Computerized Adaptive Testing and Item Banking

Jakob B Bjorner * ‡ †, MD, PhD, Mark Kosinski*, MA, John E. Ware, Jr.* ‡, PhD,

* Quality Metric, Inc.
Lincoln, RI

‡ Health Assessment Lab
Boston, MA

† National Institute of Occupational Health
Copenhagen, Denmark

Address all communications to:

Jakob B Bjorner, MD PhD, QualityMetric, Inc. 640 George Washington Hwy, Lincoln, RI,

Phone: 401-334-8800, x271, Fax: 401-334-8801, E-Mail: jbjorner@qualitymetric.com

Key words: Item response theory, computerized adaptive assessment, health status, questionnaires

Word Count: 6,026

Running Title: Computerized Adaptive Testing and Item Banking

Introduction

Improving the validity and precision of our measurement tools and making them more practical are constant challenges for the quality of life field. Computerized adaptive testing (CAT) holds great promise in helping us meet these challenges (see e.g. (Wainer *et al*., 2000)). The basic idea of a CAT is to have a computer to select the items that seem most appropriate for a particular respondent (given our knowledge so far) and to score the responses in way that allows us to compare the results with results from other respondents answering a different set of items. This results in a quicker and more accurate assessment. To function, 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 a set of statistical models building on item response theory (IRT, see also chapter 1.5 and 2.3) to select items and to score the responses. The combination of CAT and IRT provides several advantages compared to current practice:

1. By selecting the most appropriate items for each person, assessment precision is optimized for a given test length and irrelevant items can be avoided.
2. Assessment precision can be adapted to needs of the specific application. For example, for a diagnostic purpose precision should be high for scores close to diagnostic cut-points, or test precision could be set high over all the score range for purposes of follow-up of individuals.
3. Assessments can be compared even if different items have been used or different precision levels have been specified.
4. Item banks can be expanded gradually by seeding and evaluating new items, without sacrificing backwards comparability.

5. By including items from traditional questionnaires in the item bank, it is possible to cross-calibrate widely used questionnaires.
6. The response process can be monitored in real time to ensure assessment quality and aberrant response patterns explored.
7. At the end of the assessment, the respondent (or a health professional) can be given a score immediately, along the guidelines on how to interpret the score.

Although some of these advantages can be achieved with other methodology, the use of CAT and the careful analysis of items that is required for IRT modeling significantly improve assessment quality.

This chapter will be organized in the following way: We start by demonstrating the logic of CAT. We then describe how to build the item banks that underlie a working CAT and discuss practical aspects of CAT. Since discussions of CAT have almost exclusively taken place within educational testing, we briefly outline the differences between that field and applications in research on quality of life. Finally, we discuss some advanced topics and the future challenges for CAT in quality of life research. Throughout the chapter we rely heavily on examples from the CAT-based Headache Impact Test (HIT) (Ware, Jr. *et al*., 2000; Bjorner *et al*., 2003a; Bjorner *et al*., 2003b; Ware, Jr. *et al*., 2003). Readers may check out the HIT and the CAT-based assessments of generic health outcomes via the internet at https://www.amihealthy.com.

An example of CAT based on IRT methodology

The logic of a typical CAT is shown in Fig 1 (also see (Wainer *et al*., 2000)). 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.

Fig. 2 shows examples of the IRT models that are used to evaluate item information and estimate the persons IRT score (often represented by the Greek letter θ (theta), see chapters 1.5 and 2.3). The figure illustrates the IRT models for three items concerning headache impact (the three upper plots) and the corresponding item information functions (lower plot). For these examples, we used 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 headache impact. The curves are called item characteristic curves (ICC) or trace lines (see chapter 1.5). The horizontal axis is the headache impact IRT score, "normed" so that the average headache sufferer in the USA has a score of 50; a positive score means more than average headache impact. The plots show that a respondent with a score of 50 has a 52% probability of answering *A little of the time* to the question on the impact of headache on family interactions (Item 1), a 30% probability of answering *None of the* time, a 16% probability of answering *Some of the time* and low probabilities of choosing other categories. In contrast, to the question on needing help in routine tasks (Item 2) a respondent with a score of 50 has a much

higher probability of choosing *None of the time/A little of the time* (which have been combined in this analysis) than any other response. 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 (see chapter 1.5). In figure 2, item 3 has higher slope than item 2, which has higher slope than item 1.

The information functions, which express the contribution of each item to the overall test precision for various levels of headache impact, can be calculated from the IRT model (see chapter 2.3).

Fig. 3 shows an example of how the ICC and information functions are used in carrying out the tasks outlined in fig. 1. The upper part of fig. 3 shows the likelihood of various levels of headache impact for a person that has not yet answered any questions (the prior distribution). For headache impact, the prior distribution can be approximated by a normal distribution (Bjorner *et al*., 2003a). The mean (expected) IRT score is 50, but a wide range of values are possible (95% prediction interval 30 to 70). Around a score of 50, item 1 has the highest information function (fig. 2) so this item is picked as the first item. Suppose the respondent replies *Some of the time* (middle plot). The curve for this response choice (black line) is multiplied with the prior distribution, which produces the "posterior distribution". The expected value (mean) of this distribution is 57, and the prediction interval is 43 to 70, which is considerably narrower than for the prior distribution. However, this is still a wide confidence interval, and so Step 2 can be repeated by selecting the next most important item. For scores around 57, item 3 has higher information function than item 2 and is thus the

optimal second item. If one of the middle response categories is chosen, the response curve for these is multiplied with the previous likelihood – resulting in a new posterior distribution (score estimate 62, prediction interval 51 to 71). Finally, the answer to item 2 (*Some of the time*) results in an estimated score of 64 with a prediction interval of 55 to 71. No matter which or how many items are answered, the IRT score is on the same scale, and we achieve higher precision by asking more questions.

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. The methods outlined in chapters 1.5 and 2.3 are also used when developing an item bank for a CAT. We briefly review the steps in item bank development.

Construct definition and item development

Meaningful assessments require clearly defined constructs. 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). For example, for the HIT test a content analysis of previous questionnaires revealed 6 subdomains: headache pain, (impact of headache on:) role functioning, social functioning, energy, cognitive functioning, and mental health, indicating potential multi-dimensionality.

Good items are crucial for a well functioning CAT. In principle, the criteria for good items do not differ between a CAT and a traditional questionnaire. In many standard questionnaires, items are often presented in a grid to save space. However, in a CAT items are normally presented one at a time, so there is rarely a need to adapt the items to a grid format. 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. The measurement of minor disease impact is a challenge because items aimed at minor impact generally have lower slopes than items directed at major disease impact.

We recommend including items from existing questionnaires in the item bank, since this allows cross-calibrating of scores (so that the results of a CAT can be expressed in the metric of traditional questionnaires – see below). For the HIT, items were developed in several steps: the initial item pool was based on items from existing headache and migraine questionnaires; after gathering data and evaluating the IRT models for these items, we developed an additional pool of items – using the IRT results and input from clinicians.

Collecting data for item calibration and testing

To achieve precise estimates of item parameters and to check the model we need a sufficient sample size with a sufficient spread over the range of quality of life to allow for estimation of all model parameters. Sample size requirements depend on the match between the item thresholds and the IRT score levels in the sample. However, even when the match is good, the sample has to be fairly large, so we can ignore errors in the estimates of item parameters. For the dichotomous 2-parameter IRT model, simulation studies have found sample sizes of

500 – 1,000 to be sufficient (Tsutakawa and Soltys, 1988; Tsutakawa and Johnson, 1990). It is likely that similar sample sizes are required for the polytomous items. For very hard and for very easy polytomous items, some response categories are rarely used – presumably because few respondents score at the extremes covered by these item categories. In such circumstances, it might be helpful to oversample 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 *et al*., 1994) in particular for mental health questions (Bjorner *et al*., 2003c). 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.

We recommend that the following 6 steps are carried out in item bank development:

1. Basic descriptive analyses (proportion of missing, frequency distribution, skewness etc.)
2. Test dimensionality and local independence

3. Initial analyses of item characteristic curves by non-parametric methods to detect potential problems for standard parametric IRT modeling
4. Fit an item-response model and test model fit
5. Test of differential item functioning (DIF, (Holland and Wainer, 1993) and chapter 3.4)
6. Test whether unfortunate choices of items could introduce bias/multidimensionality ("random multidimensionality")

Steps 1-5 are described in other parts of this book (chapters 1.5, 2.3, 3.4) and will not be discussed in detail here. However, we will briefly mention some of the lessons learnt analyzing the HIT item pool.

Standard CAT requires that items measure only one dimension and that this dimension explains all covariation between items (the assumption of *local independence*, see chapter 1.5). Further, multidimensionality is also an important source of DIF (see chapter 3.4). Although a fully unidimensional item bank is probably not achievable for most theoretically meaningful constructs, exploration of dimensionality is one of the most important parts of 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 and local dependence between items. Many different methods exists for testing dimensionality and local dependence (Muthen and Muthen, 2001; Holland and Rosenbaum, 1986; Bock *et al*., 1988; Christensen *et al*., 2002; Stout *et al*., 1996; Muraki and Carlson, 1995) (see chapter 1.5). We rely strongly on exploratory and confirmatory factor analytic methods for categorical data (Muthen and Muthen, 2001).

There are no definitive rules for deciding when multidimensionality or local dependence is of sufficient magnitude to prose problems. However, it is easy and important to test the robustness of the model. In case of potential multidimensionality, we recommend fitting an IRT model for each subdimension to see whether the item parameters and the IRT scores differ notably from the ones achieved if unidimensionality was assumed. In case of local dependence between a pair of items, we recommend examining whether item parameter-estimates change when one item of the pair is excluded. In the analysis of the HIT item pool, we applied both approaches and found no major problems.

Based on factor analytic results and tests of robustness, we may choose to exclude items with poor loadings from the pool, or split the item pool into two or more pools. However, milder cases of multidimensionality may be handled in a CAT by specifying item selection rules to ensure that every test includes items from all subdimensions, ensuring content balance. Cases of local dependency between pairs of items (which inflates slope parameters and leads to overestimation of precision) can be handled by estimating item parameters for each item separately and specifying item selection rules to ensure that only one of the items in the pair is selected during any CAT.

A problem unique to CAT is 'random multidimensionality'. Since the items used in any particular CAT are a subset of the items in the pool it is possible that the item selection procedure may introduce bias/multidimensionality by focusing on a narrow subdomain of the item bank. This might occur even if analyses of the total item pool do not indicate multidimensionality. Such problems might necessitate more sophisticated item selection procedures (to achieve content balance). In developing the HIT item pool we examined this problem using several methods: additional factor analyses of the subset of items that were

frequently used in CAT simulations (see below), and examination of bias from different "worst case" scenarios of subsets of items. For the HIT item pool, we found no such problems.

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 or a more general IRT model used (such as the nominal categories model instead of the GPCM), or response choices that do not discriminate can be collapsed. Items with low slopes have flat information functions and will therefore rarely or never be picked in a CAT (Ware, Jr. *et al*., 2003). In this way, CAT is automatically protected against poor items. However, it is important to check whether some content areas are always omitted; if so, the CAT is assessing a narrower concept than intended.

Setting the metric

After the CAT has been developed, the researcher has to decide how the metric (IRT score) should be defined. In Rasch type models, the metric is often defined by the items: 0 is set as the mean of all item thresholds. In other IRT models, the metric is often defined by the population: the population mean is routinely set to 0, and the standard deviation to 1. However, calibration from one metric to the other is possible, as are other definitions of the metric. For generic health status measures it may be convenient to standardize the metric to a general population (e.g. US population), setting the mean to 50 and the standard deviation to 10. For disease-specific concepts, a mean defined by the general population may have little meaning. A more appropriate metric could be based on a well-defined population of people

with the given disease. Note that the population that defines the metric need not answer all the questions – only enough questions to set the metric precisely.

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 *et al*., 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 using the first item bank developed for the HIT. The line represents the standard error of measurement (SEM) for assessments using the total item bank (48 items), the dots represents SEM for CAT assessments using only the maximum information item selection rule and a stopping logic based on number of items administered (5 items for all persons). The 5-item CAT has a SEM that is below 4 points over most of the range, but higher for low impact. Additional analyses showed high concordance between the total score and the CAT score and lack of systematic deviations (bias). In worst-case CAT scenarios (where the least informative item was systematically chosen) the CAT score was much less precise but still without bias (results not shown).

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 CAT-HIT, we decided that high precision was not necessary for minor headache impact. Therefore stopping rules were based on precision, but with requirements for precision varying over the range: <40 (SEM<7.5), 40–49 (SEM<4), 50–58 (SEM<3), 59–75 (SEM<2.75) and 75+ (SEM<4.8). While the mean number of items for a CAT with these specifications was still 5, we achieve higher precision for people with severe headache impact, and people with minor impact were not burdened with many questions (see (Ware, Jr. *et al*., 2003)). Table 1 summarizes the patterns of item usage given these stopping rules: 20 items were used in the 1011 simulations, but some were rarely used. The 20 items included all the conceptual subdomains outlined for the HIT. Thus, the CAT procedure did not lead to exclusion of specific subdomains.

CAT in practice

Establishing a working CAT requires a great deal of work beyond the psychometric analyses described above. Although a few software packages allow users to develop his/her own CAT (e.g. FastTest Pro (Weiss, 2001)), to date these packages have not included models for polytomous data.[1] Thus, until now the researcher has been forced to write CAT software for quality of life research. Among the issues that need to be considered in software development are:

1. Integrating maximum likelihood item selection with other item selection principles (e.g. defining rules to achieve content balance see (van der Linden, 2000))
2. Visual appearance of items for different systems, screen setting etc.
3. Data security

[1] However, the next version of FastTEST (Version 2) is scheduled to include models for polytomous items.

4. Backup in case of system failure
5. Should the respondent have the possibility of changing previous answers?
6. Check of data quality
7. Saving data from a CAT
8. Providing feedback to respondents, clinicians and other persons
9. Integration with existing patient information systems

Detailed discussion of these issues is beyond the present chapter, but we address some aspects of data quality evaluation and presentation of results.

Evaluating data quality

Using figure 2 to evaluate the response pattern from our CAT example reveals that these responses (Item 1 - *Some of the time*, Item 2 - *Some of the time,* and Item 3 - *Mostly true*) are very likely for a person with a score estimate of 64 (the probabilities are .47, .35, and .63). However, let us consider another set of responses: Item 1 – *All of the time*, Item 2 – *Definitely false*, Item 3 – *Some of the time*. This combination of responses also leads to an estimated score of 64 (and a 95% prediction interval of 55 to 71) but figure 2 shows that these responses are much less likely (probabilities .02, .35, and .32). Another way to way to illustrate this is to compare the likelihoods of each response combination (.106 vs. 0.002). Such comparisons of likelihoods form the basis for IRT-based data quality indicators (Drasgow *et al*., 1985; van Krimpen-Stoop and Meijer, 2002). Such indicators can serve as a warning signal to indicate potential misreading of items or a too simplistic IRT model (in this hypothetical example, headache impact might not be unidimensional after all).

Presentation of results

Although the IRT score is nice from a theoretical point of view, it is advisable to do as much work as possible to make the score easy to interpret for the respondent, the clinician, and the fellow researcher. Tools to do this are "benchmarks" and cross calibration tables. One advantage of an IRT model is that it enables benchmarks from the content of the questionnaires: fig. 2 will for example quickly tell you that a typical person with a score of 50 will very rarely have need for help with routine tasks because of headaches.

Cross-calibration tables use the IRT score to predict scores on other (traditional) questionnaires on the same topic. Such tables establish comparability with results and interpretation guidelines already established from previous research using these questionnaires. If IRT parameters have been established for the items in a traditionally sum-scored scale, the expected sum score (for each level of the IRT score) can be estimated: 1. Calculate the expected item score (for each IRT score level) by calculating the product of the response choice probability and the response choice weight (e.g. *none of the time*=1, *a little of the time*=2 …) and sum over all response choices, 2. Sum the expected item scores (se e.g. (Bjorner *et al*., 2003b)). Fig. 5 shows examples of expected score values on some traditional headache scales for each level of the HIT score.

CAT in educational testing and in quality of life research

CAT was mainly developed in the setting of educational testing and most CAT research and all major books are based on this framework. Applications of CAT in QoL research differ from those in educational testing in three major areas: generation of items, choice of IRT models, and problems of item exposure.

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 quality of life 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, in particular for people with relatively good quality of life. Thus, it is a challenge to develop new items targeted at specific ranges of quality of life. In our view, the potential for developing such items has not been fully explored yet.

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 and Pashley, 2000). In contrast, quality of life research mostly uses items that are scored on a rank scale (e.g. 1-5) and analyzed by polytomous IRT models (like the graded response model or the generalized partial credit model). Such 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 prior distribution is less crucial. Furthermore, because of the many informative response choices for each item, a CAT will have power to detect response inconsistencies, even when items are targeted to the score level of the respondent (van Krimpen-Stoop and Meijer, 2002).

Item exposure

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

Advanced topics

We have dealt with a bank of items that measure one unidimensional construct. In quality of life assessment, 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; Segall, 2000). Such models can be estimated by factor analytic methods for categorical data (e.g. (Muthen and Muthen, 2001) and the parameters for these models can be converted to IRT parameters (e.g. (Gardner *et al*., 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.

Conclusion

In this chapter we have tried to illustrate how CAT works and how CAT can be used to achieve more precise, relevant, and useful measurement. We conclude with another relevant

question: What are the disadvantages of CAT? One potential disadvantage is mode of delivery: The traditional paper and pencil questionnaires that have been a robust and cost-efficient data collection method does not work with CAT (although some intermediate forms exist (Larkin and Weiss, 1975)). However, computerized administration of questionnaires will probably become much more frequent in the future, making the transition to CAT easier.

Since IRT is fairly explicit about model assumptions, criticism has been raised that the assumptions are too strong and not likely to be met with real data. In our opinion, the same assumptions are made implicitly in traditional measurement as are made explicitly in IRT. However, it is possible that CAT is less robust to violations of the measurement assumptions. Thus, careful checking of model assumptions is crucial for a successful implementation of CAT. Once model violations are identified, their impact can be evaluated and often corrected by selecting an appropriate adaptation of the CAT methodology. Thus, the real disadvantage of CAT is the cost for establishing large item banks of high quality.

Such increased costs seem justified when considering the measurement gains: reduction in response burden, increase in measurement precision, creation of a common metric, availability of real time quality control, and immediate feedback. CAT applications in the health field have been seen to have the potential to revolutionize how symptoms and treatment outcomes are assessed (NIH, 2003). It is up to the researchers in quality of life assessment to carry that revolution forward.

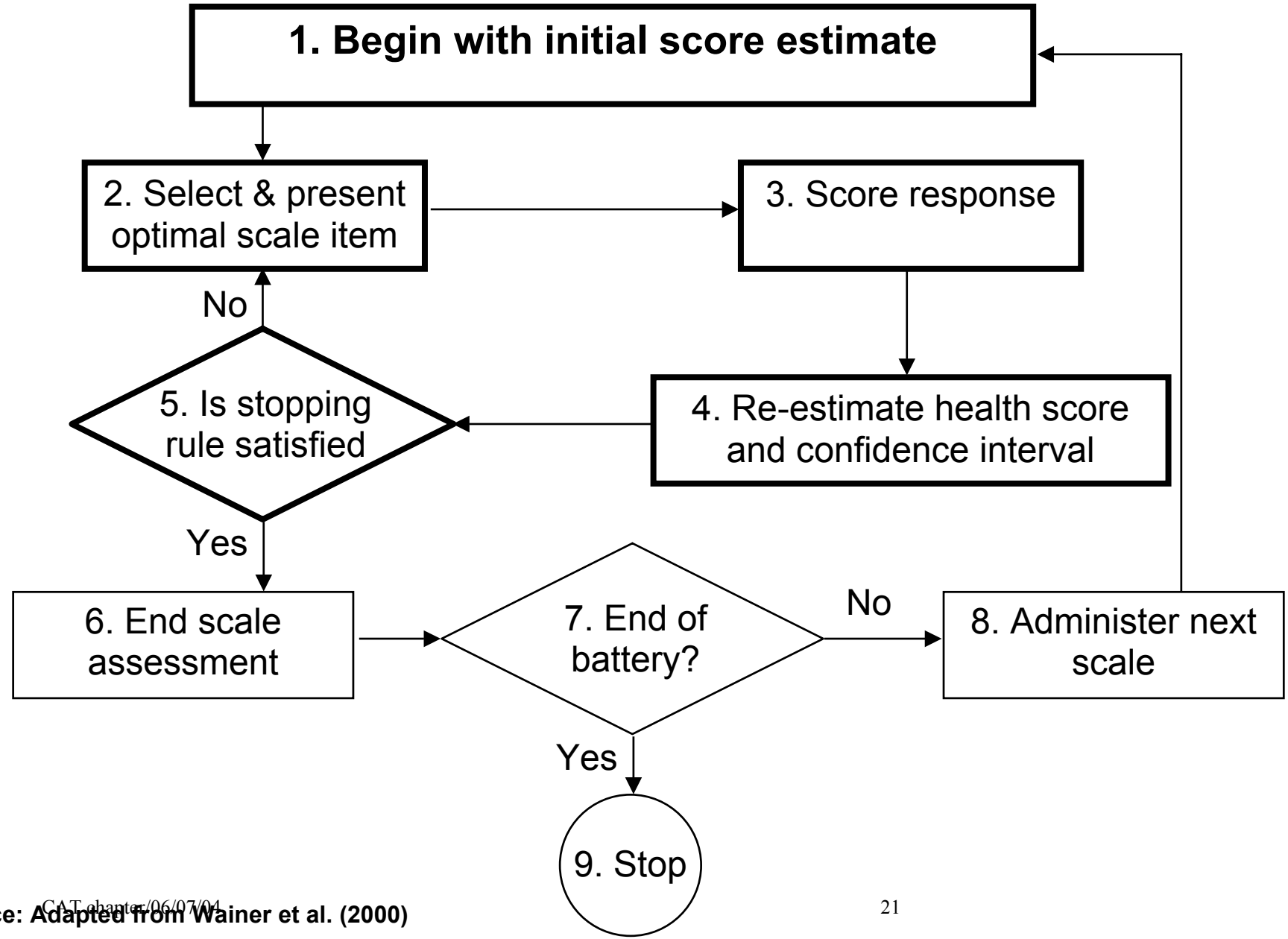
Acknowledgements

We would like to thank Morten Aa. Petersen, Christopher Dewey, and Peter Fayers for helpful comments on a previous version of this chapter.

Table 1. Characteristics of the 20 items selected in simulations of CAT-HIT.

Abbreviated Content	Domain	# Times Admin[1]	# Choices
How often is pain severe	Pain	1011	5
Restricted daily activities	Role	848	3
Feel too tired	Vit	541	5
Reduced activities, chores	Role	409	5
Fell frustrated	Emot	397	3
Restrict recreational activities	Role	397	3
Difficult achieve life goals	Role	265	3
I feel handicapped	Role	261	3
Reduced, non-work activities	Role	211	5
Limit ability to concentrate	Cog	166	5
Less likely to socialize	Soc	141	3
I am afraid to go outside	Emot	115	3
Unable social activities	Soc	76	4
Avoid social activities	Soc	63	5
Need help routine tasks	Role	48	4
Feel irritable	Emot	42	3
Difficult to focus attention	Cog	23	3
Work ability reduced	Role	18	5
Cancel work/daily activities	Role	18	4
Stress on relationships	Soc	1	3

Figure 1. Logic of computerized adaptive testing



Source: Adapted from Wainer et al. (2000)

Figure 2. Item characteristic curves and information functions.

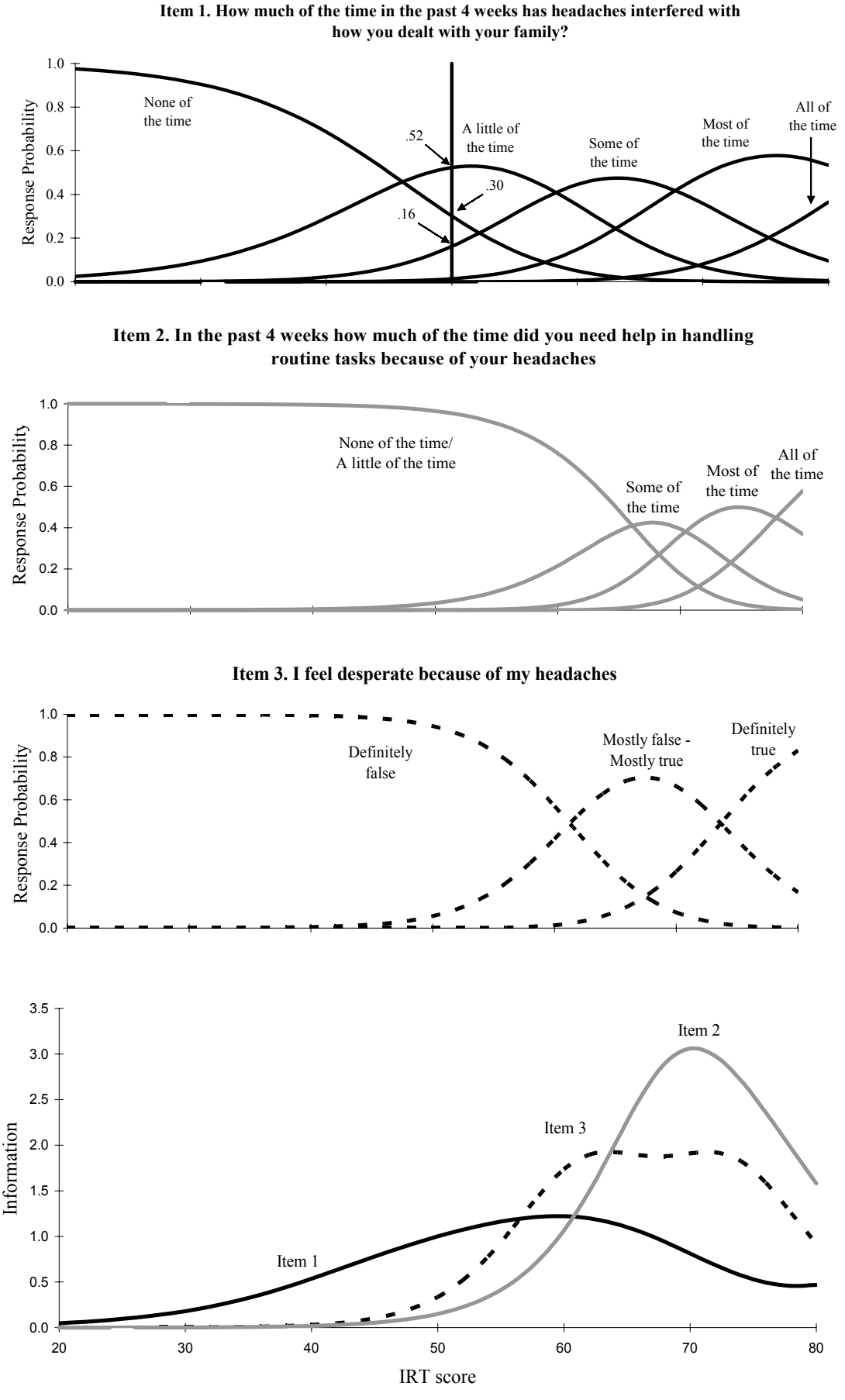


Figure 3. CAT demonstration

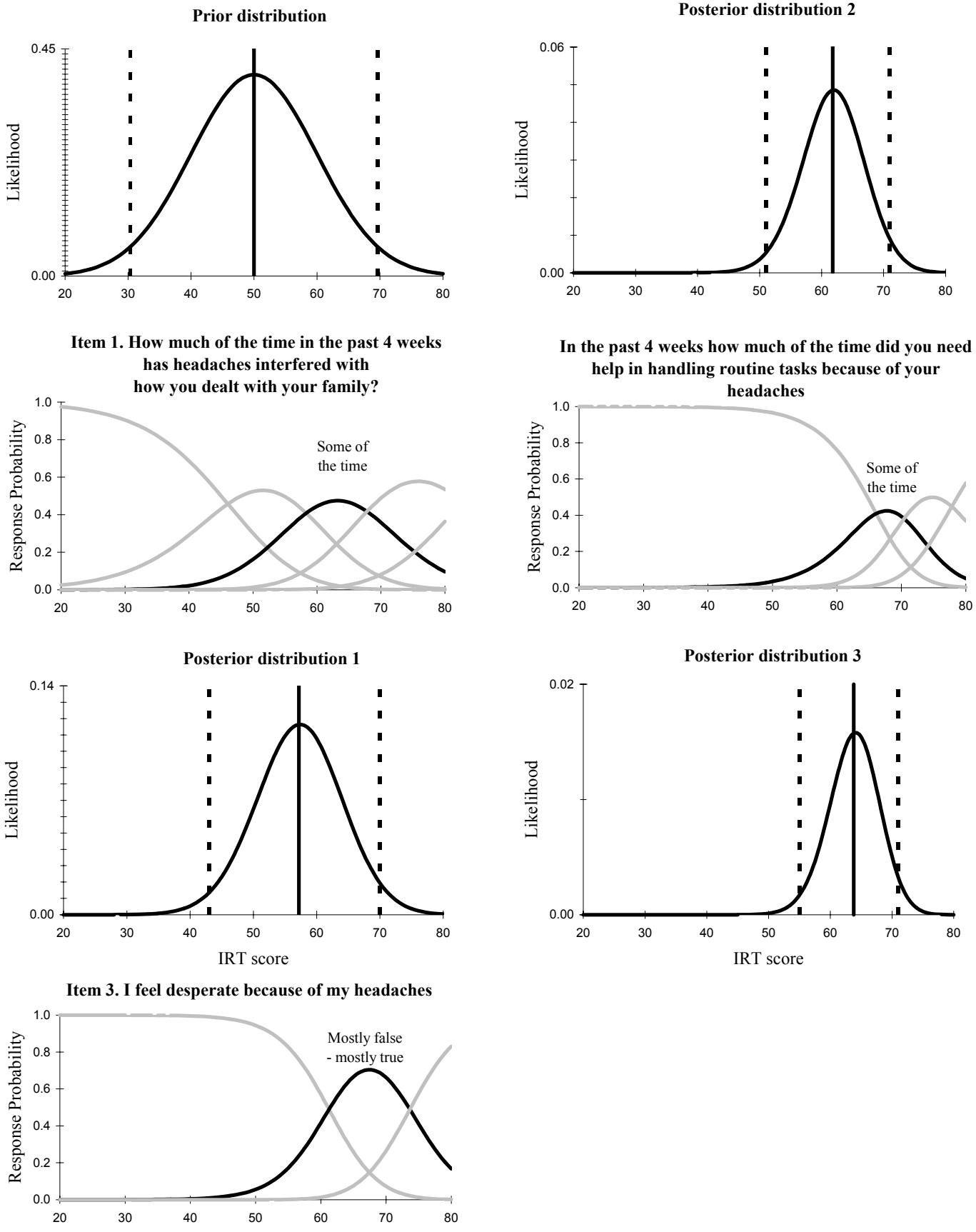


Fig 4. Standard error of measurement for the initial total HIT item bank and for a 5-item CAT.

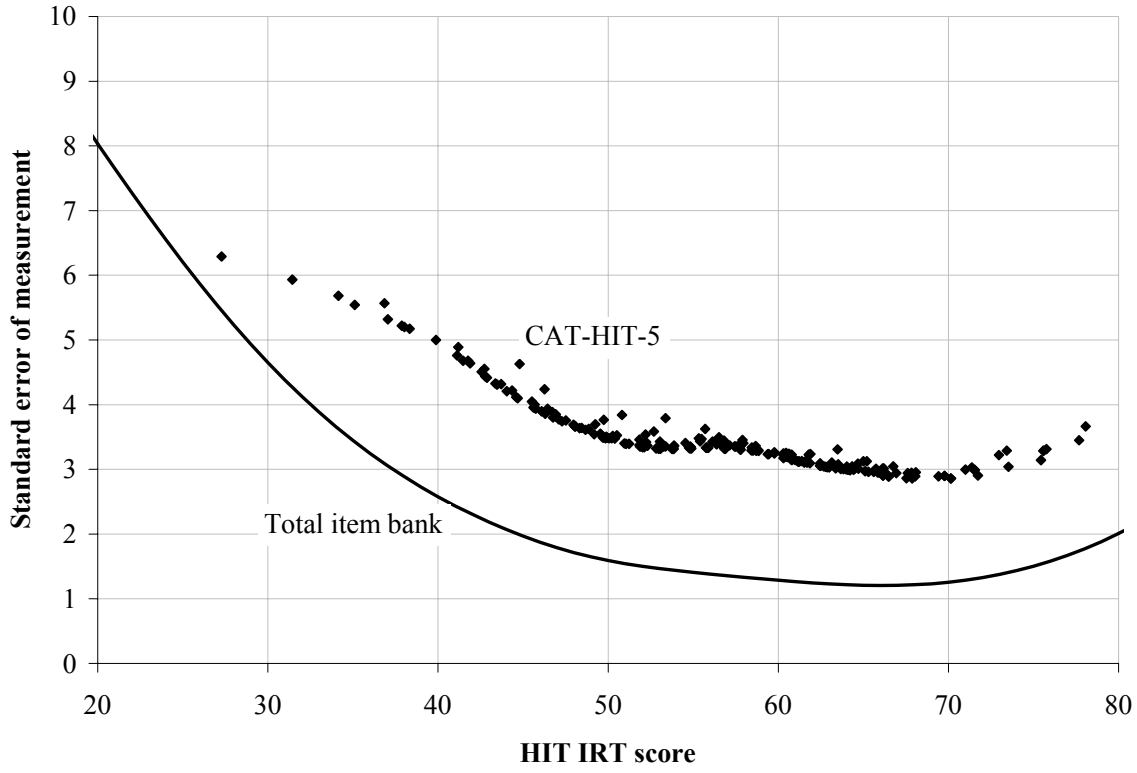
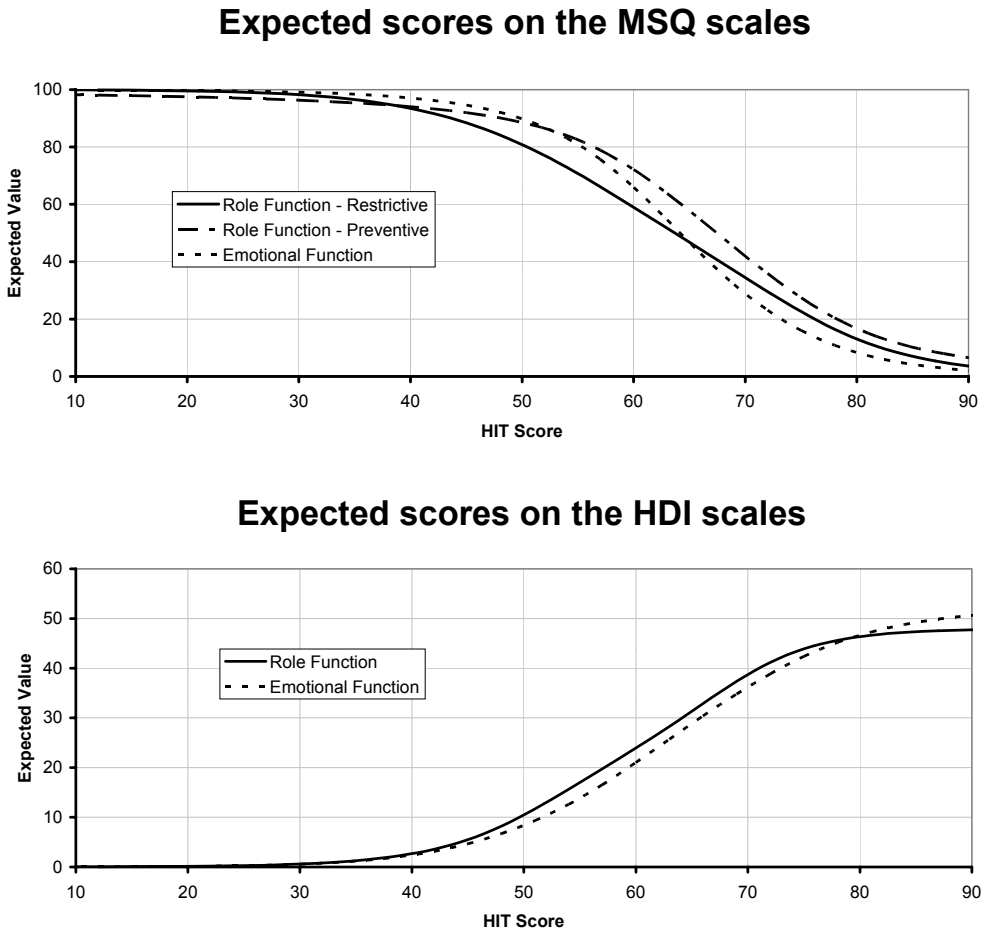


Figure 5. Scale calibration



Reference List

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 test (HIT). *Quality of Life Research, 12,* 913-933.

Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003b). Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Quality of Life Research, 12,* 981-1002.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information Item Factor Analysis. *Appl Psychol Measur, 12,* 261-280.

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

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

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

Holland, P. W. & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Ann Statist, 14,* 1523-1543.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Larkin, R. & Weiss, D. (1975). *An empirical comparison of two-stage and pyramidal adaptive testing* (Rep. No. 75-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric method program.

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. (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. & 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.

NIH (2003). *Re-Engineering the Clinical Research Enterprise*. Bethesda, MD: NIH.

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.

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W.J.van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive Testing, Theory and Practice* (pp. 53-74). Dordrecht: Kluwer Adacemic Publishers.

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

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

Tsutakawa, R. K. & Soltys, M. J. (1988). Approximation for Bayesian Ability Estimation. *Journal of Educational Statistics, 13,* 117-130.

van der Linden, W. J. (2000). Constrained Adaptive Testing with Shadow Tests. In W.J.van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive Testing, Theory and Practice* (pp. 27-52). Dordrecht: Kluwer Adacemic Publishers.

van der Linden, W. J. & Pashley, P. J. (2000). Item Selection and Ability Estimation in Adaptive Testing. In W.J.van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive Testing, Theory and Practice* (pp. 1-25). Dordrecht: Kluwer Adacemic Publishers.

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26,* 164-180.

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., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Med.Care, 38,* II73-II82.

Ware, J. E., Jr., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlof, C. G. et al. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12,* 935-952.

Weiss, D. J. (2001). FastTEST Pro (Version 1.6) [Computer software]. St. Paul: Assessment Systems Corporation.