

PROMIS SCC Data Analysis Plan (August 17, 2005)

Input into this document has been provided by

Rita Bode, Jakob Bjorner, David Cella, Karon Cook, Paul Crane, Steve Durako, Richard Gershon, Karen Gold, Elizabeth Hahn, Ron Hambleton, Ron Hays, Jin-Shei Lai, Paul Pilkonis, Bryce Reeve, Steve Reise, Dennis Revicki, Jeanne Teresi, David Thissen, Kevin Weinfurt, David Weiss

INTRODUCTION

This document describes a proposed approach for analyzing patient responses to a pool of items hypothesized to measure health domains applicable to outcomes research, clinical trials, and/or clinical practice. Also described is the application of item response theory (IRT) to the calibration of item banks. These banks will be used for developing short form instruments and enabling CAT-based assessment. Based on decisions by the Steering Committee, the first five domains targeted by the PROMIS network are physical function, fatigue, pain, emotional distress, social/role participation. A sixth domain, general health, was recently added. Those who conduct analyses for PROMIS should be sure to read the domain hierarchy protocol document authored by Fries et al.

The data analysis plan applies to two distinct sets of analyses. Initially secondary analyses of existing datasets will be performed. Later in the project, psychometric analyses of data collected under the PROMIS network will support the development and evaluation of item banks and instruments for measuring identified PROMIS domains.

SECONDARY DATA ANALYSES

Analyses will be performed on multiple datasets including the Cardiovascular Health Study (CHS), Chronic Hepatitis C (CHC) Study, Digitalis Investigation Group (DIG) QOL Sub-study, Medical Outcomes Study (MOS), Q-Score, and the WHO-QOL datasets. Other datasets considered included: American Academy of Orthopedic Surgery (AAOS) dataset, Antiarrhythmics versus Implantable Defibrillators (AVID), Atherosclerosis Risk in Communities (ARIC) Study, BIOQOL, CHD-Offspring, Diabetes Control and Complications Trial, FAMS-Bayer ESIMS Trial, GERD Study, HIV Cost and Services Utilization Study (HCSUS), Item Banking Social Bank, MEDTEP HIV Study, National Epidemiologic Survey on Alcohol and Related Conditions, National Spinal Cord Injury Database, OABq Study, PAGI-SYM, P-QOL, Pain Audit Collection System, RAND Health Insurance Experiment (HIE), Schizophrenia Study, UK-FIMFAM, Unified Medical Group Association, Veterans Health Outcomes Study, and Women's Health Questionnaire.

The secondary data analyses are opportunities to explore a number of psychometric questions, the answers to which will inform the eventual development of the PROMIS measures. These questions including what is the: (a) dimensional structure of the domains and subdomains identified by domain chairs?, (b) functional number of response categories persons discriminate in self-reporting various outcomes?, and (c) range of targeted outcomes effectively measured by items within secondary datasets. A fourth question is of particular importance: "Since different constructs have different levels of conceptual breadth, at what point along the 'construct hierarchy' will it be appropriate to scale items using item response theory since IRT assumes unidimensionality (Reise, et al, 2000). For example, PROMIS' domain map conceptualizes

Physical Health as being comprised of Function, Symptoms, and Special Senses. In turn, Function has been defined as Mobility, Dexterity, Central, and Activities of Daily Living. Additional sub-categorizations may be made within each of these subdomains. At what level will the domains be unidimensional enough for IRT modeling? At the level of Physical Health? At the level of Function? At the level of Mobility? These are questions that must be addressed conceptually and empirically. It is anticipated that the analyses of the secondary data sets will inform the decisions.

DEVELOPMENT OF PROMIS INSTRUMENTS

Ultimately, this analysis plan will be implemented on the datasets collected within the PROMIS effort and used to calibrate the items for the PROMIS core and supplemental banks. This plan also describes data collection methods to allow co-calibration of items to a common metric.

ANALYTIC METHODS

The scope of work of the PROMIS initiative is vast and methodologically challenging. Below we classify the analyses that will be undertaken and describe methodologies for accomplishing the analytic tasks. In many cases there are multiple alternatives for addressing the analytic questions and no scientific consensus regarding the best approach or the criteria by which to evaluate the results. Below we describe important issues, review available analytic options and, when evidence supports it, suggest preferred methods and criteria.

DESCRIBING THE DATA

A range of descriptive statistics will be included to examine the performance of the individual items as well as how they function within the scale. Item statistics will include means, variances, and frequencies. Item performance within the scale will be examined by inter-item correlations, item-total score correlations, and drop in coefficient alpha when the item is removed. The scale will be examined in terms of its mean, range, variance, and internal consistency.

We will examine patterns of missing data to determine reasons for missing data. For example, if missing data were more prevalent later in the sequence of administered items, this would suggest the cause was response burden or lack of time for completing the survey. The content of items that draw substantial missing responses will be examined to evaluate whether missing responses may due to sensitive item content. For calibration purposes, we propose to treat missing data as “not presented.”

EVALUATING ASSUMPTIONS OF THE IRT MODEL

Item response theory relies on strong assumptions (unidimensionality, local independence, monotonicity). Below are described methods for testing these assumptions. The order in which the subtopics are listed represents one possible sequence for conducting the analyses, but we recognize that other sequences are possible (e.g., some analysts may prefer to perform an evaluation of monotonicity earlier in the sequence).

Unidimensionality

When applying IRT models to the data, we want to assess whether scales are unidimensional “enough” to allow for the unbiased scaling of individuals on a common latent trait. The challenge of dimensionality assessment is to develop approaches that assess whether a scale has

a strong enough general factor to be considered “essentially” unidimensional. No complex item set will ever perfectly meet strictly defined unidimensionality assumptions (see McDonald, 1981). What we are really interested in assessing is whether the trait level estimates are predominantly influenced by a general factor.

A number of researchers have recommended methods and considerations for evaluating essential unidimensionality (McDonald, 1981; Roussos & Stout, 1996; Stout, 1987). Our analyses will be informed by their findings and by the work of Zinbarg, Ravelle, Yovel, and Li (in press) who recently described alternative classical indices for evaluating the degree to which test scores are influenced by a general construct.

Classical Test Theory Methods

Several classical test theory statistics will be estimated to assess dimensionality. These include inter-item correlations, item-scale correlations, and internal consistency reliability. (See Zinbarg et al., in press, for which index is best for which question). We recognize that high internal consistency can be achieved with multi-dimensional data and will not rely heavily on the results of these analyses.

Factor Analytic Methods

We will conduct confirmatory factor analysis (CFA) to evaluate the extent to which the item pool measures a dominant trait that is consistent with the content experts’ definition of the domain to be measured by the PROMIS. There are numerous options in conducting a CFA. Those conducted as part of the PROMIS data analysis will use appropriate software (e.g., MPLUS, LISREL) for categorical (ordinal) data that uses tetrachoric or polychoric correlations as the basis for item-level factor analysis.

We anticipate that, based on statistical criteria (e.g., chi-square statistics), a unidimensional model will not fit content experts’ domain definitions as well as a multi-factor model, especially when a large sample size is used. Hence, we will focus on practical fit indices such as the comparative fit index, RMSEA, factor loadings, and average absolute residual correlations. We also will conduct bifactor analysis to address these issues (see Reise, Morizot, & Hays, submitted).

With respect to fit in a CFA framework, however, a word of caution is warranted. Floyd and Widaman (1995) have called it “unreasonable” to expect satisfactory solutions from a CFA of questionnaires that contain more than a few items since pairs of items will share content and thus variance that is not a part of the factors. Their comments are particularly pertinent to factor analyses that fail to account for the categorical nature of questionnaire data (e.g., use of polychoric or tetrachoric correlation matrices). Floyd and Widaman recommend avoiding this problem by creating “testlets” (also known as “facets” or “parcels”) by summing across several items.

We can expect in our analyses of the PROMIS domains to find that a multifactor model fits better. The key issue is the degree to which multidimensionality threatens the validity of inferences based on the IRT-calibrated scores. There is no scientific consensus regarding this issue and little published literature to suggest a standard of practice for applying the results of a

CFA in assessing unidimensionality for an IRT analysis. From the health outcomes literature, two studies by Bjorner and colleagues serve as examples:

1. Bjorner, Kosinski and Ware evaluated the dimensionality of an item pool assessing burden of headache (2003a). They compared a 1-, 2- and 7-factor model and found that none of the models fit well. However, other results suggested the item pool was unidimensional *enough* for IRT modeling. The authors judged it appropriate to proceed with an IRT calibration based on evaluation of a scree plot and the finding that the first factor accounted for most of the variance (60%), the first and second factor were highly correlated (0.92), and loadings on the first factor were both high (0.62 to 0.89) and similar to the loadings for a 2-factor solution (0.63 to 0.90).
2. In a similar study, a 3-factor solution best fit a headache item pool, but the authors judged an IRT calibration appropriate based on a scree plot, the percentage of variance accounted for by the first factor (54%), high correlations among factors (0.74, 0.79, 0.83), and high loadings on the first factor (0.50-0.89 and one at 0.31 that was dropped from the item pool) (Bjorner, et. al, 2003b).

An example from educational and psychological assessment literature is a study by Schatschneider, et al. (1999). These investigators examined whether a phonological awareness measure had sufficient unidimensionality to warrant an IRT calibration. They concluded that it did, based on a scree plot, amount of variance accounted for by the first factor (65%), and the invariance of item difficulty parameters (item difficulties based on calibration from separate random halves of items correlated at 0.95).

The PROMIS plan for conducting factor analysis is as follows:

- 1) Conduct a confirmatory factor analysis (CFA) to examine how well the items fit the one factor model as described by the domain hierarchy group. The analysis will use polychoric correlations to account for the ordered-response data. An appropriate estimation procedure will be selected (e.g. WLSMV, WLSM, or DWLS).
- 2) Examine model fit using indices like the RMSEA, SRMR, TLI, and CFI.
- 3) Examine modification indices residual correlations to identify locally dependent pairs (or sets) of items.
- 4) Conduct an exploratory factor analysis if large misfit is discovered. Examine magnitude of eigenvalues for second and third factors, correlations among factors, and factor loadings to evaluate whether a unidimensional solution should be considered.

Other Potential Analyses

Below are other analyses that may be conducted. It is critical that these analyses and all others be informed by a thorough understanding of the content domains being assessed by the item pools.

- 1) Parallel Analysis. The size of eigenvalues expected by chance alone can be compared with observed eigenvalues to help determine the maximum number of underlying dimensions.
- 2) Rasch residual factor analysis. Principal components analysis of the residuals from a Rasch model can be performed to identify secondary structures or subdomains within the data (Linacre, 1998). McDonald (1981) proposed a similar strategy of analyzing residuals after extracting the first factor.

- 3) For domains in which it is assumed that a number of sub-domains load on a general factor, a bi-factor analysis may be explored to evaluate the relative weighting of items on the general and specific factors. A bi-factor model allows each item to load on a general factor and one “group” factor. The (squared) loadings indicate the proportion of variance of the item that is accounted for by the two factors. The variance components are independent and measure two different variables with independent variance components. Thus, the group factors are orthogonal to each other and the general factor and are defined by item content facets. Fitting a bi-factor model to the data allows us to evaluate the degree to which using an item set to scale individuals on a common factor is distorted by the presence of small secondary group factors (see Appendix A for more details).
- 4) For each scale, compute a unit-weighted composite and use item-level factor analytic results to compute how much of the variance in this unit-weighted composite is due to the general factor and how much to the group factors (nuisance content dimensions). (Raykov, 2004).
- 5) Apply non-parametric methods such as DIMTEST and poly-DIMTEST.

Available software for use in assessing dimensionality assessment:

- 1) SAS, SPSS or STATA for item-scale correlations and internal consistency reliability estimates.
- 2) TESTFACT for item-level factor analyses of dichotomous items.
- 3) MPLus and EQS for EFA and CFA analyses of ordinal data.
- 4) PRELIS/LISREL or Microfact (unless there is too much missing data)

Local Independence

We will assess whether there are locally dependent item sets that influence IRT model calibration of the data (Steinberg & Thissen, 1996; Wainer & Thissen, 1996; Yen, 1993).

Analyses of local independence will include:

- 1) Identifying locally dependent (LD) item sets by examining the residual correlation matrix produced by the factor analyses. Histograms of residuals will be examined and high residual correlations will be noted and considered possibly LD. In addition, we will estimate Yen’s Q_3 statistic. Also, we will look at modification indices in MPLUS to estimate the improvement in model fit that would occur if residual correlations were estimated. (MPLUS has some limitations with large sample sizes and many items that may lead to difficulties in computing modification indices.)
- 2) Looking for serious violations (e.g., two or more items that are essentially the same content), we will consider marking the items as “enemies,” preventing them from appearing on the same assessment in the future.
- 3) Exploring whether the item parameters are consistent when analyzed with different subsets of items, and equivalently, looking for discrimination parameters that are unusually high.

Other potential analyses include:

- 1) Inspecting the item content of group factors that emerge in the exploratory bi-factor model analyses or item sets that load on factors other than the dominant factor in an EFA.
- 2) Determining overall impact on item banking, e.g., in the development of CAT-based and short form instruments. Potential LD item sets will be flagged. Possible causes of LD

include similar item wording and items sharing similar content that is not part of the construct being assessed.

Available software for evaluating item dependency for dichotomous response items includes IRTNEW (Chen, 1998).

Monotonicity

The probability of endorsing or selecting an item response indicative of better health should increase as underlying level of health increases (monotonically). Monotonicity means that, apart from sampling fluctuation, the proportion of people “passing each step” on the response scale is larger for those with a higher scale score (correcting for item overlap with the scale score). If the predicted order is reversed, this is a “violation” of monotonicity. Evaluating monotonicity is important, particularly since some programs (e.g., Parscale) will not run when this assumption is violated.

One approach for studying monotonicity is to examine graphs of item mean scores conditional on rest-scores (total raw scale score minus the item score). Other approaches are to fit a non-parametric IRT model to the data or apply a “lowess” function commonly available on SAS and SPSS. This method provides initial IRT trace line estimates to go in the next phase of IRT-calibration.

Examples of the evaluation of monotonicity are noted in Appendix B.

Available software for monotonicity assessment include:

- 1) MSP (Molenaar & Sijtsma, 2000) calculates rest-score functions and provides statistical tests of monotonicity violations. MSP handles both dichotomous and polytomous item responses and also provides a Mokken scaling analyses.
- 2) Ramsay’s TESTGRAF software fits non-parameter IRT models to the data.

We will also conduct analyses of the robustness of IRT estimates to violations of model assumptions including simulation studies (e.g., using software developed by David Weiss).

ESTIMATE IRT ITEM PARAMETERS AND MODEL FIT

We will fit Samejima’s Graded Response Model (GRM) to the response data and examine the variation in difficulty and discrimination among the item parameter estimates. Model fit to the response data will be examined. The GRM will be used to calibrate the items of the item bank. Analyses of item parameters and model fit will include:

- 1) IRT model estimation will use the two-parameter polytomous GRM. The GRM offers an attractive and flexible framework for modeling the participant responses. The model is relatively easy to understand and illustrate to PROMIS investigators and “consumers” and retains its functional form when response categories are merged. The GRM also is easy to implement in a CAT-based application. See Appendix C for the construction of the GRM.
- 2) Assess IRT model fit. (Note: if model assumptions are supported by the data, then strict adherence to model fit statistics is not vital given the limits of acceptable fit indices). We will compare observed and expected response frequencies by item and response category. In addition, we will compare fit for different models based on analyses of the size of the

differences (residuals). We will examine common fit statistics such as Q_1 (Yen, 1981), Bock's chi-square, and others (van der Linden & Hambleton, 1997). We also will consider generalizations of Orlando and Thissen's work $S-\lambda^2$ (2000, 2003) to polytomous data. It is important to note that chi-square based fit statistics are known to have distributional problems and are highly affected by sample size. The ultimate question is to what degree misfit affects model performance in terms of the valid scaling of individual differences. (Simulation can help answer this question.)

- 3) The psychometric properties of the items will be examined by review of their item parameter estimates, item characteristic curves (ICCs, also called trace lines), and item information curves. ICCs model, in probabilistic terms, the relationship between a person's response to a question and his or her level on the construct (θ , theta) being measured by the scale. The steepness of the curves are defined by the discrimination power (a parameter; also referred to as the item's slope parameter). Intersections between category curves are defined by item difficulty (b parameter; also called location or threshold parameter). ICCs allow us to examine how well each response category functions for measuring different levels of the measured construct. Information curves indicate the range over θ where an item is best at discriminating among individuals. Higher information denotes more precision (or reliability) for measuring a person's trait level. The height of the curves (denoting more information) are a function of the discrimination power (a parameter) of the item. The location of the information curves is determined by the threshold (b) parameter(s) of the item. Information curves allow us to identify which item(s) is (are) most useful for measuring different levels of the measured construct. This is critical for the item selection process in CAT administrations and in the development of short-form surveys.
- 4) Items that do not fit the model or have poor discrimination should be reviewed by content experts before the item bank is established. Misfitting items may be retained or revised when they are identified as clinically relevant and flagged with an identifier. Low discriminating items in the tails of the theta distribution also may be retained or revised to add information for extreme scores.
- 5) The theta-metric will be standardized by setting the population mean to zero and the variance to one (reference group determined by PROMIS steering committee). This will allow interpretation of difficulty (threshold) parameter(s) relative to the population mean and the discrimination parameters relative to the population standard deviation. For example, a difficulty parameter estimate of $b = 1.5$ suggests, in the dichotomous response case, that a person who is 1.5 standard deviations above the mean will have a 50% probability of endorsing the item. This metric will facilitate the conversion of the IRT z-score metric to the T-score distribution adopted by the PROMIS Steering committee. For the purposes of computing the proportion of the norming/calibration sample that score below each theta level and finding the z-score corresponding to that percentage from a normal distribution, we will treat the maximum likelihood estimates as raw scores. These pseudo-normalized z-scores will be converted to T-scores.

Other possible analyses include:

- 1) Alternate IRT models will be examined in a methodological study to explore the effect of modeling data using various parametric IRT models as well as multi-dimensional models.

- 2) Multidimensional CAT, in which the response in one domain informs the score estimation in other domains, will also be considered. (However, it should be noted that multidimensional IRT has all the rotation problems and complexity that factor analysis does, it greatly complicates DIF analyses, and the meaning of scores is often unclear when subscales are highly correlated. In addition, essentially unidimensional constructs are often desirable from a theoretical perspective.)

Available Software:

- 1) Both PARSCALE and MULTILOG can estimate parameters for one and two-parameter polytomous response data.
- 2) We will work with Bjorner to implement a new model fit software program he developed which provides S- λ^2 , S-G2, and plots the expected and observed response proportions for various levels of the simple sum score.
- 3) Freeware (http://work.psych.uiuc.edu/irt/dif_dtf.asp).

Differential Item Functioning Studies

An item displays differential item functioning (DIF) if the probabilities of responding in different categories differ by population for the same underlying level of the attribute (Teresi, 2001). Items can be evaluated for DIF by contrasting the IRT difficulty or location (b_i) and slope (a_i) parameters between two groups. Determination of DIF is optimized when the samples are as representative as possible of the populations from which they are drawn.

There are numerous approaches to assessing DIF. Below we list some (not all) methods for DIF testing (see Millsap & Everson, review of invariance). It is prudent to evaluate DIF using multiple methods, and identify those items that are flagged by multiple methods. Inclusion or exclusion of DIF items, or controlling the bias using separate IRT calibrations, will be determined by group consensus within the PROMIS domain-working group with feedback from clinical experts.

Likely DIF methods to be implemented include:

- 1) Raju's signed and unsigned area tests and multiple-group analyses to estimate item parameters simultaneously for each subgroup, as well as the mean and standard deviation on the latent trait. By imposing a constraint that the average item locations (or discriminations) are equal across groups, we will examine the interaction between group membership and the item parameter. Measurement equivalence within Raju's Differential Functioning of Items and Tests (DFIT) framework means that the true score differences are equal to zero at both item and scale levels. This framework includes both noncompensatory DIF (NCDIF) and compensatory DIF (CDIF) indices. NCDIF reflects the average squared difference between the item-level true scores for the focal and reference groups. CDIF is an item level index representing an item's net contribution to the differential test ("scale") functioning (DTF). A chi-square test will be used to test whether DTF is significantly different from 0. When significant DTF is found, we will set the items aside and estimate DTF again, iterating until DTF is no longer significant. After completing the process items with $NCDIF \neq 0$ will be reported as potentially biased items. Determination of whether these indices are "significant" is conditional on a significant chi-square and a difference that exceeds a specified critical value.

- 2) Thissen's IRT-based Likelihood-Ratio (IRT-LR) test to identify both uniform and non-uniform cases of DIF. Uniform DIF suggests DIF in the threshold (or difficulty) parameter of the model, which indicates that for all levels of the underlying construct, the test and reference groups have different response probabilities for the tested item. Non-uniform DIF appears in the discrimination parameter and suggests interaction between the underlying measured variable and group membership (Teresi, Kleinman, & Ocepek-Welikson, 2000). In other words, the degree to which an item relates to the underlying construct depends on the group being measured. The IRT-LR DIF procedure compares hierarchically nested IRT models with one allowing the test item parameters to be freely estimated between groups and the other constraining the parameters to be equal between the two groups. Thissen (IRTLRDIF v. 2.0b June 30, 2001) notes that "because IRT-LR procedures make no use of the simple summed score on any set of test items, but instead rely on IRT characterization of the posterior distribution of ability as does a computerized adaptive test (CAT), IRT-LR procedures may make an easy transition to the detection of DIF for data collected in a CAT environment." The IRT-LR DIF procedure can be easily implemented using the free IRTLRDIF software program made available on the following web-site: <http://www.unc.edu/~dthissen/dl.html>.

A sample of at least 200 per group is generally recommended for IRT-based DIF detection. One approach is to calibrate with the larger group and apply the results to the smaller sample to evaluate how well the observed responses correspond to the modeled responses (residual analysis).

Other possible DIF detection methods include:

- 1) Obtaining separate item calibrations for each subgroup. After transforming item parameters to the same mathematical metric, if the difference between the separate item parameter values is greater than twice the combined group standard errors, the item may have an unstable location on the underlying continuum/construct, therefore, it will be flagged as exhibiting DIF on the characteristic being examined.
- 2) Ordinal logistic regression (OLR) approach. In this approach a series of logistic models predicting the probability of item response are run and compared. The independent variables in Model 1 are the trait estimate (e.g., raw scale score), group, and the interaction between group and trait. Model 2 includes main effects of trait and group, and Model 3 includes only the trait estimate. Non-uniform DIF is detected if there is a statistically significant difference in the likelihood for Model 1 and Model 2. Uniform DIF is evident if there is a significant difference in the likelihoods for Models 2 and 3. Crane et al. (2004) suggested that, in addition to statistical significance, the relative change in beta coefficients between Model 2 and 3 should be considered. Based on simulation by Maldonado and Greenland (1993), a 10% change in beta has been recommended as a criterion for uniform DIF.
- 3) Multi-Group MIMIC modeling. Multiple-indicator, multiple cause structural equation models (MIMIC) can be used to examine if differences in observed variables exist beyond differences in the latent variable of interest (Fleishman & Lawrence, 2003).

Available software for these analyses includes DFIT (Raju et al., 2005) LINKDIF, IRT-LRDIF (Thissen, 2001), and DIFDETECT (Crane et al., 2003).

LINKING OF DATASETS

Analogous to structural equation modeling, the estimation of IRT model parameters requires an identification constraint. That is, in order to identify the scale for the item parameters, the scale for person parameters must be fixed in some manner – typically by specifying that the mean in the population is zero and the standard deviation is one.

As a consequence of the identification problem, when IRT item parameters are estimated using different samples of individuals, the item parameters will not be on a common or comparable metric unless the samples are random samples from the same population. We can seldom assume random samples, therefore some type of scale “linking” procedure will be required to make item parameter estimates comparable.

In educational assessment, the issue of linking to a common metric is of great concern because different cohorts are administered different versions of an instrument, and there is a compelling need for scores to be on the same scale. Also, new items frequently are added to the item pool and need to be calibrated to the same mathematical metric as the old ones.

In PROMIS, the need for linking occurs in several contexts. Consider these examples. In context A, a common measure is administered to two samples of individuals, and item parameters are estimated separately in each. Because the samples may be drawn from populations that differ in mean and standard deviation on the latent trait, a linking procedure is necessary before the degree to which item parameters are similar across groups can be investigated. This situation is completely analogous to the assessment of DIF. (Alternatively, multiple group simultaneous estimation methods may be implemented, such as those available in PARSCALE, BILOG-MG, or MULTILOG.)

In context B, a common group of individuals respond to a measure consisting of several pre-calibrated items plus some new that have yet to be calibrated (i.e., experimental items). The objective of the analyses is to estimate the item parameters for the new items, such that their item parameters are on the established scale. There are two approaches to accomplishing this. First, all item parameter (for both old and new items) can be estimated in the sample. Then, standard procedures for finding transformation constants (e.g., the test characteristic curve method, mean and sigma linking, etc.) can be used on the anchor items. Assuming no DIF, these transformation constants can then be used to transform the item parameter estimates for the new items onto the old scale. A second approach is to fix the item parameters for the anchor items to their pre-existing values. Under this approach, it is hoped that the fixing of values for the common items correctly sets the metric for the estimation of the parameters for the new items.

In context C, two different samples of individuals have responded to a set of common items and some unique items. The common items can serve as an anchor in a simultaneous calibration, which, in theory, can set the metric for the unique items administered to each group. This type of situation is common in aptitude assessment where achievement tests are administered to students at multiple grade levels. (Under some circumstances, data collected as described above in context B may, for statistical reasons, be treated as though they came from context C, considering the “old” items to be the common items and the “new” items to be the unique items. This is done, for example, between years in the ongoing National Assessment of Educational Progress.)

PROMIS sampling plan for item co-calibration [This plan will be adjusted as we learn about the number of legacy items]:

- 1) For each unidimensional item pool, the total number of available items will be randomly divided into k blocks of x items. Item blocks will be reviewed to make sure there is representation in terms of item difficulty across the trait continuum. (Note: The size x of the linking blocks will change as we determine both the limits of patient burden, the expected sample size of respondents participating in the first wave of PROMIS testing, and budgetary and practical constraints.)
- 2) Questionnaires will be constructed to contain two item blocks from each measured domain/item pool (thus patients will respond to two blocks per item pool). For example, Questionnaire 1 will contain item blocks 1 and 2, Questionnaire 2 will contain item blocks 2 and 3, Questionnaire 3 will contain item blocks 3 and 4, Questionnaire T-1 will contain item blocks $k-1$ and k , and Questionnaire T will contain item blocks k and 1. The goal is to have 1000 respondents (minimum 500) for each linking block. Each block of items should also be administered to patients of each disease/disorder classification to ensure generalizability.
- 3) Randomization of people (either by overall study population or by target disease population) to each questionnaire form will ensure that an adequate number of people will respond to all k blocks of items to permit simultaneous calibration of all items.

 Example: Item Pool A has 120 items. The 120 items are divided into ($k =$) 6 blocks of ($x =$) 20 items. Each person will respond to 40 items representing 2 blocks of items. Thus the sampling plan will look something like this (each group of respondents consists of patients from all major disease/disorder populations):

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Group 1	*	*				
Group 2		*	*			
Group 3			*	*		
Group 4				*	*	
Group 5					*	*
Group 6	*					*

If each group above has 500 people in it, it provides 1000 responses per item.

-
- 4) Evaluation of psychometric properties of response data, assessment of model assumptions, and IRT co-calibration continues as described earlier in this document. The sample sizes are large enough for parameter estimation as well as estimation of mean differences and standard deviation ratios among study populations.

CAT Simulations

To determine if any measurement gaps or deficiencies exist along the theta continuum for CAT-based assessment based on some pre-determined settings of minimum standard error or maximum number of item presentations, we will perform CAT simulations.

Item Drift

We will examine item drift by comparing item parameter estimate for items administered at multiple points in time (DeMars, 2004). We will compare estimates of item difficulty and discrimination for the same item at multiple time points to assess if the properties of items appear to have changed over time.

References Cited

Bjorner, J.B., Kosinski, M., Ware, J.E., Jr. (2003a) Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact (HIT). Quality of Life Research, 12:913-935.

Bjorner, J.B., Kosinski, M., Ware, J.E., Jr. (2003b) The feasibility of applying item response theory to measures of migraine impact: a re-analysis of three clinical studies. Quality of Life Research, 12:887-902.

Chen, W. (1998). IRTNEW: A computer program for the detection of local item dependence. Chapel Hill: N.C.: L. L. Thurston Laboratory, University of North Carolina at Chapel Hill.

Crane, P. K., Jolley, L., & van Belle, G. (2003). DIFDETECT. University of Washington, Seattle, WA.

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. Statistics in Medicine, 23, 241-256.

DeMars, C. (2000). DRAWICC: Modules to graph item response functions and item information functions with SAS GPLOT. Applied Psychological Measurement, 24, 224.

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: True differences or differential item functioning. Medical Care, 41, III-75--III-86.

Floyd, F.J., Widaman, K.F. (1995) Factor analysis in the development and refinement of clinical assessment instruments. Psychological Assessment, 7:286-299.

Fries, J. et al. (in preparation). Conceptual framework for PROMIS dimensions and domains. March 2, 2005.

Kolen, M. J., & Brennan, R. L. (2004). Test equating: Methods and practices, 2nd edition. New York: Springer.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? Journal of Outcomes Measurement, 2, 266-283.

McDonald, R.P. (1981). The dimensionality of test and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17, 297-334.

Molenaar, I.W., & Sijtsma, K. (2000). User's manual MSP5 for windows. Groningen: iecProGAMMA.

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

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, 289-98.

Raju, N.S., van der Linden, W. J., & Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement, 19, 353-368.

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modelling approach. British Journal of Mathematical and Statistical Psychology, 57, 21-27.

Reise, S. P., Morizot, J., & Hays, R. D. (submitted). The role of the bifactor model in resolving dimensionality issues in health outcomes measures.

Reise, S.P., Waller, N.G., Comrey, A.L. (2000). Factor analysis and scale revision. Psychological Assessment, 12, 287-297.

Roussos, L, & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, 20, 355-371.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our view of the state of the art. Psychological Methods, 7, 147-177.

Steinberg, L, & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. Psychological Methods, 1, 81-97.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Teresi, J. A. (2001). Statistical methods of examination of differential item functioning with applications to cross-cultural measurement of functional, physical and mental health. J Mental Health and Aging, 7, 31-40.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning.

van der Linden, W. J., & Hambleton, R. K. (1997). Handbook of modern item response theory. New York: Springer.

Wainer, H. & Thissen, D., (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practice, 15, 22-29.

Ware, J.E., Jr, Kosinski, M., Bjorner, J.B., Bayliss, M.S., Batenhorst, A., Dahlof, C.G., Tepper, S., Dowson, A. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Qual Life Res.* 2003 Dec;12(8):935-52.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (in press). Cronbach's α , Revelle's β , and McDonald's $\hat{\omega}_h$: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*.

Appendix A: Using Bifactor Model to Assess Dimensionality

Figure 1 shows four possible latent variables models. Model A is the most restricted and Model D is the least. Model A is the standard unidimensional model where we hypothesize that the covariance among item responses is explained by one common factor. This is the model we hope provides an acceptable fit to our data if our objective is to apply a unidimensional IRT model.

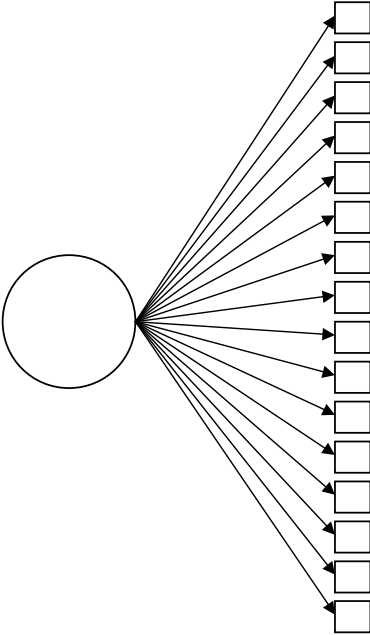
In Model B, the data matrix contains more than one common dimension, however, the dimensions are uncorrelated. This is a trivial case of multidimensionality and is easily addressed by forming subscales and then fitting unidimensional IRT models to the separate subscales. This is essentially equivalent to assuming the dimensions are uncorrelated.

Model C also has more than one common factor among the items, however the factors are correlated. We refer to such a representation as a non-hierarchical multidimensional model. Although this type of structure can also be handled by forming subscales and then fitting separate unidimensional IRT models, non-hierarchical multidimensional IRT models may be used to more efficiently assess individuals on multiple dimensions simultaneously.

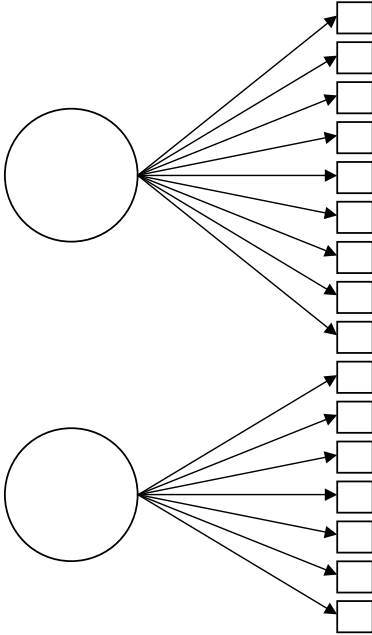
Finally, Model D is a bifactor model, namely, there is a general factor that explains the item intercorrelations, but in addition, there are also so-called “group” factors that attempt to capture the residual variation due to secondary dimensions. Model D, a hierarchical model, is also shown in Figure 2. Each item is allowed to have a positive loading on a general trait that is assumed to underlie all the items. In addition, each item can load on one or more “group” factors. In most applications, a bifactor model is specified so that each item loads on only one group factor, and the general and group factors are all orthogonal to each other. We will look to see if the item loadings on the general factor (λ_{13} - λ_{83}) are sufficiently high after accounting for the group factor loadings (λ_{11} - λ_{41} and λ_{52} - λ_{82}).

Figure 1. Four Possible Latent Variable Models

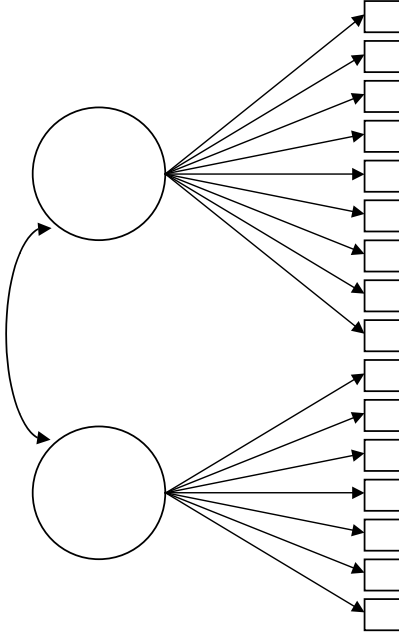
MODEL A



MODEL B



MODEL C



MODEL D

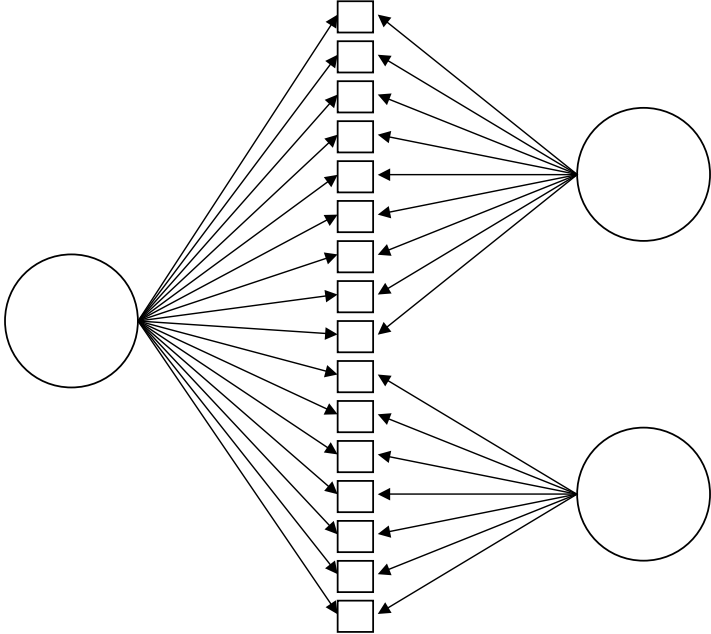
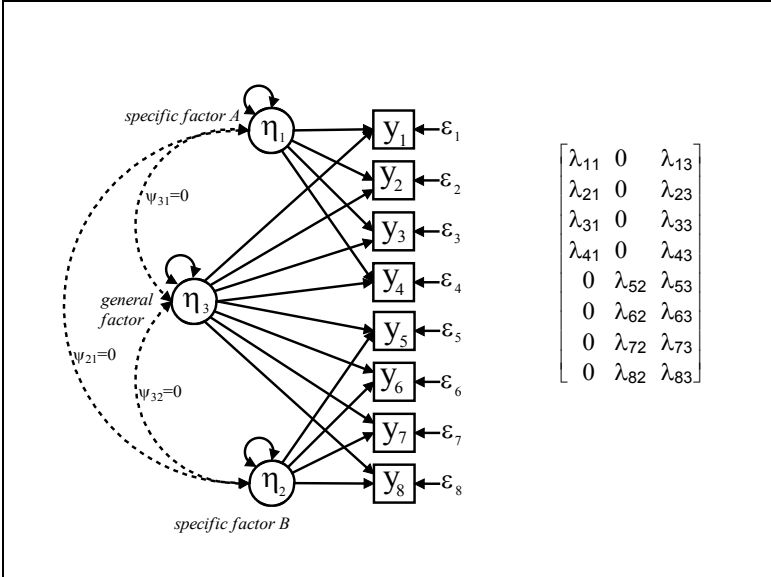


Figure 2: Model for Mcdonald's (1999) Hierarchical Factor Analysis (courtesy of Rich Jones)



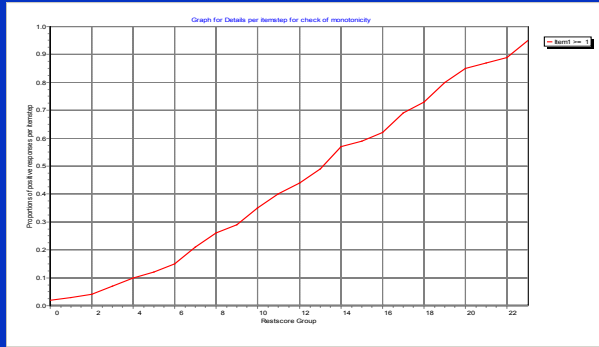
Appendix B: Examples of Evaluating Monotonicity

The rest-score function for the item in Figure 1 shows a well functioning item. The conditional item means increase rapidly with each one-unit change in rest scores. In addition, at low rest score levels, the conditional item mean approaches zero, and at high levels, the conditional item means approaches one. This item is well suited for IRT modeling. The item graphed in Figure 2 also satisfies the monotonicity assumption, but the rest-score function illustrates a poorly functioning item. Clearly, this item is not very useful in differentiating between higher and lower scorers on this scale, because response endorsement rates (conditional means) do not increase very much across the rest-score continuum. The rest-score function for Figure 3 indicates an item that violates the monotonicity assumption.

Analysis of monotonicity can be conducted using Mspwin 5.0 (Mokken Scaling Program, Molenaar & Sijtsma, 2000a). Monotonicity is assessed in MSP by creating an “empirical” IRF. Specifically, restscores (total scale score minus the response to the item – the total of the “rest” of the items) are calculated for each item, and then the proportion of respondents in each restscore group endorsing the item is tabulated. Restscore groups are defined by specifying the minimum number of individuals, *Minsize*, to be grouped together by restscore value in order to calculate proportion endorsed. “Adjacent restscore groups are joined if they contain less than *Minsize* individuals, in order to avoid instable proportion estimates.” With large sample sizes, we can use a *Minsize* parameter of 200 for determining the number of restscore groupings. If monotonicity is attained by the empirical IRF, then the proportion endorsed in each restscore group should increase as restscores increase. If the proportion endorsed in a restscore group is lower than the preceding restscore group(s) then a violation of monotonicity has occurred. A certain amount of minor violations are expected due to random fluctuation so Mspwin calculates if the probability of the violation is less than 5% and records any violation falling in this range as a “real” violation. The Mspwin program also combines information about the item’s scalability (i.e., degree to which item responses conform to a Guttman scale) and monotonicity (the frequency, the size of the monotonicity violations and their statistical significance) into a single test of monotonicity violation per item (*Crit*). Values of *Crit* in excess of 80 strongly suggest that the monotonicity assumption has been violated. Thus, any item with a *Crit* value above 80 will be flagged as violating monotonicity and will be deleted from further analyses.

Figure 1

Item satisfying monotonicity assumption

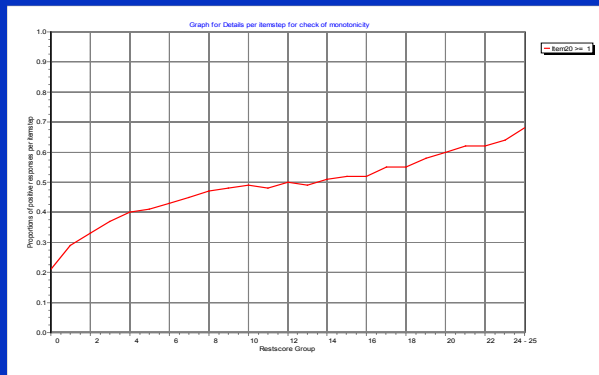


1 12/19/2004

Figure 2

Poorly functioning item

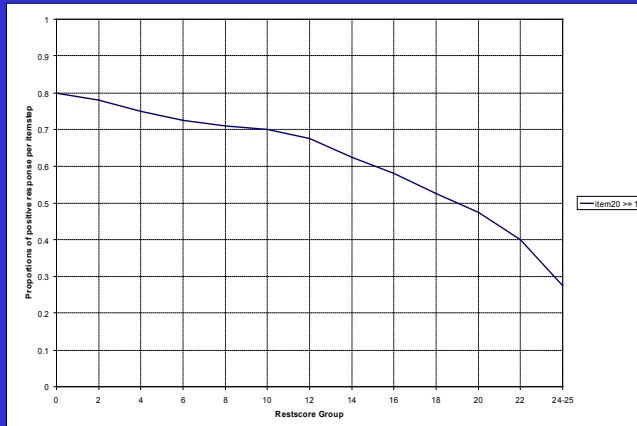
Figure 2



1 12/19/2004

Figure 3

Poorly functioning item



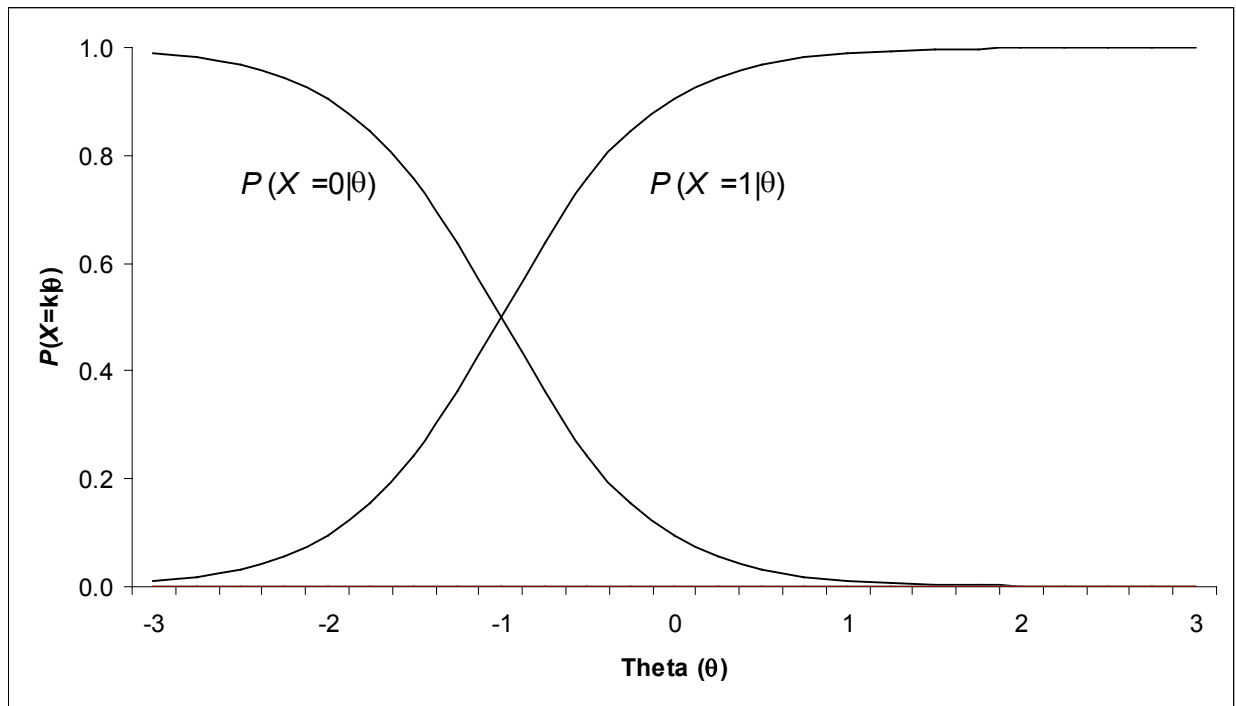
Appendix C: Constructing the Graded Response Model

For dichotomous response data, the IRT two-parameter logistic (2PL) model trace line for the probability of a positive response to item i for a person with latent trait level θ is:

$$P(X_i = 1|\theta, b_i, a_i) = \frac{1}{1 + \exp[-a_i(\theta - b_i)]},$$

where both the items threshold parameter (b) and discrimination parameter (a) vary from item to item.

Figure 5: IRT 2PL model curves with $a = 2.26$ and $b = -1.00$. Two curves (one a linear transformation of the other ($1 - P(X)$)) are presented representing a “false” or “no” response ($P(X = 0|\theta)$) and a “true” or “yes” response ($P(X = 1|\theta)$).



Samejima’s (1969) graded response model (GRM) is a generalization of the 2PL model described above. The GRM is based on the logistic function giving the probability that an item response will be observed in *category k or higher*. For ordered responses $X = k$, $k = 1, 2, 3, \dots, m$, where response m reflects the highest θ value, the graded model trace line is:

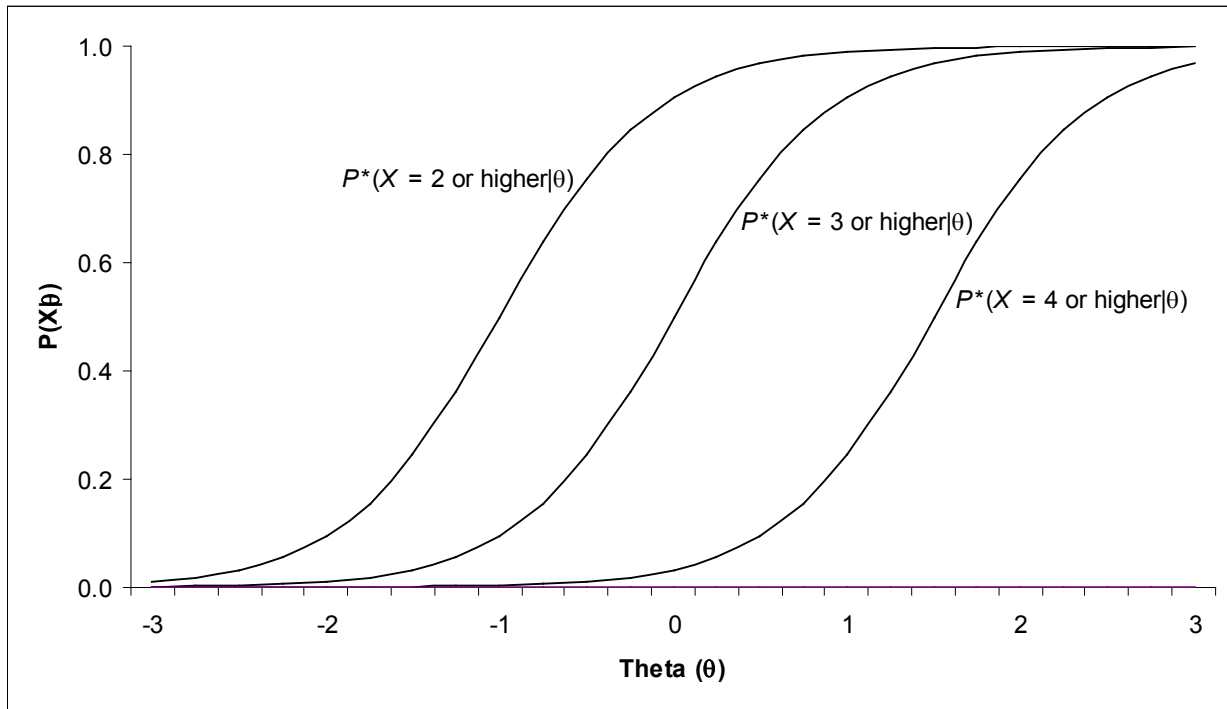
$$P(X_i = k|\theta, b_i, a_i) = \frac{1}{1 + \exp[-a_i(\theta - b_{i,k})]} - \frac{1}{1 + \exp[-a_i(\theta - b_{i,k+1})]}$$

The trace line models the probability of observing each response alternative as a function of the underlying construct. The slope a_i varies by item i , but within an item, all response curves share the same slope (discrimination). This constraint of equal slope for responses within an item keeps trace lines from crossing, thus avoiding negative probabilities. The threshold parameters

b_{ik} varies within an item with the constraint $b_{k-1} < b_k < b_{k+1}$. At each value $\theta = b_k$, the respondent has a 50% probability of endorsing the category.

There are two stages for computing the category response probabilities in the GRM. The first step is estimating the probability $P^*(k|\theta, b_i, a_i)$ that a respondent at any particular level of θ will respond in that scoring category (k) or a higher category. Figure 6 presents the Operating Characteristic Curves for each $P^*(k|\theta, b_i, a_i)$ function for a four-response category item. The curve representing the probability for any respondent to select response one or any higher category is not presented in Figure 6 because that probability is one.

Figure 6: Operating characteristic curves for a four-response category item under the GRM with $a = 2.26$, $b_1 = -1.00$, $b_2 = 0.00$, and $b_3 = 1.50$. Each curve from left to right shows the probability of being in category “two or higher”, “three or higher”, and “four or higher”, respectively, given a person’s level on theta (θ). The curve (not shown) representing the probability of being in the first category or higher is simply 1 (i.e., 100% probability you will respond to any response category).



The second step is to estimate the GRM category response curves $P(X_i = k|\theta, b_i, a_i)$; which represent the proportion of participants responding to that category across θ . This is represented by the following function:

$$P(X_i = k|\theta, b_i, a_i) = P^*(X_i = k|\theta, b_i, a_i) - P^*(X_i = k + 1|\theta, b_i, a_i),$$

which matches the same structure as presented in the equation above. Each category response curve, $P(X_i = k|\theta, b_i, a_i)$, will be a nonmonotonic curve, except for the first and last response categories.

For the first response category $k = 1$, $P^*(X_i = 1 | \theta, b_i, a_i) = 1$; therefore, the trace line $P(X_i = 1 | \theta, b_i, a_i)$ will have a monotonically decreasing logistic function with the lowest threshold parameter:

$$P(X_i = 1 | \theta, b_i, a_i) = 1 - P^*(\text{category 2 or higher} | \theta), \text{ alternatively,}$$

$$P(X_i = 1 | \theta, b_i, a_i) = 1 - \frac{1}{1 + \exp[-a_i(\theta - b_{i,2})]}.$$

For the last response category $k = m$, $P^*(X_i = m+1 | \theta) = 0$; therefore, the trace line $P(X_i = m | \theta, b_i, a_i)$ will have a monotonically increasing logistic function with the highest threshold parameter:

$$P(X_i = m | \theta, b_i, a_i) = P^*(\text{category } m \text{ or higher} | \theta) - 0, \text{ alternatively,}$$

$$P(X_i = m | \theta, b_i, a_i) = \frac{1}{1 + \exp[-a_i(\theta - b_{i,m})]}.$$

Figure 7 presents the category response curves derived from the operating characteristic curves for the four-response category item presented above.

Figure 7: Category response curves for a four-response category item with IRT GRM parameters: $a = 2.26$, $b_1 = -1.00$, $b_2 = 0.00$, and $b_3 = 1.50$.

