continuum and thus informative for differentiating among people with low depression. Among the cognitive and vegetative item sets, items selected based on information also met our requirement of minimizing content overlap (cognitive items 5, 9, and 17; vegetative items 2 and 7). As can be seen from the standard error functions of the 19- and 10-item scales depicted in Fig. 4, the short form appears to maintain adequate measurement precision. Although there is a fairly even increase in the standard error across the entire depression continuum when going from 19 to 10 items, the loss in precision of the short form never exceeds .11 standard errors. In terms of reliability, at the center of the continuum, in the theta range –.5–.5, the loss in reliability from the 19- to the 10-item form is less than .10 and the 10-item form has reliability values greater than .76 in this range.

## Preliminary validation of short form

It is preferable, whenever possible, to examine the validity of IRT-derived short scales by comparing the short and long forms in an independent sample. We set aside half the AddHealth sample for this purpose. IRT scores ($\theta$'s) based on the original and 10-item short form were generated for each respondent in the validation sample. The correlation between these two scores was .96 and the mean difference in scores was –.013. Although this difference was significantly different from 0 ($t_{(3251)} = -2.87$, $p = .004$), there was a lot of statistical power for this test and the difference is small in magnitude, lending support for the validity of the short form.

Based on the scoring and cut score from the CESD, we also determined that an observed total score of 10 on the 19-item three-category response scale could be considered indicative of risk for depression. Maintaining a similar percent of the total score,[3] the corresponding observed total score for the 10-item short form is about five. The cross-classification of respondents above and below the two cut

[3] For the purposes of this demonstration, we elected not to conduct more sophisticated analyses for linking observed scores to one another and to IRT scores based on IRT calibrations [68].

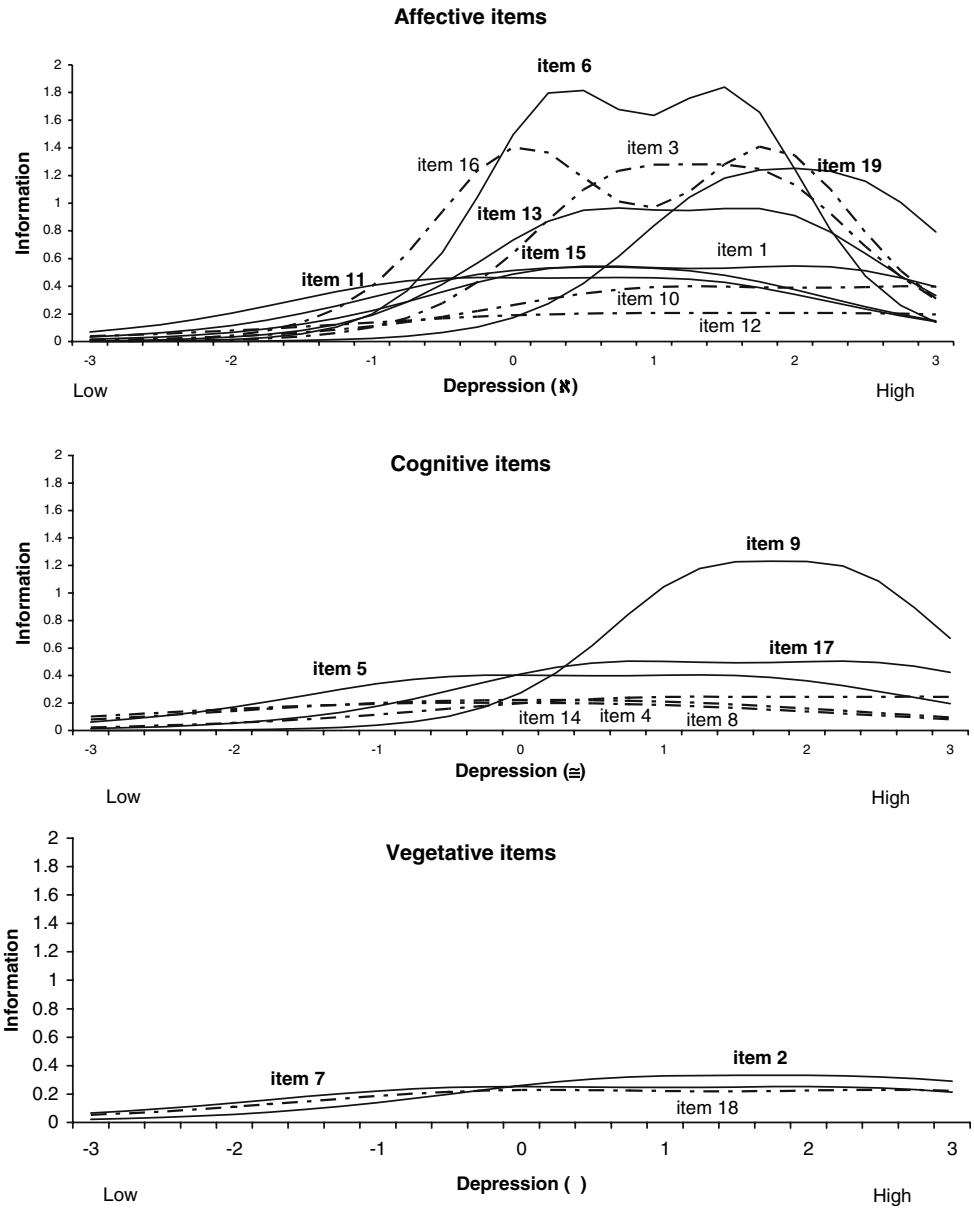


**Fig. 3** Information functions for 19 Feelings Scale items grouped according to content sub-domain

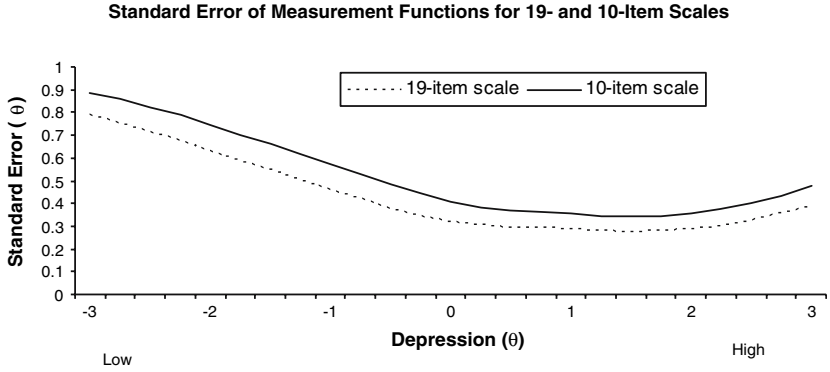


**Fig. 4** Standard error of measurement functions for full scale and 10-item short form

**Table 4** Cross-classification of respondents above cut-points for probable depression based on the 19- and 10-item forms ($n$ = 3,227)

| 10-item form | 19-item form | | |
|---|---|---|---|
| | Below 10 | At or above 10 | |
| Below 5 | 48.53 | 4.71 | 53.24 |
| At or above 5 | 4.77 | 41.99 | 46.76 |
| | 53.30 | 46.70 | |

scores in the validation sample is displayed in Table 4, and indicates that nearly 90% of the sample is classified in the same way regardless of the form used. The kappa statistic for this concordance is .810. The extent of classification correspondence between the two versions lends additional support for the use of the 10-item form, especially in applications where researchers wish to decrease respondent burden.

## Discussion

In addition to providing a brief overview of IRT, this paper has demonstrated how IRT methodology can be used for scale development, refinement, and evaluation. Critics assert that CTT methods frequently lead to similar conclusions, and question the value added by IRT. In our example, we used IRT methods and item content considerations to select a 10-item short form. A straightforward CTT approach might use a Principal Components Analysis to identify the 10 items that accounted for the most variance in the 19-item scale score. If we had adopted this approach, we would have selected seven of the same items (6, 9, 11, 13, 15, 17, 19) and three different items (1, 3, and 16 rather than 2, 5, and 7). Thus our results would not have been markedly different, especially if we had considered item content in addition to the PCA results.

Indeed, in many applications, the value added by IRT is not in the end result per se, but rather in the process taken to reach conclusions. Results from an IRT calibration contain detailed item-level information that can be considered from many useful perspectives (e.g., ICCs, item and test information plots, rank ordering of item slopes). The set of interpretive tools associated with IRT applications allows researchers, clinicians, and scale developers to work together in an informed manner in a way that is not readily achievable with a traditional CTT approach.

Despite our assertion that IRT offers several advantages over CTT, we do not advocate the abandonment of CTT methods in favor of IRT. In fact, insights from IRT analyses are most useful when they are complemented by a familiarity with the basic properties of the data from classical analyses, as well as input from content experts, which is essential if the implications of results are to be evaluated according to their statistical as well as their clinical significance.

Our example highlights the complementary relationship between IRT and CTT methods in several ways. First, we examined the basic properties of the items using traditional descriptive statistics and used CTT statistics (i.e., Cronbach's alpha, average item–total correlation) to determine that a more parsimonious representation of the item responses that collapsed the two most extreme categories would be appropriate. We also evaluated the performance of the short form relative to the original by comparing the classification rates of the two forms based on observed total score cut points. In addition, we used EFA and CFA to ensure that the set of items was sufficiently unidimensional for application of IRT.

Incidentally, there is some debate over the use of both EFA and CFA approaches for the purpose of evaluating the unidimensionality assumption that merits discussion. In cases where the dimensionality of the instrument is not well established, it is most reasonable to examine this question with EFA. In the absence of underlying theory, results from an EFA can provide information about possible additional factors that may be present in the item set. However, the lack of formal fit evaluation for EFA solutions leads some to favor CFA when there is theory or prior knowledge of the instrument's factor structure. It is our opinion that even in cases where there is some prior knowledge of dimensionality, as was the case in this example, the use of EFA in conjunction with CFA can be more informative than using CFA alone. For example, whereas the modification indices from a 1-factor CFA do not include consideration of the addition of correlated errors, the results from a multi-factor EFA can reveal the potential need for these parameters. In our example, results of the EFA informed model modifications in the CFA and led to a pointed examination of the local dependence assumption in the IRT application. Of course, it is not appropriate to perform both EFA and CFA on the same sample. Therefore, these approaches should only be used in a complementary manner when the sample is large enough to be split into two for this purpose.

The primary purpose of our example was to demonstrate the utility of IRT in selecting a shortened scale. Our available data allowed us to set aside half the sample to compare the short form to the original. However, the 19-item Feelings Scale should not be considered a gold standard for identifying depression in adolescents, and the cut scores we selected were approximate. Thus our cut score examination should not be considered evidence for discriminant validity. Full validation of the short scale would require extensive analyses not included in this example [69]. For instance, it is critical to evaluate the performance

of a short form in a completely independent sample of respondents who are administered only the short form items (as opposed to having received the full item set).

This paper provided a general introduction to IRT to familiarize potential new users with this methodology. Although this introduction is brief, hopefully it conveyed the potential that IRT has to offer to health measurement and will encourage readers to explore this methodology further. More detailed discussions and comparisons of CTT and IRT are available [15, 70, 71] and several of the publications referenced in this paper may also be helpful resources for readers who wish to learn more about applying IRT.

## References


1. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J.-S., & Cella, D. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care,* in press.
2. Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341–349.
3. Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
4. Lord, F. M., (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Earlbaum.
5. Wainer, H., Dorans, N. J., Flaugher, R. et al. (1990). *Computerized adaptive testing: A primer*. Hillsdale NJ: Lawrence Earlbaum Associates.
6. Abrahamowicz, M., & Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika, 57*(1), 5–27.
7. Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics, 27*(3), 291–317. .
8. Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25–36.
9. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
10. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monography, 34*.
11. 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.
12. Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577.
13. Hambleton, R. K., Lipscomb, J., Gotay, C. C., & Snyder, C. (2005). Applications of item response theory to improve health outcomes assessment: Developing item banks, linking instruments, and computer-adaptive testing. In Outcomes assessment in cancer: Measures, methods, and applications (pp. 445–464). Cambridge University Press.
14. Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, (this issue).
15. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
16. Cattell, R. B. (1966). The screen test for the number of factors. *Multivariate behavioral Research, 1*, 245–267.
17. Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum.
18. Loehlin, J. C. (1987). *Latent variable models*. New Jersey: Lawrence Erlbaum Associates.
19. Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.
20. Teresi, J., & Fleishman, J. (2007). Assessing measurement equivalence across populations: Differential item functioning (DIF). *Quality of Life Research*, (this issue).
21. 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.
22. Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*:561–573.
23. Andrich, D. (1978). Application of a psychometric rating model to ordered categories, which are scored with successive integers. *Applied Psychological Measurement, 2*, 581–594.
24. Muraki, E. (1992). A generalized partial credit model: Application of the EM algorithm. *Applied Psychological Measurement, 16*, 159–176.
25. Muraki, E. (1997). A generalized partial credit model. In: van der Linden W & Hambleton RK (eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer.
26. Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.
27. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.
28. Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164–184.
29. Du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood IL: Scientific Software International.
30. Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.
31. Ramsay, J. O. (1995). *TestGraf – a program for the graphical analysis of multiple choice test and questionnaire data [computer software]*. Montreal: McGill University.
32. Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.
33. Anderson, E. (1973). A goodness of fit test for the rasch model. *Psychometrika, 38*, 123–140.
34. Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53*(4), 525–546.
35. Rost, J., & von Davier, M. (1994). A conditional item-fit index for rasch models. *Applied Psychological Measurement, 18*, 171–182.
36. 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.
37. Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23–48.
38. McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 19*, 49–57.
39. Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262.

40. Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.
41. Orlando, M., & Thissen, D. (2003). Further examination of the performance of *S-X*$^2$, an item fit index for dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298.
42. Bjorner, J. B., Christensen, K. B., Orlando, M., & Thissen, D. (2005). Testing the fit of item response theory models for patient reported outcomes. Poster presented at the annual meeting of the International Society of Quality of Life Research. San Francisco, CA, October (2005). .
43. Drasgow, F., Levine, M. V., Tsien, S. et al. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143–165.
44. Kingston, N., & Dorans, N. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*, 281–288.
45. Mislevy, R. J., & Bock, R. D. (1986). *Bilog: Item analysis and test scoring with binary logistic models*. Mooresville, Indiana: Scientific Software.
46. Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher et al. (Eds.), *Computerized adaptive testing: A primer* (pp. 65–101). Hillsdale NJ: Lawrence Earlbaum Associates.
47. Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277–298.
48. McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*(2), 121–137.
49. Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29*(1), 26–44.
50. Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.
51. Linacre, J. M. (1994). Sample size and item calibration stability, *Rasch Measurement Transactions, 7*(4), 328.
52. Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371–390.
53. 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), 50–59.
54. Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118–128.
55. Thissen, D. (2003). Estimation in multilog. In M. du Toit (Ed.), *IRT from SSI: Bilog-MG, multilog, parscale, testfact*. Lincolnwood, IL: Scientific Software International.
56. Bearman, P. S., Jones, J., & Udry, J. R. (1997). http://www.cpc.unc.edu/projects/addhealth/design/html, The National Longitudinal Study of Adolescent Health: Research Design.
57. Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401.
58. Goodman, E., & Capitman, J. (2000). Depressive symptoms and cigarette smoking among teens. *Pediatrics, 106*, 748–755.
59. McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring*. Mahwah, New Jersey: Lawrence Earlbaum & Associates.
60. Stout, W. A. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 28.
61. Muthén, L. K., & Muthén, B. (1998–2004). Mplus user's guide. Los Angeles, CA: Muthen & Muthen.
62. Steiger, J. H., & Lind, J. (1980). Statistically based tests for the number of common factors. Paper presented at the Psychometrika Society Meeting, Iowa City.
63. Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.
64. Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246.
65. Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Kollen & J. S. Long (Eds.), *Testing structural equation models*. Thousand Oaks, CA: Sage.
66. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.
67. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57*, 289–300.
68. 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.
69. Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111.
70. Reeve, B. B., & Mâsse, L. C. (2004). Item response theory modeling for questionnaire evaluation. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Sinter (Eds.), *Methods for testing and evaluation survey questionnaires* (pp. 247–273). Hobeken, NJ: Wiley.
71. Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: Comparison with the classical test theory approach. *Health Education Research, 21*(1), i19–i32.