Developing Tailored Instruments: Item Banking and Computerized Adaptive Assessment

Jakob B Bjorner MD, PhD,

QualityMetric Incorporated, Lincoln, RI, and

Health Assessment Lab, Waltham, MA

The basic idea of a Computerized Adaptive Test (CAT) is to have a computer select the items that seem most appropriate for a particular respondent (given our knowledge so far) and to score the responses in a way that allows comparison with respondents answering a different set of items (Wainer et al., 2000). This results in a quicker and more accurate assessment. The logic of a typical CAT is shown in Fig 1. The test begins with an initial estimate of the respondent's score (Step 1). This could be based on the response to an initial global question that is asked of all respondents, or on previous information about the respondent. A global question should be informative for the average person and have appropriate content for a first item. The initial score is used to select the most informative item, which is administered at Step 2. The answer is used at Step 3 to re-estimate the score. At Step 4, a respondent-specific confidence interval (CI) is computed for the score estimate. At Step 5, the computer determines whether any stopping rules have been fulfilled. If the stopping rule is test-precision the computer evaluates whether the CI is within specified limits. Once the standard is met, the computer either begins assessing the next concept or ends the battery. Otherwise, step 2 is repeated for the next most informative item.

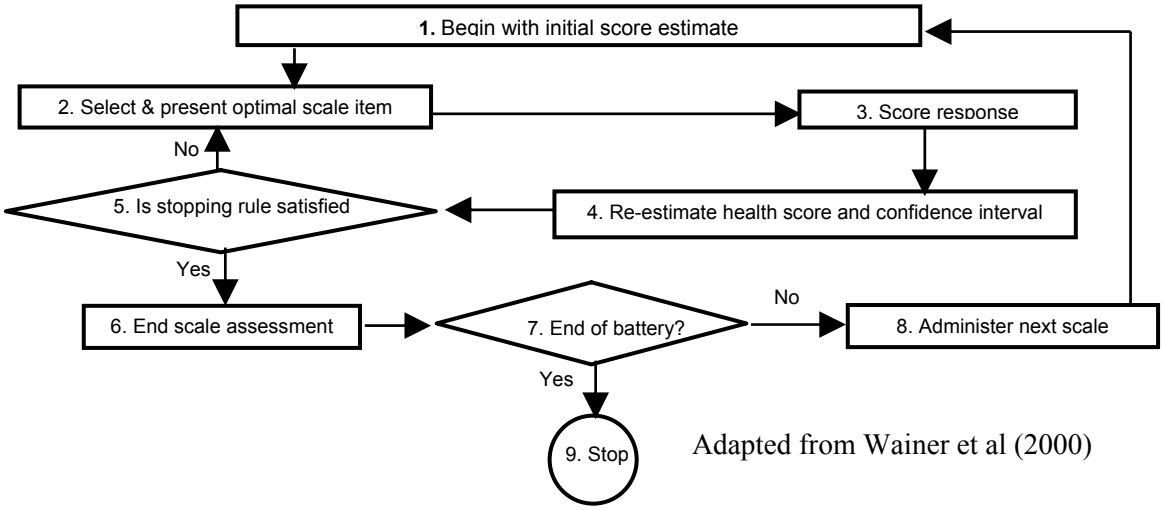


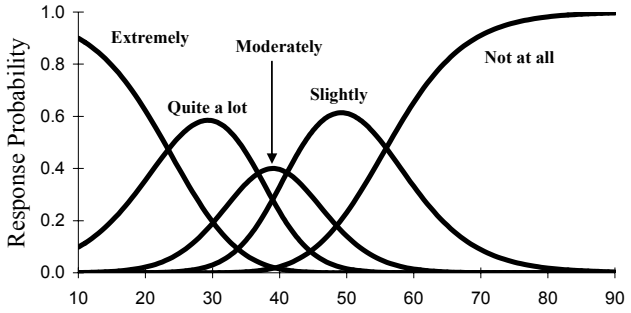
Figure 1. Logic of Computerized Adaptive Testing

The CAT needs a set of rules for selecting the most appropriate items and scoring them on a common ruler, and a bank of items that can be chosen for the test. Item banks contain information on the wording of each item, the concept it measures, and its measurement characteristics according to a measurement model. Most CAT-based assessments utilize Item Response Theory (IRT) to select items and to score the responses.

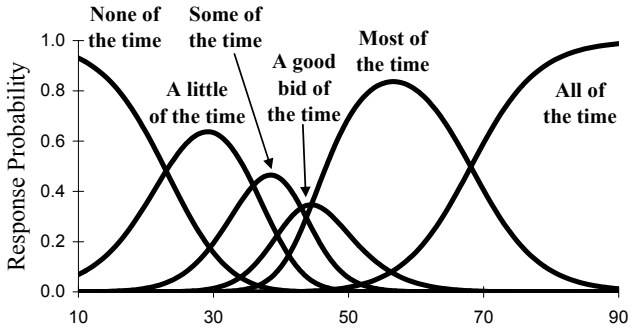
IRT models are statistical models of the relationship between a person's score on the concept being measured and their probability of choosing each response choice for each item measuring that concept. IRT models can evaluate how informative an item is for a specific range of scores. Fig. 2 shows examples of the IRT models that are used to evaluate item information and estimate the persons IRT score. The figure illustrates the IRT models for three items concerning mental health (the three upper plots) and the corresponding item information functions (lower plot). These examples use the Generalized Partial Credit IRT model (GPCM) (Muraki, 1997). Each curved line in the upper plots represents the models' prediction of the probability of choosing each of the item response categories for various degrees of mental health. The curves are called item characteristic curves (ICC), option characteristic curves, or trace lines. The horizontal axis is the mental health IRT score, "normed" so that the average adult in the USA has a score of 50; a positive score means better mental health. The figure show that for the average IRT score of 50, the most likely response on SF8MH (*... how much have you been bothered by emotional problems...*) is *slightly* (probability 0.61), the most likely response on MHP01 (*... how much of the time have you been a happy person*) is *most of the time* (probability 0.69) and the most likely response on MHC01 (*... felt so down in the dumps that nothing could cheer you up*) is *none of the time* (probability 0.83). MHC01 is most relevant for persons with poor mental health and can be said to be the 'easiest' item, because the probability of getting a high item score for a given IRT score level is higher on MHC01 that on the other items. In contrast MHP01 can be said to be the hardest item.

The GPCM is characterized by two types of item parameters: thresholds and slopes. The item threshold parameters are the values on the horizontal axis where the item characteristic curves for two adjacent categories intersect, and the slope parameter (only one for each item) is a function of the slope of the curves. In figure 2, item 3 has higher slope than item 2, which has higher slope than item 1.

SF8MH. During the past 4 weeks, how much have you been bothered by emotional problems (such as feeling anxious, depressed or irritable)?



MHP01. During the past month, how much of the time have you been a happy person?



MHC01. How much of the time, during the past month, have you felt so down in the dumps that nothing could cheer you up?

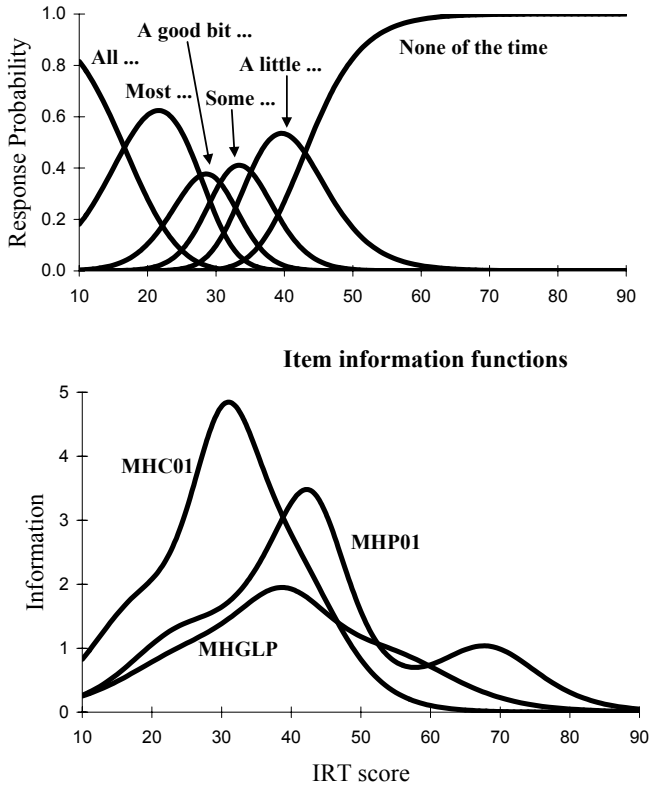


Figure 2. Trace lines (item characteristic curves) and item information functions for three items on mental health.

The information functions shown in the lower part of figure 2 express the contribution of each item to the overall test precision for various levels of mental health. These functions can be calculated from the IRT model (Muraki, 1993). Figure 2 shows that MHC01 is most informative for people with poor mental health and that MHP01 is the most informative items for people with good mental health.

The IRT model can be used to estimate the score of the person, once an answer is given. Unlike the traditional sum scoring approach, IRT score estimation can be performed when only a minor subset of the items is answered. Figure 3 illustrates two possible sequences of score estimation and item selection in a CAT. Let us assume that we have no prior knowledge of the respondents and that all answer SF8MH as the first item. The Bell curve in the first row represents our prior assumption about the distribution of mental health in the population. The mean (expected) IRT score is 50, but a wide range of values are possible (95% prediction interval 30 to 70). If the answer to SF8MH is *extremely* (second row, left column) the curve for this response choice (black line) is multiplied with the prior distribution, which produces the "Posterior distribution 1" (row three). The new IRT score estimate is 30; the mean of the posteiror distribution. The prediction interval is 17-42 and considerable narrower than for the prior distribution. At an IRT score of 30, MHC01 provides much more information than MHP01 (Figure 2) and would thus be the logical choice for the next item. If the respondent answers *a good bit of the time* to this item (row four) the curve for this response choice is multiplied with posterior distribution 1 to produce posterior distribution 2 (row 5). The IRT score estimate is now 29 with a 95% prediction interval of 21-36. If we had access to a large item bank and wanted more precision, we could continue to ask questions to narrow down the prediction interval.

If another respondent answered *not at all* to the first item, SF8MH, the CAT would take a different route (row two, right column). Multiplying the prior distribution and the curve for this response choice leads to an IRT score estimate of 58 with a 95% prediction interval of 44-74 (row three). In this score range, the MHP01 item provides more information and would be the logical choice. If the respondent responded *most of the time* to this items, out IRT score estimate after two items would be 57 with a 95% prediction interval of 46-70. Again, we could ask more questions to get more precision. However, the MHC01 would be of little value or relevance for assessing this person. The respondent would be highly likely to select the response *none of the* time and the item would add very little information for this range of IRT scores (see figure 2).

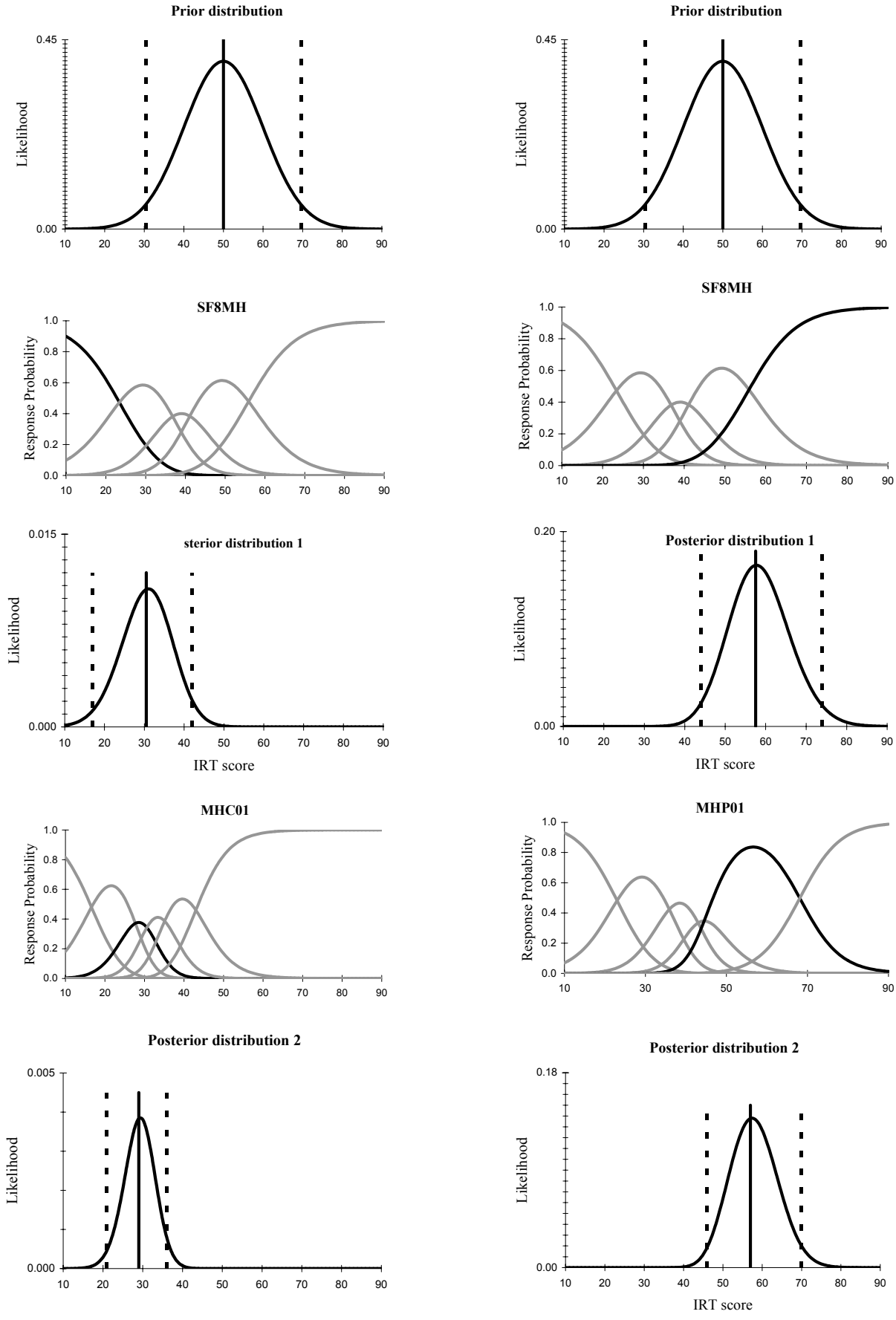


Figure 3. Two possible CAT scenarios.

Note that although the two respondent have answered different questions, their scores are on the same scale and can be compared. No matter which or how many items from the item bank are answered, the IRT score is on the same scale.

The combination of CAT and IRT provides several advantages:

1. Test relevance and precision can be optimized for a given respondent burden.
2. Precision can be adapted to the needs of the specific application. If we do not require high precision for a given purpose the assessment can be stopped early to reduce respondent burden, if high precision is required, more items can be administrered.
3. Scores are placed on a uniform metric regardless of which items in the bank are used.
4. Item banks can be expanded gradually by seeding and evaluating new items.
5. The response process can be monitored in real time to ensure assessment quality and that inconsistent response patterns are explored.

Examples of CAT-based assessments of generic and disease-specific health outcomes can be found at https://www.amihealthy.com.

Steps in the development of an item bank for CAT

To achieve a CAT of high quality, we need an item bank (or "item pool") containing a sufficient number of items fitting an IRT model. Developing an item bank for a CAT involves the following steps (see (Bjorner, Kosinski, & Ware, Jr., 2003), for an example in the outcomes field):

Construct definition and item development

Meaningful assessments require clearly defined constructs and good items. Careful specification of the subdomains of the constructs and the domains that are not part of the constructs ensures that the item bank covers all relevant aspects of the constructs. Often this involves specifying hypotheses to be tested in later stages, e.g. whether some domains can be seen as part of a common construct (dimension) or whether they should be treated as two separate constructs (dimensions).In developing an item bank for mental health, we reanalyzed data on a well researched tool, the 34-item *Mental Health Inventory (MHI)* (Veit & Ware, Jr., 1983). This questionnaire builds on a conceptual model for mental health that includes five subdomains: Anxiety, Depression, Behavioral/Emotional Control, Positive Well-being and Loneliness/Belonging. We tested whether these subdomains could be seen as part of one common domain (see below). The analyses of this item bank will be used here to illustrate some of the steps in item bank development. In subsequent data collection and analyses, we expanded the item bank by 40 additional items.

If high measurement precision throughout the range of IRT scores is required, steps have to be taken to develop items that are relevant for the extremes of the scale. Such items often have poor item-total correlations or other psychometric properties, and therefore they tend to be lacking from traditional questionnaires. For the mental health item bank, we collected new data to calibrate a broad set of additional mental health items into the item bank.

The item information functions showed in figure 2 are fairly typical in the sense that the items provide most information for people with poorer than average health. It is often a challenge to develop items that provide high precision for people with better than average health.

Collecting data for item calibration and testing

We need a sufficient sample size with a sufficient spread over the range of outcomes to allow for estimation of all model parameters. For IRT models like the GPCM, sample sizes of 500 – 1,000 are probably sufficient (Tsutakawa & Johnson, 1990). For simpler models (e.g. the Rasch type models where the slope parameter is constrained to be equal for all items) even smaller sample sizes may work. For very hard and for very easy items, some response categories are rarely used – presumably because few respondents score at the extremes covered by these item responses (for example, the response *all of the time* to MHC01, see figure 2). In such circumstances, it might be helpful to over-sample respondents in the ranges for which we want to establish good measurement precision.

Ideally, the data collection method for item calibration should be the same as the data collection method in the final CAT (most often a computer interface, although interviews, phone interviews, and automated phone interviews may also work well with a CAT). People tend to give more positive responses (better QoL) in personal telephone interviews than in self-administered postal surveys (McHorney, Kosinski, & Ware, Jr., 1994) in particular for mental health questions (Bjorner, Ware, Jr., & Kosinski, 2003). We therefore expect differences (although not major) between IRT parameters for data gathered by interviews and by postal/computerized surveys - mainly for mental health and mainly for the threshold parameters. Major differences between computerized and paper and pencil surveys seem unlikely.

Fitting an IRT model and testing model assumptions

The statistical analyses of the item pool serves several purposes: test model assumptions, to identify items and item response choices that do not function well, to select the best IRT model, and to get item parameters. The steps in the IRT analysis are described later.

Setting the metric

After the CAT has been developed, the researcher has to decide how the metric (IRT score) should be defined. For generic health status measures it may be convenient to standardize the metric to a general population (e.g. the U.S. population), setting the mean to 50 and the standard deviation to 10. For disease-specific concepts, the metric could be based on a well-defined population of people with the given disease. The population that defines the metric need not answer all the questions in the item bank – only enough questions to set the metric precisely. For the mental health item pool, we used a five-item subset of the mental health inventory (the MHI-5) to define the metric. These five items were administered to a representative sample of the US general adult population. The metric was then set so this population achieved a mean score of 50 and a standard deviation of 10.

CAT design and pretesting by simulations

To use the item bank in a CAT, item selection rules and stopping rules must be defined. Simulated CAT runs are very effective in evaluating the impact of various rules on test length, precision, and validity. One approach is to run simulations of a CAT on the data already collected (so-called "real simulations") (Sands, Waters, & McBride, 1997). These simulations can implement the steps shown in fig 1. The total set of responses used to develop the item pool are used as input, but during the simulation the computer only reads the responses that correspond to the questions that would have been asked during a real CAT. Another possibility is to simulate item responses based on the IRT model, and use these simulated responses as input to the CAT; this is particularly useful when the item bank has been developed by linking items across several studies, so that no respondent has actually answered all items.

Figure 4 shows results of a 'real' simulation of the Mental Health CAT based on our initial item pool. In this simulation we specified that each 'respondent' should be administrered the five most informative items. The figure shows that this method has excellent agreement with the score based on 31 items. Within the limits set by the total information in the item pool, the precision of a given CAT can be set to meet the need of a given situation. For the Mental

Health CAT, we decided that higher precision we necessary for people with poor mental health (thus in high risk for having e.g. depression (Berwick et al., 1991)). Therefore we defined our standard stopping rules to be based on precision, but with requirements for precision varying over the range: <42 (SEM<3), 42–60 (SEM<4), >60 (SEM<6.6).

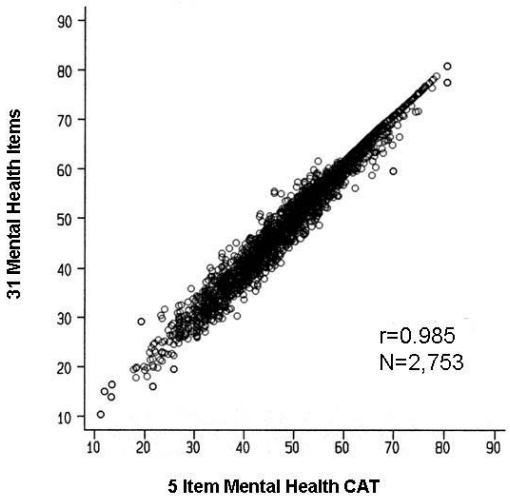


Figure 4. 'Real' simulation of a 5 item Mental Health CAT.

Statistical analyses: Fitting an IRT model and testing model assumptions

Testing dimensionality and local independence

Standard CAT requires that items measure only one dimension and that this dimension explains all covariation between items (the assumption of *local independence*). Although a fully unidimensional item bank is probably not achievable for most theoretically interesting constructs, exploration of dimensionality is crucial in item analysis. The bank needs to be sufficiently unidimensional to make a single score meaningful and to ensure that item parameter estimates (and in turn person IRT scores) are not unduly influenced by problems of multidimensionality or local dependence between items. Exploratory and confirmatory factor analytic methods for categorical data represent strong and flexible approaches to testing dimensionality and local dependence (Muthen & Muthen, 2001), but many other methods exist (e.g. (Christensen, Bjorner, Kreiner, & Petersen, 2002; Stout, Habing, Douglas, & Kim, 1996; Muraki & Carlson, 1995; Chen & Thissen, 1997). If problems are identified, possible solutions include item exclusion, splitting the item pool into two or more pools, using multidimensional CAT, or, in milder cases of multidimensionality, using item selection rules to ensure content balance.

Table 1. Factor correlations for subdomains of Mental Health. N= 2717.

	Anxiety	Depression	Behavioral	Positive	Loneliness
Anxiety	1				
Depression	0.89	1			
Behavioral/Emotional Control	0.88	0.96	1		
Positive Well-being	0.81	0.90	0.91	1	
Loneliness/Belonging	0.71	0.82	0.85	0.86	1

Data from the Medical Outcomes study (Ware, Jr., Bayliss, Rogers, Kosinski, & Tarlov, 1996)

For the mental health item pool, we evaluated dimensionality by categorical data factor analysis, comparing a unidimensional model to a five dimensional model (the five original subdomains). Table 1 shows some of the results from these analyses – factor correlations in a five-factor model run in the combined data set across 5 diseases (we also ran analyses that treated the diseases separate groups). The factor correlations are high, except for the Loneliness/Belonging domain and for the correlation between Anxiety and Positive Well-being. Based on these and other results, we felt it justified to fit a unidimensional IRT model for the items. We excluded three items from the Loneliness/Belonging domain since they did not load strongly on the factor. Further, we defined item selection rules to ensure content balancing between the three main domains in the pool (Anxiety, Depression/Control, and Positive Well-Being). Thus, no respondent will be given a Mental Health assessment that builds only on Anxiety items, only on Depression items, or only on Positive Well-Being items.

Initial analyses of item characteristic curves by non-parametric methods

Before fitting a parametric model, it is useful to examine non-parametric IRT models that allow visual inspection of the empirical item characteristic curves (Ramsay, 1995). This allows further identification of poor items and response choices. Items can be excluded, a more general IRT model can be used, or response choices that do not discriminate can be collapsed in the IRT analyses.

Fit an item-response model and test model fit

Many different models are available for IRT analyses (van der Linden & Hambleton, 1997; Fischer & Molenaar, 1995; Muraki & Bock, 1996; Thissen, 1991) and the choice of model is sometimes hotly debated. In general, the fit of the model and the model assumptions need to be tested and misfit needs to be dealt with – either by using a more general IRT model or by removing items from the bank.

Test of differential item functioning (DIF)

One of the basic assumptions in outcomes measurement is that items function the same way in different disease and demographic groups. For a given scale or IRT score level, item responses should be independent of group membership. Although DIF is a general measurement problem, it is easiest conceptualized and detected using IRT or similar methods. Evaluation of DIF should be a part of item bank development. Items with DIF can be excluded from the item bank, but if the DIF is well understood, IRT methodology can be used to correct for DIF.

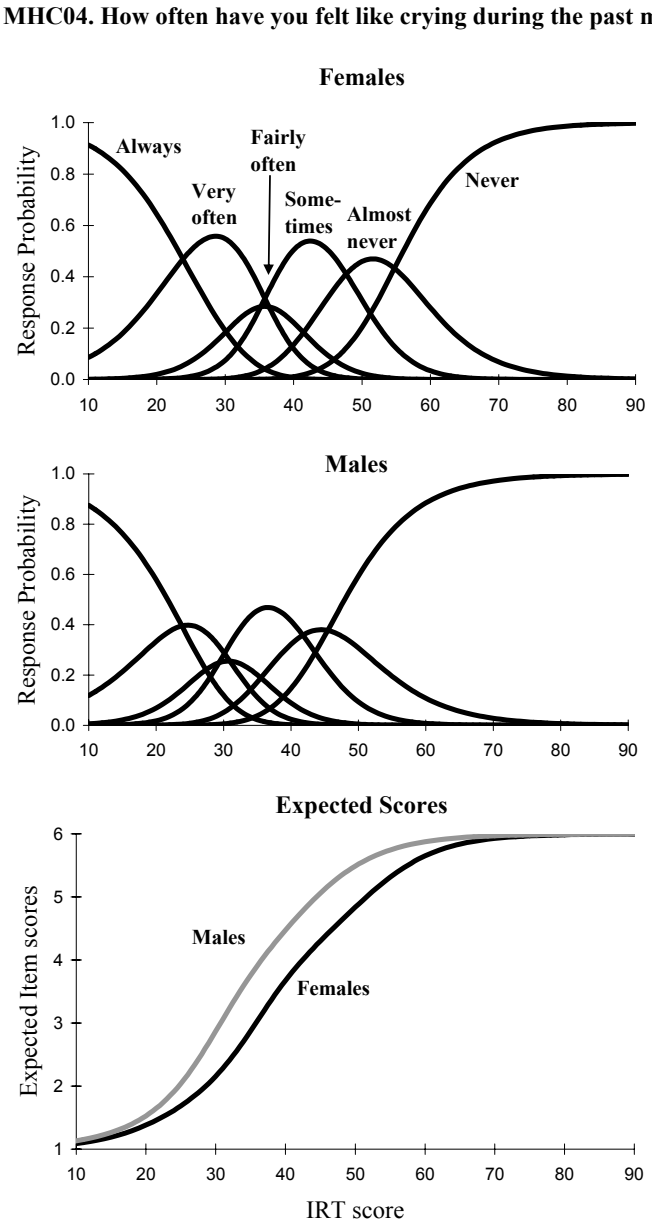


Figure 5. Illustration of Differential Item Functioning (DIF) for gender.

Figure 5 illustrates an example of DIF for gender found in the mental health item bank: the items MHC04 (*How often have you felt like crying during the past month*). The figure shows that this item functions differently for men and women. For a given level of mental health, men are less likely to cry (i.e., the trace lines are shifted to the left). The bottom row of figure 5 summarizes the difference in trace lines between men and women by showing the expected item score (when *Always* is coded 1, *Very often* coded 2, etc). At a mental health IRT score of 50, the expected score on MHC04 is 0.65 higher for men than for women. Given no other information, a woman who has chosen *Very often* on this item would be expected to have a mental health score around 29 (figure 5, first row), while a man who has chosen *Very often* would be expected to have a mental health score around 25 (second row). This instance of DIF can be corrected by using separate IRT models for men and women (as shown in figure 5) or by deleting this item from the bank.

Multidimensional CAT

For patient reported outcomes, the researcher will often want to measure several related constructs and might want to gain measurement precision by utilizing information on the association between the different dimensions. Further, it might sometimes be more realistic to assume that some items are measuring more than one dimension. Both of these tasks can be accomplished by multidimensional CAT, which allows simultaneous measurement of multiple dimensions (Segall, 1996). Such models can be estimated by factor analytic methods for categorical data (e.g. (Muthen et al., 2001) as has been done for a mental health instrument (Gardner, Kelleher, & Pajer, 2002). Multidimensional CAT is an exciting area for future development, but can also be very computer intensive. Currently only a small number of dimensions can be handled within reasonable computational time. Also, the interpretation of scores is more complex.

CAT in educational testing and in outcomes research

CAT was mainly developed for the assessment of abilities (see e.g. (Sands et al., 1997)) and CAT applications in outcomes research builds heavily on the methods developed in educational testing (Wainer et al., 2000). However, applications of CAT in outcomes research differ in three major areas: choice of IRT models, generation of items, and problems of item exposure.

IRT model

Educational tests most frequently use multiple-choice items that are scored right/wrong and analyzed by dichotomous IRT models. Such items are only informative over a narrow range of the scale and uninformative at other levels, which can create problems if the CAT is started at an inappropriate level (van der Linden & Glas, 2000). In contrast, outcomes research mostly uses items that are scored on a rank scale (e.g. 1-5) and analyzed by IRT models such as the GPCM model shown in figure 2 and 3. Such 'polytomous' items provide more information over a much broader range of scores. Therefore, the same level of precision can be attained with fewer items and the choice of starting point is less crucial.

Generation of items

To achieve precision over the full range of a scale, the total item bank needs a large number of items with sufficient diversity. In an educational test, generation of new items is done routinely and the pool of potential items can be seen as unlimited for many topics. In contrast, the number of ways questions can be asked about patient reported outcomes may be limited. Item banks based on pooling items from existing questionnaires may provide good measurement precision in some ranges, but insufficient precision at the extremes.

Item exposure

In educational testing, the assessment needs to take place in a controlled environment and item content needs to be kept secret to avoid cheating. Countering these problems necessitates special test sites, large item pools, and complex procedures for item exposure control (Wainer et al., 2000). For patient reported outcomes, items are not kept secret and item exposure is thus much less of a problem. Thus, CAT in research on patient reported outcomes can be simpler and much more cost-efficient than in educational testing.

Summary

In CAT, a computer selects the items from an item bank that are most relevant for the particular respondent; thus optimizing test relevance and precision. Development of an item bank involves a clear definition of the construct to be measured, good items, a careful statistical analysis of the items (including evaluation of dimensionality, fitting and testing an IRT model for the items, and evaluation of DIF), and a clear specification of the final CAT.

References

Berwick, D. M., Murphy, J. M., Goldman, P. A., Ware, J. E., Jr., Barsky, A. J., & Weinstein, M. C. (1991). Performance of a five-item mental health screening test. *Medical Care, 29,* 169-176.

Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003). Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). *Quality of Life Research, 12,* 913-933.

Bjorner, J. B., Ware, J. E., Jr., & Kosinski, M. (2003). The potential synergy between cognitive models and modern psychometric models. *Quality of Life Research, 12,* 261-274.

Chen, W.-H. & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Educational and Behavioral Statistics, 22,* 265-289.

Christensen, K. B., Bjorner, J. B., Kreiner, S., & Petersen, J. H. (2002). Tests for Unidimensionality in Polytomous Rasch Models. *Psychometrika, 67,* 563-574.

Fischer, G. H. & Molenaar, I. W. (1995). *Rasch Models - Foundations, Recent Developments, and Applications*. (1 ed.) Berlin: Springer-Verlag.

Gardner, W., Kelleher, K. J., & Pajer, K. A. (2002). Multidimensional adaptive testing for mental health problems in primary care. *Med.Care, 40,* 812-823.

McHorney, C. A., Kosinski, M., & Ware, J. E., Jr. (1994). Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. *Medical Care, 32,* 551-567.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Appl Psychol Measur, 17,* 351-363.

Muraki, E. (1997). A Generalized Partial Credit Model. In W.J.van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 153-164). Berlin: Springer.

Muraki, E. & Bock, R. D. (1996). Parscale - IRT based Test Scoring and Item Analysis for Graded Open-ended Exercises and Performance Tasks (Version 3) [Computer software]. Chicago: Scientific Software Inc.

Muraki, E. & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Appl Psychol Measur, 19,* 73-90.

Muthen, B. O. & Muthen, L. (2001). Mplus User's Guide (Version 2) [Computer software]. Los Angeles: Muthén & Muthén.

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

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized Adaptive Testing: From Inquiry to Operation*. Washington (DC): American Psychological Association.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

Stout, W., Habing, B., Douglas, J., & Kim, H. R. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement,* sychological.

Thissen, D. (1991). Multilog - Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory (Version 6) [Computer software]. Chicago: Scientific Software Inc.

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

van der Linden, W. J. & Glas, C. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer Academic Publishers.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. Berlin: Springer.

Veit, C. L. & Ware, J. E., Jr. (1983). The Structure of Psychological Distress and Well-Being in General Populations. *J Consult.Clin.Psychol., 51,* 730-742.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J. et al. (2000). *Computerized Adaptive Testing: A primer*. (2 ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Ware, J. E., Jr., Bayliss, M. S., Rogers, W. H., Kosinski, M., & Tarlov, A. R. (1996). Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service systems. Results from the Medical Outcomes Study. *JAMA, 276,* 1039-1047.