Published in final edited form as:

Qual Life Res. 2009 May ; 18(4): 447–460. doi:10.1007/s11136-009-9464-4.

Having a Fit: Impact of Number of Items and Distribution of Data on Traditional Criteria for Assessing IRT's Unidimensionality Assumption

Karon F. Cook, PhD,

Department of Rehabilitation Medicine, University of Washington, Seattle, WA, 801 Cortlandt St, Houston, TX 77007, karonc2@u.washington.edu 713.291.3918

Michael A. Kallen, PhD, MPH, and

Department of General Internal Medicine, University of Texas M. D. Anderson Cancer Center, PO Box 301402, Houston, TX 77230-1402

Dagmar Amtmann, Ph.D

Department of Rehabilitation Medicine, University of Washington, Box 357920, Seattle, Washington 98195-7920

INTRODUCTION

For health outcome domains such as fatigue, pain, physical functioning, and emotional functioning there exist numerous instruments purporting to measure the same or similar constructs. The use of different measures in different studies produces results that are difficult to compare and restrict generalizability across studies. An alternative is to develop banks of items measuring outcomes of interest and calibrate responses to an item response theory (IRT) model.[1,2] Once an item bank is developed, tested, and calibrated, and assuming that the data fit the IRT model, a number of assessment approaches become possible, including computerized adaptive tests and short forms. Multiple short forms that target particular measurement settings, goals, and demands can be developed while producing scores on a common mathematical metric. Items also can be administered dynamically using computer adaptive testing (CAT) to achieve greater measurement precision and reduced response burden. [3–6]

The potential benefits of having a calibrated item bank to support the measurement of patient reported outcomes (PROs) are considerable, but so are the challenges in developing such a bank. In addition to defining the construct of interest, developing and testing items in target populations, and evaluating the soundness of the final item bank with respect to its psychometric properties, it is essential to ensure that item parameter estimates obtained in the IRT calibration are accurate. The accuracy of parameter estimation depends substantially on the degree to which data meet the assumptions of IRT and fit the chosen IRT model. To date, the vast majority of item banks have been calibrated using IRT models that assume a single, latent trait drives how persons respond to items. Health constructs have considerable conceptual complexity, however, and often require a broad range of indicators; item banks developed to measure health constructs never perfectly meet strictly-defined unidimensionality assumptions.[7–10] The pertinent question is whether the presence of secondary dimensions threatens the accuracy of unidimensional IRT calibrations.

Factor analytic approaches are commonly used to evaluate dimensionality in the context of IRT. In this paper we briefly review confirmatory factor analytic approaches and published fit statistic standards. We also briefly describe parallel analysis and bifactor analysis. We then present results of an evaluation of the impact of number of items and of non-normal trait distribution on traditional factor-analysis-based fit standards for unidimensionality. The data were responses to a bank of items developed by the Patient Reported Outcome Measurement Information System (PROMIS)[11] to measure pain impact across multiple diseases and conditions. In addition to traditional factor analytic methods, we describe and apply bifactor models for representing the dimensional structure of the PROMIS pain impact data and argue that such models complement traditional factor analytic approaches to assessing unidimensionality.

EVALUATIONS OF DIMENSIONALITY FOR IRT

FACTOR ANALYTIC APPROACHES TYPICALLY USED TO EVALUATE DIMENSIONALITY

Factor analysis is commonly used to investigate whether item responses are “unidimensional enough” for calibration using a unidimensional IRT model.[7,12–14] Though exploratory factor analysis (EFA) often is used, confirmatory factor analysis (CFA) has the distinct advantage of allowing hypothesis testing.[14] Despite its common application for evaluating unidimensionality, EFA does not allow for direct testing of the latent construct.

When CFA is used to test IRT’s unidimensionality assumption, a model is specified in which all items load on a single factor. Typically, the fit of the model is evaluated using factor model fit statistics and traditional cutoff values for these statistics.[14] A potential problem with using factor analytic approaches to test the unidimensionality assumption of IRT is that factor analyses themselves make certain assumptions. These include, for example with maximum likelihood-based estimation, multivariate normality of the data, an assumption seldom met by PRO data because their distributions often are quite skewed. Though some research has suggested that factor analytic approaches are robust to violations of normality, [15] other studies indicate potential problems.[16]

An additional challenge to the use of CFA fit statistics in evaluating the unidimensionality of PRO item banks is the fact that, as Floyd and Widaman[17] have noted, “It may be unreasonable to expect that lengthy questionnaires with many items assessing each factor will show satisfactory solutions when the individual items are submitted to confirmatory factor analysis” (p. 293).[17] Difficulty obtaining “satisfactory solutions” with longer static instruments may be challenging, but in the context of item bank development, the challenge is greatly magnified. Because of concerns about response burden, a static PRO instrument with 15–20 items could be considered a rather long instrument in many measurement contexts. [18] But with item banking and CAT administration, response burden is managed through adaptive item selection. The efficiency of CAT is obtained by having an item bank with a large number of items that collectively extend across the measurement continuum. Item banks seldom have fewer than 35 items and can have many more than that. For example, the PROMIS physical function item bank has 125 items (<http://www.nihpromis.org>). With large numbers of items there are many opportunities for subsets of items to have shared variance not accounted for by the dominant trait.

If CFA fit statistics are overly sensitive to number of items and inconsequential multidimensionality, their appropriateness for evaluating IRT’s unidimensionality assumption in PRO item banks diminishes. This is especially the case if, rather than using CFA as a tool for evaluating level of unidimensionality, it is used primarily to generate fit statistic values that are mechanically compared to published criteria invoked as arbiters of whether data are unidimensional “enough” for calibration using unidimensional IRT models.

CONFIRMATORY FACTOR ANALYSES FIT STATISTICS AND CRITERIA

Statistical methods for judging model fit, conducting model comparisons, and evaluating the adequacies of specific models have been applied successfully in structural equations modeling (SEM). McDonald and Mok[19] demonstrated the possibility of using SEM-based indices to inform evaluations of dimensionality in the context of IRT. From this work evolved the use of fit indices and associated rules-of-thumb to judge data dimensionality in the context of CFA.

A number of fit indices have been suggested for evaluating model fit with CFA.[20–25] The χ^2 statistic is a “badness-of-fit” index in that larger values represent worse fit. It measures the discrepancy between the observed covariance matrix and the matrix the model predicts. When the model holds, this statistic follows a χ^2 distribution; however, its value is exceptionally sensitive to sample size and usually indicates significant misfit even for models that fit well for practical purposes.[24]

Other fit indices are descriptive and though many are designed to range from zero to one, their sampling distributions are not well-specified.[22] The Tucker Lewis Index or Non-Normed Fit Index (NNFI) [26] and Comparative Fit Index (CFI) [27] both are incremental relative fit indices. They estimate differences between the examined model and a hypothetical (null) model where none of the components in the model are related. The NNFI differs from the CFI in that it penalizes lack of parsimony in the hypothesized model.

The Root Mean Square Error of Approximation (RMSEA)[21] is complimentary to the NNFI and CFI in the sense that it estimates the difference between the examined model and a hypothetical model where every component in the model is related to every other component. [21] The RMSEA provides an answer to the question, “How well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?”[21] It is based on an estimate of the population discrepancy function, which assesses the error of approximation in the population. The RMSEA is thus a measure of discrepancy; because it presents discrepancy per degree of freedom, it is sensitive to model complexity (i.e., number of estimated parameters).[28] The RMSEA is based on *covariance* matrices and is affected by the metric of the input variables.

The Standardized Root Mean Square Residual (SRMR), proposed in 1995[29] is based on average differences between observed and predicted *correlation* matrices. It represents the average of all standardized residuals and can be interpreted as being the average discrepancy between the correlation matrices of the observed sample and the hypothesized model.

The most recent addition to the root mean square residual “family” is the Weighted Root Mean Square Residual (WRMR) introduced by Muthen and Muthen. The WRMR uses a variance-weighted approach specially suited for models whose variables measured on different scales or have widely unequal variances. (23, 24) The WRMR has been tested with categorical variables, and its developers suggest the WRMR is also highly appropriate for data that are not distributed normally.[24,25]

It is important to note that the fit indices described here, and other proposed fit indices, may not perform uniformly across conditions. They may vary by factors such as parameter estimation method, sample size, or matrix analyzed (covariance vs. correlation). The χ^2 statistic presents a recognizable example. With some parameter estimation methods (e.g., maximum likelihood, generalized least squares, weighted least squares), the χ^2 statistic may be correctly distributed as a chi-square distribution; with other estimation methods (e.g., unweighted least squares, diagonally weighted least squares), the χ^2 statistic needs to be adjusted.[30] Even with a correct (or corrected) chi-square value, however, it may prove unrealistic to assume that a model tested holds exactly, rather than approximately, in the population. In this case an

obtained chi-square value is more appropriately compared with a non-central than with a central chi-square distribution.[28]

FIT STATISTIC “STANDARDS”

Much is yet to be learned about the performance of fit indices particularly regarding their use with samples of small and moderate size, with models of different types and complexity, with variables of differing distributional characteristics, and with the increasing number of parameter estimation methods available. But despite the challenges to “universal” application of fit indices, standards for CFA fit statistics have been proffered and debated. In 2007, a description of the PROMIS analysis plan was published.[31] The originally submitted manuscript did not list criteria for what constituted “good enough” fit to a unidimensional model based on factor analytic results, but at the insistence of the editors, PROMIS investigators offered such standards. Based on prior published criteria by Hu and Bentler, [22,23,27] McDonald, [32] and others, [21,33,34] the following standards for “good fit” were offered: CFI >0.95, RMSEA <0.06, NNFI >0.95, and SRMR <0.08.[31] For the WRMR values <1.0 have been suggested as indicative of adequate model fit.[25]

CONFIRMATORY FACTOR ANALYSES OF ORDINAL DATA

An additional challenge to factor analytic evaluations of item bank data is the fact that the measurement of patient reported outcomes is typically based on categorical item responses. Conducting CFAs of ordinal data requires departures in thinking and methodology from approaches designed for interval or interval-like data. To analyze continuous data, covariance or product-moment matrices can be computed and model parameters correctly estimated using, for example, maximum likelihood or generalized least squares methods. Following this approach with categorical data, however, may produce misleading results including distortions in parameter estimates and incorrect standard errors, χ^2 statistics, and chi-square-based goodness-of-fit measures.[35] Advances in weighted least squares estimation offer researchers sophisticated methods for investigating ordinal data, based on the computation of a polychoric correlation matrix. However, the performance of fit indices under such “new” estimation methods and other challenging modeling conditions has not yet been fully investigated.[36]

Parallel Analysis—Parallel analysis is a statistical approach for evaluating the optimal number of factors/components in a factor analysis or principal component analysis.[37] In this approach, values of eigenvalues expected by chance alone are compared to values of observed eigenvalues to help determine the justifiable number of components or factors. Factors with eigenvalues greater than those expected by chance are extracted. Available computer algorithms allow for simulations based on normal distributions or based on the distribution of the observed data.

Bifactor Analysis—Reise and colleagues have pointed out that multidimensionality can occur when there is item content redundancy and when there are diverse indicators of a complex construct.[38,39] They recommend the use of bifactor analysis, arguing that comparing unidimensional and multidimensional models using confirmatory and exploratory factor analyses is a poor method of judging whether a dataset is sufficiently unidimensional for IRT analysis.

In a bifactor model, all items are allowed to load on a general factor that is assumed to be the latent trait being measured.[39,40] In addition, responses to items may load on zero, one, or more group factors that may or may not be correlated. In most practical applications, responses to items are modeled as loading on a general factor and on one (and only one) group factor; the general and all group factors are modeled as being orthogonal to one another (i.e., uncorrelated).

Based on the factor loadings obtained for the general and group factors, estimates can be made about the “saturation” of the data by the general factor and the relative impact of group factors. In addition, the factor loadings from a one-factor model can be compared to loadings on the general factor in a bifactor model to assess the level of disturbance due to multidimensionality in the data. An advantage of the bifactor model over a second order factor model is that unlike in the latter, in the bifactor model the relation between an item and the general factor is not constrained to be proportional to the relation between the first and second-order factors (“proportionality constraint”). Thus, the relationship between items and factors is simpler to interpret than in a second order model. Reise and colleagues argue that bifactor analysis is an alternative to non-hierarchical multidimensional models, useful for estimating the level of distortion that results when multidimensional data are modeled using unidimensional models, and helpful in deciding the relative merits of developing subscales.[39,40]

METHODS

MEASURE

In 2004 the National Institutes of Health (NIH) funded the PROMIS initiative.[11] This initiative established a national collaborative network to create a publicly available system for measuring PROs. Initial efforts by the PROMIS network focused on measuring the domains of pain, fatigue, emotional distress, physical function, and social function, as well as their subdomains (14 subdomains in all). Details of this initiative have been reported elsewhere. [11,31,41,42] Candidate item banks were developed to measure each subdomain, including Pain Impact. The items of the Pain Impact bank were used for the current study. The time frame for these items was, “In the past seven days...” Based on psychometric analyses and consultation with content experts, a 56-item candidate item bank was reduced to 47-items (Appendix). (Note: Description of a more recently revised PROMIS Pain Impact item bank is available on the PROMIS website: <http://www.nihpromis.org>).

OBSERVED AND SIMULATED DATA

Analyses were conducted both on simulated and observed full datasets and item subsets. The names and characteristics of each item set and subset are described in Table 1.

47-Item Sets-PROMIS Pain Impact Data—An extensive data collection effort was undertaken in 2007 to obtain responses to candidate PROMIS item banks.[31] Responses to the pain impact items were gathered using two sampling arms. In the “bank” arm, participants completed all 56 items of the candidate Pain Impact bank. A total of 944 respondents met inclusion criteria, namely: 1) did not give repetitive strings of ten or more non-extreme responses (e.g. responding “3” to 10 items in a row) and 2) had a response greater than one second per item. In the other arm (“block arm”), participants completed subsets of 7 items representing each of the PROMIS subdomains. There were a total of eight non-overlapping seven-item Pain impact subsets to account for the total of 56 items. A total of 14,584 persons in the block arm met above inclusion criteria. Participants were recruited from the PROMIS network’s primary research sites and by Polimetrix (www.polimetrix.com, also see www.pollingpoint.com), a polling firm based in Palo Alto, California.

After initial evaluations and review by content experts, the 56-item bank was reduced to 47 items (Appendix). Responses to these items were calibrated using Multilog (version 7)[43] and Samejima’s graded response model (GRM).[44] For calibration purposes, respondents in both the bank (those administered 56 pain impact items) and block arms (those administered 7-item subsets) were included. Fit was evaluated using IRTFIT, [45] a SAS Macro that produces a number of statistics that evaluate fit to IRT models, including polytomous extensions of the S-X² (Pearson’s chi-square) and the S-G² (likelihood ratio) statistics.[46] With these two

statistics, the predicted distribution of item responses for each simple sum score level is compared to observed frequencies. To account for multiple tests, we chose a significance level of 0.01 to indicate item misfit. Of the persons in the bank arm, 745 had no missing responses to the 47 items. These data (PROMIS₄₇) constituted the observed dataset to which we compared results from simulations. We limited the analyses to persons with no missing responses because, as described below, comparison item subsets were “constructed” out of the full 47-item set. By limiting the sample to persons who responded to all items, we were able to create cleaner comparisons of number of items (e.g., all 10-item subsets had 10 responses per record).

Simulated 47-Item Sets—Using WinGen2[47,48], we simulated 40,000 sets of responses to the 47 Pain Impact items based on the parameters of the GRM calibration of the PROMIS Pain Impact items. We then obtained scores (theta) for each record based on the item parameters using Multilog (version 7)[43]. We divided both the observed score distribution (range = -1.06 to 3.01 logits) and the 40,000 simulated scores into 19 theta ranges of 0.203 logits each. To create SIM.PROMIS₄₇, we randomly selected from the simulated data the number of observations in each range equal to the number observed in the PROMIS distribution in the same range. This insured that the simulated distribution approximately mirrored the observed distribution. Because there was one observed score in the uppermost theta range, but no simulated scores in this range, one extra simulee was randomly selected from the penultimate theta range to be included in SIM.PROMIS₄₇. A normal distribution of simulated responses (SIM.NORM₄₇) was obtained by randomly selecting simulees in the theta ranges in proportion to the number expected in a normal distribution with similar range.

Ten-Item CATs—We simulated a CAT administration based on the PROMIS₄₇ data. The simulation was accomplished using Firestar (version.1), a computer program for simulating CAT with polytomous items.[49] We simulated a fixed-length CAT of ten items using minimum expected posterior variance to select items and expected a priori (EAP) estimation to calculate theta. All respondents received the same first item. We identified each unique subset of ten items administered during the simulated CAT and identified the ten most frequently administered ten-item subsets. We then created datasets by extracting responses to these ten, ten-item CATs for each respondent from the PROMIS₄₇, SIM.PROMIS₄₇, and SIM.NORM₄₇ datasets. For convenience we refer to these 10-item CATs, respectively, as PROMIS-CAT_{a-j}, SIM.PROMIS-CAT_{a-j}, and SIM.NORM-CAT_{a-j} (see Table 1).

Randomly-Selected 10-Item Subsets—It could be argued that items selected on the basis of a CAT simulation represent the “best of the bank”. Therefore, the factor analytic results obtained using these item subsets could be affected not only by the number of items, but by the quality of the items. To distinguish the impact of item number and item quality, we randomly selected ten subsets of ten items each from the original PROMIS₄₇ item set. We refer to these as PROMIS-RAND₁₋₁₀ (see Table 1).

Item Subsets with Fewer than Ten Items—To investigate the impact on factor analytic results of having fewer than ten items, we constructed 5-, 6-, 7-, 8-, 9-, and 10-item subsets (PROMIS-PROMIS-SUB₅₋₁₀) based on the observed data (PROMIS₄₇). These subsets were constructed based on the most frequently selected items in the 10-item CAT simulation. We eliminated the starting item in the CAT simulation because it was administered to all respondents. From the remaining items, we selected the five items most often chosen in the CAT simulation to constitute PROMIS-PROMIS-SUB₅. For PROMIS-PROMIS-SUB₆, we selected the six most often chosen items. We followed the same procedure to identify items for PROMIS-PROMIS-SUB₇₋₁₀ (see Table 1).

ANALYSES

Unidimensional CFA's were conducted to obtain fit statistics for data based on each of the observed and simulated item sets. For each, the polychoric correlation matrices were analyzed using MPlus software.[24] Weighted Least Squares with Mean and Variance adjustment (WLSMV) was used as the parameter estimation method. Results from the following fit statistics were collected and compared: the χ^2 test of model fit, [22] the NNFI, [26] the CFI, [27] the RMSEA, [21] the SRMR, [33] and the WRMR.[24,25]

We conducted a parallel analysis of the raw PROMIS data based on principal axis/common factor analysis and on the distribution of the raw data set.[37] The size of eigenvalues expected by chance alone were computed and compared to observed eigenvalues to help determine the maximum number of underlying dimensions. Based on the results, we conducted bifactor analysis allowing all items to load on a general factor and on the single group factor for which the item had the highest loading. We estimated the saturation of the models by the general factor by calculating the ratio of the variance accounted for by the super-ordinate factor to total observed variance in the data.[50] In addition, we compared the magnitude of factor loadings based on the first-order CFA and on the bifactor models. [38,39]

RESULTS

DATASETS

PROMIS₄₇ Sample Characteristics—Demographics for the PROMIS₄₇ sample are reported in Table 2. Participants were allowed to indicate more than one category of race/ethnicity. The sample was largely white/not Hispanic (73.3%) and female (54.4%). Age of participants ranged from 18–90, with a mean of 51 years and a standard deviation of 19 years. More than half the sample reported having a college degree or higher.

IRT Calibration and Evaluation of Fit—Multilog centers the calibration metric on persons; thus, the latent mean for the calibration sample was zero with standard deviation of one. However, as described above, PROMIS₄₇ was a subsample extracted from the calibration sample that included persons who responded to seven-item subsets. The mean of the theta distribution of PROMIS₄₇ varied slightly from the calibration sample with a mean of -0.04 with standard deviation (SD) of 0.81. The mean of the theta distribution of SIM.PROMIS₄₇ was -0.03 with SD of 0.80. The mean of SIM.NORM₄₇ was 1.07 with SD of 0.52.

Item content and item parameter estimates are reported in the Appendix. Item calibration was accomplished using both participants who took all the Pain Impact items (bank arm) and those who took a subset of seven items (block arm). Pain Impact items were scored so that higher scores indicated greater pain. Threshold estimates ranged from -0.5 to 2.8. Discrimination parameters ranged from 2.4 to 6.5 (mean=4.1). No items were found to be misfitting. Probability values for S-G² statistics ranged from 0.020 to 1.00 (mean = 0.686). Probabilities for S-X² statistics ranged from 0.013 to 1.000 (mean = 0.659). The distribution of theta estimates for PROMIS₄₇ is displayed as Figure 1. Skewness was 0.418; kurtosis was -0.216 . The distribution reflects the fact that 25% (183) endorsed the lowest category (indicating least impact) for all items. On the other end of the continuum, one person endorsed the highest category for every item (indicating greatest impact).

SIMULATION OF CATS

CAT with a stopping rule of ten items was simulated using the PROMIS₄₇ dataset. The correlation between the 10-item CAT scores and full scores was 0.97. Among the subsets of ten items administered to 745 respondents, there were 46 unique, ten-item CATs. The items

of the ten most frequently administered subsets of items (CAT_{a-j}) are reported in Table 3. PI56 was in all subsets because it was the starting item for the CAT simulation. Of the respondents in the CAT simulation, 83.0% received one of the ten most frequently administered CATs.

As is evident in Table 3, many of the ten most frequently occurring CATs varied little with respect to the specific items included. For example, CAT_c and CAT_d are identical except for one item. There were 21 unique items among the items of CAT_{a-j} , and of these, seven appeared in more than half of these 10-item CATs. This similarity in CAT item composition across the most frequently occurring CATs is typical because the CAT algorithm favors the most highly discriminating items. In fact, the average discrimination of the nine most frequently administered items (excluding the starting item) was 5.2. The average discrimination of the remaining items was 3.8. There was variability in item usage, however. All but twelve items were selected in at least one of the simulated CATs.

To select items to comprise the five- through ten-item subsets (PROMIS-PROMIS-SUB₅₋₁₀), we computed the number of times each item was administered in the CAT simulations. The most frequently administered items were used to construct PROMIS-SUB₅₋₁₀. In order of frequency, these items were the starting item, item PI56 ($n=745$, 100.0%); PI3 ($n=574$, 77.0%); PI10 ($n=503$, 67.5%); PI9 ($n=485$, 65.1%); PI22 ($n=461$, 61.9%); PI24 ($n=460$, 61.7%); PI12 ($n=436$, 58.5%); PI20 ($n=410$, 55.0%); PI5 ($n=384$, 51.5%); and PI39 ($n=359$, 48.2%). So, for example, PROMIS-SUB₅ included responses to PI56, PI3, PI10, PI9, and PI22 extracted from the PROMIS₄₇ dataset.

FIRST ORDER CONFIRMATORY FACTOR ANALYSIS

Summary CFA fit statistics for all datasets are reported in Table 4. The probabilities for the chi-square statistics for all 47-item datasets, CAT item sets, and PROMIS-SUB₅₋₁₀ indicated poor fit. All were <0.001 and are not reported in the table.

CFA Fit Values for 47-Item Datasets—If published fit criteria are used as arbiters, the fit of the PROMIS₄₇ data to a unidimensional model depends upon choice of fit statistic. The NNFI/TLI was very high (0.989) and well beyond the >0.95 criterion for good fit. Likewise, the SRMR value of 0.062 was well below the published criterion of <0.08 . However, based on CFI, RMSEA, and WRMR values, the PROMIS₄₇ data would be judged insufficiently unidimensional for calibration using an IRT model. Notable in particular are the values of the RMSEA and WRMR for the PROMIS₄₇ data. The RMSEA of 0.159 and WRMR of 2.837 are well over twice as large as the values of published criteria.

As expected, the data simulated to fit a normal distribution and the GRM conformed to fit criteria for the CFI, NNFI, RMSEA, SRMR, and WRMR. The values of fit statistics for SIM.PROMIS₄₇, a dataset simulated to fit the GRM and to mimic the PROMIS₄₇ distribution, were minimally worse than those obtained for SIM.NORM₄₇, suggesting minimal impact of distribution on the 47-item datasets. As reported above, the chi-square criterion was not met for any of the datasets.

CFA Fit Values for Observed and Simulated 10-Item CATs—Several CFA fit statistics for 10-item CATs based on the PROMIS₄₇ data subsets (PROMIS-CAT_{a-j}) were substantially better than those for the full 47-item bank. For example, the mean CFI value for the most frequently administered 10-item CATs was 0.980 (compared to 0.913 for PROMIS₄₇). Only one of the CATs had a CFI value below, albeit minimally, the published criteria of >0.95 (PROMIS-CAT_a CFI=0.949). Neither the PROMIS-CAT_{a-j} mean RMSEA of 0.157 nor the mean WRMR of 2.373 conformed to the published fit criteria of <0.06 and <1.0 , respectively; however, the WRMR was closer and the RMSEA was substantially closer than the values for the full 47-item dataset. The mean SRMR value for PROMIS-CAT_{a-j}

(0.034) was almost half the value obtained for the full 47-item PROMIS dataset (0.062). The RMSEA values were an exception. There was little difference between the mean RMSEA value for PROMIS-CAT_{a-j} (0.157) and the value obtained for the full 47-item PROMIS dataset (0.159).

With the exception of the chi-square statistic, all fit statistics for each of the SIM.NORM-CAT_{a-j} conformed to published criteria. Results for the ten-item datasets were very similar to those obtained for SIM.NORM₄₇. In contrast to the comparison between results for PROMIS₄₇ and PROMIS-CAT_{a-j}, number of items had little observable impact on fit values for these data simulated both to fit the GRM and to be normally distributed as assumed in factor analysis.

The SIM.PROMIS-CAT_{a-j} results provide an interesting contrast because they are based on data that fit the model but have a skewed distribution that mirrors the observed data. All SIM.PROMIS-CAT_{a-j} results conformed to published CFI, NNFI, and SRMR criteria, but all did not meet RMSEA and WRMR criteria. RMSEA values ranged from 0.057 to 0.136, with values for four of the ten above the published cut off of <0.06. WRMR values ranged from 0.660 to 1.971, with values for three of ten above the published cut off of <1.0. These results suggest that the RMSEA and WRMR are affected by the distribution of the data—even data generated to fit the GRM model, and that the direction of influence when data are skewed is toward putative indication of poorer fit. This result for the WRMR is especially notable because it has been reported to be robust to the impact of data skewness. [24,25] There was no similar effect of skewness on observed values of CFI, TLI, or SRMR values. However, because the values for these statistics were at the end of the range in each case, it is not possible to ascertain whether this result proceeds from the robustness of these statistics or their lack of sensitivity.

CFA Fit Values for Randomly Selected 10-Item Subsets—The average fit values for the randomly-selected 10-item subsets (PROMIS-RAND_{a-j}) were better than those for the full 47-item PROMIS bank and similar to the values for PROMIS-CAT_{a-j}. This suggests that the better fit of the 10-item CATs compared to the full bank was not due to the CAT algorithm choosing the “best” and, presumably, better-fitting items.

CFA Fit Values for PROMIS-SUB₅₋₁₀—The results for subsets of items comprised of five through ten items indicated an impact of number of items on values of the fit statistics. CFI, RMSEA, SRMR, and WRMR values improved with each reduction in the number of items from ten to six items. The five-item subset was an exception. All fit values were the same or worse for the five-item set. NNFI values improved with each reduction from ten items to nine and to eight items, but not for seven items. The trend toward better fit values with fewer items is expected since fewer items allow fewer opportunities for shared variance between pairs of items. The values of the five-item subset varied from this trend. The impact of having such a small number of items to identify the factor may have outweighed the impact of decreased opportunity for unmodeled, shared variance between item pairs. Though the trend toward better fit with fewer items did not follow with the reduction from six to five items, all values for PROMIS-SUB₅ (as well as PROMIS-SUB₆₋₁₀) were substantially better than those obtained for the 47-item bank. For example, for the 5-item subset, the RMSEA value was 0.108, compared to 0.159 for the full 47-item bank; and the WRMR value for the 5-item subset (1.357) was less than half that of the full bank (2.873)

PARALLEL ANALYSIS

The first three eigenvalues generated from the parallel analysis were 30.7, 2.2, and 1.4. The first three PROMIS₄₇ data eigenvalues were 36.2, 2.1, and 1.2. These results support a one-factor solution because only the first PROMIS₄₇ eigenvalue is greater than those generated in

the parallel analysis. Recall that the parallel analysis algorithm accounted for the skewness of the distribution as it was based on the distribution of the observed data.[37] The results support a one-factor solution.

BIFACTOR ANALYSIS

We conducted two bifactor analyses. One modeled a general factor and two group factors. The other modeled a general factor and three group factors. The particular group factor each item loaded on in these analyses is reported in the Appendix. The results were similar for the model with two and with three group factors. In the model with two group factors, the group factors accounted for 10.0% of the total variance (11.9% of the common variance); the general factor accounted for 74.4% of the total variance (88.1% of the common variance). In the model with three group factors, the group factors accounted for 10.4% of the total variance (12.2% of the common variance); the general factor accounted for 74.8% of the total variance (87.8% of the common variance). For both the two and three group bifactor models, the loadings on the general factor were very similar to the factor loadings of a first order one-factor solution. The absolute differences ranged from 0.00 to 0.06, suggesting little disturbance in factor loadings by unmodeled secondary factors. Fit values for the two and three group bifactor models were, respectively, CFI = 0.923, 0.934; NNFI/TLI = 0.995, 0.995; RMSEA = 0.109, 0.102; SRMR=WRMR = 1.596, 1.519.

DISCUSSION

This study explored the application of CFA fit statistics in a very specific context—assessing whether an item bank dataset is “unidimensional enough” for modeling using IRT. Though reasonable on the face of it, such an application encounters challenges in the context of item banking where responses to large numbers of items are modeled and where data often are skewed. Thus we designed a study to demonstrate how fit statistics values are influenced under these conditions.

The results call into question the common practice in item bank development of using published CFA fit standards as arbiters of whether data are “unidimensional enough” for calibration using unidimensional IRT models. Relying on these standards to determine whether responses to the PROMIS Pain Impact bank met IRT’s unidimensionality assumption would leave us in a quandary since our conclusion would depend upon the particular fit statistics invoked. The bank could be argued to have very good fit, based on the NNFI/TLI value of 0.989, or very poor fit based on the RMSEA of 0.159 and WRMR of 2.873.

The CFI and SRMR values for the real data were substantially improved with reduction in the number of items in the bank. RMSEA and WRMR values also improved, but not as substantially as those of the CFI and SRMR. Because the design of our study did not allow us to separate the effects of number of items from the effects of multidimensionality, we cannot say what drives these differences in fit statistic values. It may be that RMSEA and WRMR values are more robust than those of CFI and SRMR to reduction in number of items. However, an alternative conclusion is that CFI and SRMR are more sensitive to reduction in number of items, and RSMEA and WRMR are more sensitive to the presence of secondary dimensions. The bifactor results give some support for the latter conclusion. All fit statistics improved for the 2- and 3-group bifactor models, but the improvements were greater for the RMSEA and WRMR. The CFI value was 0.913 for the one factor CFA of the PROMIS₄₇ data and 0.923 and 0.934 for the 2- and 3-group bifactor models, respectively. Differences for the RMSEA values were more substantial. For the one factor CFA, the RMSEA was 0.159 for the one factor CFA and 0.109 and 0.102, respectively for the 2- and 3-group bifactor models, respectively. The WRMR also improved substantially when subdimensions were modeled in the bifactor

models. WRMR values were 2.873 for the one factor CFA, and 1.596 and 1.519 for the 2- and 3-group bifactor models.

The only fit statistic whose values were high across all conditions was the NNFI/TLI. Our previous experience with NNFI/TLI values in our studies and those of others is that this statistic tends to be insensitive to model misfit making it a robustly “agreeable” statistic. This may explain the frequency with which NNFI/TLI values are reported in the literature, but we question whether it is a statistic of much value in this context.

In summary, this study demonstrated that two of the most distinctive characteristics of health outcome item banks, large numbers of items and skewed data distributions, can impact CFA fit results dramatically. We should clarify that these results do not lead us to critique CFA, per se, or its use in exploring the dimensionality of a dataset. Our critique is of the mechanical use of CFA fit criteria as a “permission slip” for modeling data using IRT. Nor do we claim to have “discovered” anything new about CFA or fit statistics. Rather we have demonstrated the responsiveness of commonly used fit statistic to characteristics of data typically observed in item banking contexts.

Our results lead us to concur with Reise and colleagues’ contention that nonhierarchical CFA, though it may be useful in exploring content heterogeneity, is unsatisfactory for judging whether data are sufficiently unidimensional for IRT analyses[38,39]. We further concur that bifactor analysis is a viable alternative that allows evaluations more conceptually proximal to the question at hand. After modeling the PROMIS Pain Impact data with bifactor models, we were able to estimate the saturation of the data by a dominant factor and to evaluate the relative impact of secondary dimensions. By comparing factor loadings from the first order and bifactor models, we estimated the degree to which unmodeled multidimensionality distorted results.

Though investigators and journal reviewers understandably appreciate the simplicity and putative clarity of comparing CFA results to traditional cut-offs and standards, we cannot recommend this approach. Our results demonstrated how sensitive CFA fit values are to factors other than dimensionality of the data. This leads us to favor a more investigative approach when evaluating whether data are sufficiently unidimensional for IRT modeling.

Our final conclusion concerns the appropriateness of the PROMIS₄₇ data for analysis using IRT models. As has already been noted, were we to appeal to CFA fit criteria in this decision, our conclusion would depend on the fit statistic selected. Based on published NNFI/TLI and SRMR standards, the data fit a unidimensional model. Based on published CFI, RMSEA, and WRMR standards, the data do not fit a unidimensional model, and with respect to the latter two statistics are far from fitting. However, the results of the bifactor analysis provided strong support for modeling the data using a unidimensional model. The secondary dimensions combined accounted for no more than 10% of the total variance and 12% of the common variance. The general factor clearly dominated, accounting for, in the two- and three-group models, respectively, 74% and 75% of the total variance and 88% of the common variance. We judge this result to be adequate evidence for the essential unidimensionality of these data and the appropriateness of applying a unidimensional IRT model.

LIMITATIONS AND DIRECTIONS FOR FUTURE RESEARCH

A limitation of this study is that only one skewed and one normal distribution were simulated and evaluated. A more comprehensive approach was outside the scope of this study—namely, to conduct analyses on multiple simulated datasets. This would allow more thorough examination of number of items and the level of skewness on CFA results. Though we think such a study would be informative, we do not think its chief purpose should be to construct arguments regarding which fit statistic or cutoff values are most appropriate; nor do we think

the purpose should be to develop a set of criteria for large item banks that have skewed distributions. In our opinion, such a goal is neither wise nor attainable. We recommend the use of bifactor analysis as an adequate and informative approach to evaluating the degree to which item banks conform to the IRT assumption of essential unidimensionality.

Another limitation is the nature of the real dataset. A large portion of persons reported having no pain impact. Future studies should evaluate the substantive meaning and methodological impact of including or not including responses of persons who report none of the trait being measured.

Acknowledgments

The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Evanston Northwestern Healthcare, PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project are Deborah Ader, Ph.D., Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, and Susana Serrate-Sztejn, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See the web site at www.nihpromis.org for additional information on the PROMIS cooperative group.

References

1. Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. Mahway, NJ: Lawrence Erlbaum Associates, Publishers; 2000.
2. Hambleton, R.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: Sage Publishing, Inc.; 1991.
3. Cook KF, O'Malley KJ, Roddey TS. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health Serv Res* 2005;40:1694–1711. [PubMed: 16179003]
4. Dodd BG, De Ayala RJ, Koch WR. Computerized adaptive testing with polytomous items. *Applied Psychological Measurement* 1995;19:5–22.
5. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Medical Care* 2000;38:II28. [PubMed: 10982088]
6. Wainer, H. Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.
7. Cook KF, Teal CR, Bjorner JB, Cella D, Chang CH, Crane PK, Gibbons LE, Hays RD, McHorney CA, O'Cepek-Welikson K, Raczek AE, Teresi JA, Reeve BB. IRT health outcomes data analysis project: an overview and summary. *Qual Life Res* 2007;16(Suppl 1):121–132. [PubMed: 17351824]
8. McDonald R. The dimensionality of test and items. *British Journal of Mathematical and Statistical Psychology* 1981;34:100–117.
9. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. *J Pers Assess* 2005;84:228–238. [PubMed: 15907159]
10. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. *Psychological Assessment* 2000;12:287–297. [PubMed: 11021152]
11. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical Care* 2007;45:S3–S11. [PubMed: 17443116]
12. Bjorner JB, Kosinski M, Ware JE Jr. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). *Qual Life Res* 2003;12:913–933. [PubMed: 14651412]

13. Lai JS, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Qual Life Res* 2006;15:1179–1190. [PubMed: 17001438]
14. Brown, TA. *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press; 2006.
15. Gorsuch, RL. *Factor analysis*. Vol. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.
16. Hu LT, Bentler PM, Kano Y. Can test statistics in covariance structure analysis be trusted? *Psychol Bull* 1992;112:351–362. [PubMed: 1454899]
17. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment* 1995;7:286.
18. Mannion AF, Elfering A, Staerkle R, Junge A, Grob D, Semmer NK, Jacobshagen N, Dvorak J, Boos N. Outcome assessment in low back pain: how low can you go? *Eur Spine J* 2005;14:1014–1026. [PubMed: 15937673]
19. McDonald RP, Mok MMC. Goodness of fit in item response models. *Multivariate Behavioral Research* 1995;30:23–40.
20. Bentler PM, Mooijart A. Choice of structural model via parsimony: a rationale based on precision. *Psychol Bull* 1989;106:315–317. [PubMed: 2678203]
21. Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. *Testing Structural Equation Models*. Newbury Park, CA: Sage Publications, Newbury Park, CA.; 1993. p. 136-172.
22. Hu, L.; Bentler, PM. Evaluating Model Fit. In: Hoyle, RH., editor. *Structural Equation Modeling: Concepts, Issues and Applications* (76–79). Thousand Oaks, CA: Sage Publications; 1995.
23. Hu LT, Bentler P. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling* 1999;6:1–55.
24. Muthen, BO.; Muthen, LK. *Mplus User's Guide*. Los Angeles, CA: Muthen & Muthen; 2001.
25. Yu, CY. Doctoral dissertation. Los Angeles: University of California; 2002. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.
26. Tucker L, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 1973;38:1–10.
27. Bentler P. Comparative fit indices in structural models. *Psycho Bull* 1990;107:238–246.
28. Joreskog, KG.; Sorbom, D. *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Lincolnwood, IL: Scientific Software International, Inc.; 1993.
29. Bentler, PM. *EQS structural equations program manual*. Encino, CA: Multivariate Software; 1995.
30. Browne MW. Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 1984;37:62–83. [PubMed: 6733054]
31. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS, 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* 2007;45:S22–31. [PubMed: 17443115]
32. McDonald, RP. *Test Theory: A Unified Treatment*. Mahway, NJ: Lawrence Earlbaum; 1999.
33. Kline, RB. *Principles and Practice of Structural Equation Modeling*. New York, NY: The Guilford Press; 1998.
34. West, SG.; Finch, JF.; Curran, PJ. *SEM with Nonnormal Variables*. Thousand Oaks, CA: Sage Publications; 1995.
35. Joreskog, KG. Lincolnwood, IL: Scientific Software International, Inc; 2005. *Structural equation modeling with ordinal variables using LISREL*.
36. Yuan KH, Bentler PM. Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association* 1997;92:767–774.
37. O'Connor BP. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers* 2000;32:396–402.
38. Reise S, Widaman K, Pugh R. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin* 1993;114:552. [PubMed: 8272470]

39. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual Life Res* 2007;16(Suppl 1):19–31. [PubMed: 17479357]
40. Yung YF, Thissen D, McLeod LD. On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika* 1999;64:113–128.
41. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. *Medical Care* 2007;45:S12–21. [PubMed: 17443114]
42. Reeve BB, Burke LB, Chiang YP, Clauser SB, Colpe LJ, Elias JW, Fleishman J, Hohmann AA, Johnson-Taylor WL, Lawrence W, Moy CS, Quatrano LA, Riley WT, Smothers BA, Werner EM. Enhancing measurement in health outcomes research supported by Agencies within the US Department of Health and Human Services. *Qual Life Res* 2007;16(Suppl 1):175–186. [PubMed: 17530449]
43. Thissen, D.; Chen, W-H.; Bock, RD. Multilog (version 7). Lincolnwood, IL: Scientific Software International; 2003.
44. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. No. 17. 1969
45. Bjorner, JB.; Smith, KJ.; Stone, C.; Sun, X. IRTFIT: A macro for item fit and local dependence tests under IRT models. Lincoln, RI: QualityMetric; 2007.
46. Orlando M, Thissen D. Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement* 2003;27:289–298.
47. Han KT. WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement* 2007;31:457–459.
48. Han, KT.; Hambleton, RK. User's Manual: WinGen (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education; 2007.
49. Choi SW. Firestar: Computerized adaptive testing (CAT) simulation program for polytomous IRT models *Applied Psychological Measurement*. 2008in press
50. Zinbarg RE, Barlow DH, Brown TA. Hierarchical structure and general factor saturation of the Anxiety Sensitivity Index: Evidence and implications. *Psychological Assessment* 1997;9:277–284.

Figure 1.
Distribution of Estimated Theta in Observed PROMIS Sample (N=745, Persons with No Missing Responses)

Table 1

Item Set/Subset Names and Characteristics

Item Set/Subset Names	Number of Item Sets	Number of Items per Item Set	Origin	Description
PROMIS ₄₇	1	47	Observed	Pain Impact items and item responses collected as part of PROMIS wave I data collection (N=745)
SIM.PROMIS ₄₇	1	47	Simulated	Same items as PROMIS ₄₇ . Item responses simulated to mirror the PROMIS ₄₇ distribution and fit GRM- calibrated parameters of PROMIS ₄₇ items (N=745)
SIM.NORM ₄₇	1	47	Simulated	Same items as PROMIS ₄₇ . Item responses simulated to fit a normal distribution and fit GRM-calibrated parameters of PROMIS ₄₇ items (N=745)
PROMIS-CAT _{a-j}	10	10	Observed	Items of the 10 most frequently presented item subsets in a simulation of CAT based on PROMIS ₄₇ (N=745)
SIM.PROMIS-CAT _{a-j}	10	10	Simulated	Same items as PROMIS-CAT _{a-j} but responses extracted from SIM.NORM ₄₇ (N=745)
SIM.NORM-CAT _{a-j}	10	10	Simulated	Same items as PROMIS-CAT _{a-j} but responses extracted from SIM.PROMIS ₄₇ (N=745)
PROMIS-RAND10 _{a-j}	10	10	Observed	10 subsets of 10 items each (and item responses) selected at random from among the PROMIS ₄₇ items (N=745)
PROMIS-SUB ₅₋₁₀	4	6-9	Observed	5- 6-, 7-, 8-, 9-, and 10-item subsets based on most frequently administered single items in CAT simulation. Responses based on PROMIS ₄₇ (N=745)

PROMIS = Patient-Reported Outcomes Measurement Information System

GRM = Graded response model

Table 2Demographics of the PROMIS₄₇ Sample

	N	% of those reporting
Gender:		
Male	340	45.6%
Female	405	54.4%
TOTAL	745	100.0%
Race/Ethnicity: ~		
Asian	7	0.9%
Black	66	8.9%
Hispanic	83	11.1%
Native American	26	3.5%
Pacific Islander	1	0.1%
White/not Hispanic	546	73.3%
Other	54	7.2%
TOTAL	783	105.1% ^a
Education:		
Less than high school degree	2	0.4%
High school or GED	15	3.0%
Some college/technical degree	111	22.1%
College degree	216	42.9%
Advanced degree	159	31.6%
TOT	503	100.0%
Household income:		
Less than \$20,000	69	9.5%
\$20,000 to \$49,999	254	35.0%
\$50,000 to \$99,000	282	38.8%
\$100,000 or more	121	16.7%
TOT	726	100.0%

^aDemographic percentages sum to over 100% because participants could endorse more than one race/ethnicity.

Table 3
Items Occurring in Most Frequently Administered 10-Item Computer Adaptive Tests (CATs)

Unique CATs	ITEMS	Frequency	Percent of Total (n=745)
CAT _a	PI12, PI13, PI16, PI20, PI26, PI34, PI35, PI39, PI53, PI56	227	30.5%
CAT _b	PI3, PI5, PI6, PI8, PI14, PI20, PI24, PI30, PI46, PI56	122	16.4%
CAT _c	PI3, PI5, PI6, PI8, PI14, PI24, PI30, PI43, PI46, PI56	91	12.2%
CAT _d	PI3, PI5, PI6, PI8, PI14, PI24, PI30, PI41, PI46, PI56	44	5.9%
CAT _e	PI3, PI5, PI13, PI14, PI16, PI20, PI28, PI34, PI46, PI56	38	5.1%
CAT _f	PI3, PI5, PI13, PI14, PI16, PI20, PI28, PI43, PI46, PI56	24	3.2%
CAT _g	PI3, PI5, PI13, PI14, PI16, PI20, PI28, PI34, PI53, PI56	20	2.7%
CAT _h	PI3, PI5, PI6, PI14, PI20, PI24, PI28, PI30, PI46, PI56	18	2.4%
CAT _i	PI3, PI13, PI16, PI20, PI26, PI28, PI34, PI35, PI53, PI56	17	2.3%
CAT _j	PI3, PI13, PI14, PI16, PI20, PI26, PI28, PI34, PI53, PI56	17	2.3%
TOTAL:		618	83.0%

Note: PI56 was the starting item for the CAT simulations

Table 4

Confirmatory Factor Analyses Fit Statistics, All Datasets

CAT	CFI	NNFI (TLI)	RMSEA	SRMR	WRMR
Standard Cutoff Values:	> .95	> .95	< .06	< .08	< 1.0
PROMIS ₄₇	0.913	0.989 [*]	0.159	0.062 [*]	2.873
SIM.PROMIS ₄₇	0.990 [*]	1.000 [*]	0.029 [*]	0.013 [*]	0.697 [*]
SIM.NORM ₄₇	0.993 [*]	1.000 [*]	0.025 [*]	0.012 [*]	0.662 [*]
PROMIS-CAT _{a-j} Mean	0.980 [*]	0.995 [*]	0.157	0.034 [*]	2.373
Minimum	0.949	0.986 [*]	0.114	0.020 [*]	1.671
Maximum	0.991 [*]	0.998 [*]	0.199	0.051 [*]	3.055
SIM.PROMIS-CAT _{a-j} Mean	0.995 [*]	0.999 [*]	0.057 [*]	0.017 [*]	0.955 [*]
Minimum	0.975 [*]	0.994 [*]	0.032 [*]	0.012 [*]	0.660 [*]
Maximum	0.999 [*]	1.000 [*]	0.136	0.034 [*]	1.971
SIM.NORM-CAT _{a-j} Mean	0.996 [*]	0.999 [*]	0.027 [*]	0.015 [*]	0.685 [*]
Minimum	0.994 [*]	0.998 [*]	0.020 [*]	0.014 [*]	0.629 [*]
Maximum	0.998 [*]	0.999 [*]	0.034 [*]	0.018 [*]	0.754 [*]
PROMIS.RAND _{a-j} Mean	0.975 [*]	0.990 [*]	0.135	0.048 [*]	2.208
Minimum	0.945	0.973 [*]	0.081	0.021 [*]	1.259
Maximum	0.992 [*]	0.998 [*]	0.226	0.082	3.849
PROMIS-SUB ₁₀	0.965 [*]	0.989 [*]	0.131	0.052 [*]	2.111
PROMIS-SUB ₉	0.971 [*]	0.990 [*]	0.128	0.050 [*]	1.999
PROMIS-SUB ₈	0.986 [*]	0.995 [*]	0.101	0.034 [*]	1.558
PROMIS-SUB ₇	0.988 [*]	0.994 [*]	0.088	0.032 [*]	1.346
PROMIS-SUB ₆	0.994 [*]	0.996 [*]	0.085	0.024 [*]	1.225
PROMIS-SUB ₅	0.994 [*]	0.994 [*]	0.108	0.025 [*]	1.357

* Cutoff criteria met

Abbreviations:

CFI – Comparative Fit Index

NNFI (TLI) – Non-Normed Fit Index (Tucker Lewis Index)

RMSEA – Root Mean Square Error of Approximation

SRMR – Standardized Root Mean Square Residual
WRMR – Weighted Root Mean Square Residual