Quality of Life Research **12**: 913–933, 2003.

Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT™)

Jakob B. Bjorner[1,2], Mark Kosinski[1] & John E. Ware Jr[1,3]
[1]*QualityMetric Incorporated, Lincoln, RI, USA (E-Mail: jbjorner@qualitymetric.com);* [2]*National Institute of Occupational Health, Copenhagen, Denmark;* [3]*Health Assessment Lab, Waltham, MA, USA*

Abstract

Background: Measurement of headache impact is important in clinical trials, case detection, and the clinical monitoring of patients. Computerized adaptive testing (CAT) of headache impact has potential advantages over traditional fixed-length tests in terms of precision, relevance, real-time quality control and flexibility. *Objective*: To develop an item pool that can be used for a computerized adaptive test of headache impact. *Methods*: We analyzed responses to four well-known tests of headache impact from a population-based sample of recent headache sufferers (n = 1016). We used confirmatory factor analysis for categorical data and analyses based on item response theory (IRT). *Results*: In factor analyses, we found very high correlations between the factors hypothesized by the original test constructers, both within and between the original questionnaires. These results suggest that a single score of headache impact is sufficient. We established a pool of 47 items which fitted the generalized partial credit IRT model. By simulating a computerized adaptive health test we showed that an adaptive test of only five items had a very high concordance with the score based on all items and that different worst-case item selection scenarios did not lead to bias. *Conclusion*: We have established a headache impact item pool that can be used in CAT of headache impact.

Keywords: Computerized adaptive testing, Disability, Headache, Health status, Impact, Item response theory, Migraine, Quality of life, Questionnaires, Severity, Tension headache

Abbreviations: CAT – computerized adaptive testing; DIF – differential item functioning; DynHA™ – dynamic health assessment; EAP – expected a posteriori; HDI – Headache Disability Inventory; HIMQ – Headache Impact Questionnaire; HIT™ – Headache Impact Test; IRT – Item Response Theory; MIDAS – Migraine Disability Assessment Score; MSQ – Migraine Specific Questionnaire; MSQOL – Migraine Specific Quality Of Life; RMSEA – root mean square error of approximation

Introduction

Headache is an extremely common symptom that may have profound impact on peoples' functioning and quality of life. The two most common primary headache disorders, migraine and tension-type headache, have one-year prevalences of ≈11% [1] and 40% [2] in an adult population. The disorders are heterogeneous conditions that result in a spectrum of disability within and among different individuals [3]. Although there are effective treatments for most migraine patients [4], migraine is currently under-diagnosed and under-treated [5]. Assessment of headache disability is important to inform decisions about acute and preventive therapy [6]. Several standardized questionnaires for migraine and headache impact have been developed [7–15]. Such questionnaires have at least two potential uses: (1) assessment of groups, e.g. outcome evaluation in clinical studies, (2) assessment of individuals, e.g. identification of patients in need of treatment (screening) or monitoring of the individual patient. Although the available instruments have been proven valuable for group

comparisons, better reliability and measurement precision would be optimal for the purpose of assessing individuals. This problem is not specific to the headache field, but a general potential problem for most widely used health outcomes measures [16]. One solution would be to administer very lengthy surveys in order to increase measurement precision at the level of the individual patient. However, in practice this is not a feasible solution due to the burden it places on a respondent.

A more promising solution to the dilemma between respondent burden and test precision is to use computerized adaptive testing (CAT). In this approach, we start with a large pool of questions (items) and let a computer select the items that are most appropriate for a given person (evaluated from his/her previous answers). Further, the computer scores the responses on a scale that allows comparison with persons answering other questions from the same pool. Various decision rules can be specified for the evaluation of item appropriateness, but the most important factor is how informative the item is in assessing the level of headache impact.

Evaluation of item information is based on modern item response theory (IRT) models (see also [17]). The models are illustrated in Figure 1 that shows the IRT model for two items from the Headache Disability Inventory (HDI) [10, 11]. The solid lines represent the models prediction of the probability of choosing each of the item response categories for various degrees of headache impact. The horizontal axis is the headache impact IRT score; normed so that the average headache sufferer in the USA has a score of zero and a positive score means more than average headache impact. Figure 1 shows that a respondent with a score of zero has a 94% probability of answering *definitely false* to the question on feeling desperate (HDI09E) and a 6% probability of choosing one of the middle categories (*mostly false – mostly true*, categories have been collapsed in this analysis). In contrast, to the question on feeling irritable (HDI11E) a respondent with a score of zero has 20% probability of answering *definitely false*, 70% probability of choosing one of the middle categories *mostly false – mostly true*, and 10% probability of answering *definitely true*. The information functions (the broken lines in Figure 1) express the contribution of each item to the overall test precision for various levels of headache impact. These functions are calculated from the IRT model [18]. Figure 1 shows that the item HDI09E is most informative for a headache impact score in the range 1.0–2.5 (from one to two and a half standard deviation above the mean) while the item HID11E is most informative in the score range −0.9 to 1.1 (also the maximum information value is lower for this item). All other things equal, the computer would choose HDI09 for a person that is believed to have a score around 2 (severe headache impact) but choose HDI11E for a person that is believed to have a score around 0.

The logic of CAT is shown in Figure 2 (also see [19]). The test begins with an initial estimate of the respondent's score (Step 1) that is based on the response to an initial global question that is asked of all respondents. This question should be informative for the average person and should also have an appropriate content for a first item (often the first item is a very general one). The response to the first item is used to select the most informative item from the pool, which is administered at Step 2. The answer is used at Step 3 to re-estimate the score. This estimation is based on the IRT model and the principle of maximum likelihood [20, 21] (see also [17]). For example, if no other information is available a person who answered '*Mostly true*' to HDI09E would get a score of 1.7, the most likely IRT score given that response (see Figure 1). At Step 4, a *respondent-specific* confidence interval (CI) is computed. The confidence interval can be computed as the inverse square root of the sum of the item information functions for the items that have been answered. 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.

Computerized adaptive testing have several noticeable advantages:

1. By selecting the most appropriate items for each person, test precision is optimized for a given test length and irrelevant items can be avoided.
2. Test precision can be adapted to needs of the specific application. For example, for a diag-

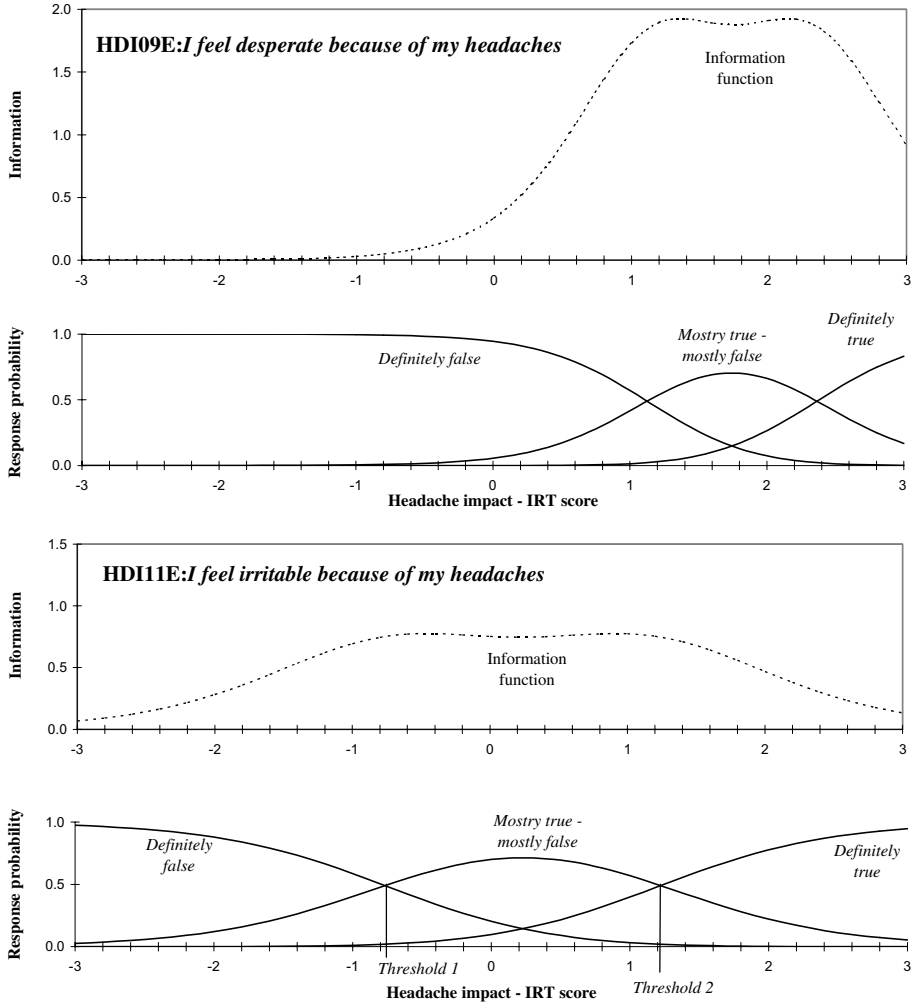

Figure 1. IRT trace lines and information functions for two items on headache impact. The items originate from the HDI [10, 11]. Response choices have been modified from the original. The curved full lines are the trace lines (option characteristic curves) for each response option. HDI09E has a higher slope parameter than HDI11E and thus the trace lines are steeper. The intersections between two adjacent trace lines are the IRT thresholds. They are marked with vertical straight lines. The item information functions (dashed lines) shows the amount of information each item provides at a given level of the IRT score. The information function for HDI09E peaks at a higher level than the item information function for HDI11E. This is because HDI09E has higher slope and because the thresholds are closer together.

nostic purpose test precision could be high for scores close to diagnostic cut-points, or test precision could be set high only for those individuals who are selected for follow-up.

3. By calibrating all items onto a common ruler, test scores can be compared, even if different items have been used or different precision levels have been specified.
4. Item pools can be expanded gradually by seeding and evaluating new items, without sacrificing backwards comparability.
5. The response process can be monitored in real time to ensure assessment quality and explore further in case of aberrant response patterns.

These advantages are well documented within education and psychology [19]. In the medical

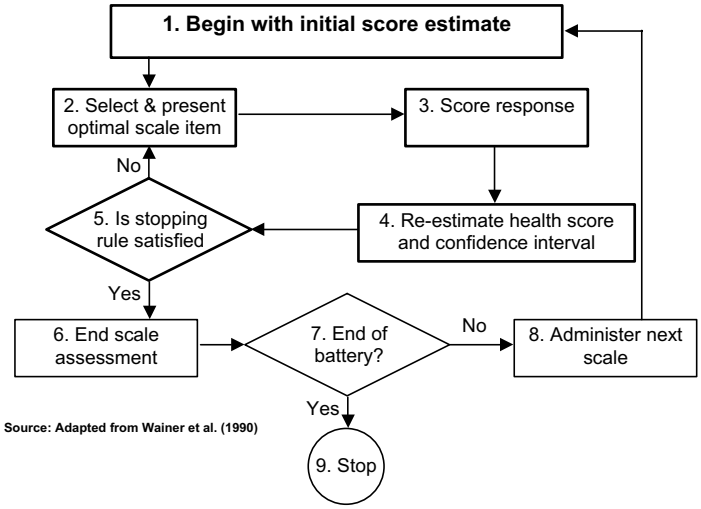

Figure 2. Logic of computerized adaptive testing of headache impact.

field, we have shown for patients in a clinical trial that scoring a headache questionnaire using IRT methodology leads to gains in test precision and responsiveness to change over time compared to the traditional sum score approach [17, 22]. One of the disadvantages of CAT is that very elaborate data collection and analysis is required in order to establish the CAT item pool and to specify the rules for item selection and stopping logic.

This paper documents the psychometric work underlying the development of the headache impact test (HIT), to our knowledge the first computerized adaptive test of any health outcome. Specifically we present results regarding the analyses to develop an item pool for headache impact, and studies to assess the agreement between a short adaptive test and scores based on the total item pool.

Material and methods

Sample

We used the National Survey of Headache Impact, a telephone survey [23]. The study was longitudinal with a baseline interview and a three-month followup. For item pool development, we used the baseline data only. Interviews were completed for 1016 adults between April 21, 1999 and May 12, 1999. The sampling frame was a randomly generated list of telephone numbers from the 48 contiguous US states. Trained interviewers conducted interviews with a convenience sample of eligible respondents after obtaining verbal informed consent. Individuals were eligible for participation if they met all of the following criteria: (1) 18–65 years of age; (2) permanent resident of the household called; and (3) had at least one headache in the four weeks prior to interview (not from a hangover, cold, or flu). In addition, the respondents had to be: (1) able to converse easily in English; and (2) physically and mentally competent to be interviewed. A total of 7510 households were contacted and 5360 (71%) agreed to be interviewed. The prevalence of headache in the past four weeks was 45.7%. All in all, 1533 persons were eligible and 1016 (66.3%) of these completed their interview. The mean interview duration was 21.5 min (range 17–27 min). Of the individuals who did not report having a headache in the past four weeks, 35% reported having a headache in the past three months.

Measures

To benefit as much as possible from prior work and to maintain comparability with current questionnaires for measuring headache impact, our strategy for building the HIT item pool began with using items from four widely used measures of headache impact: the Migraine Specific Question-

naire (MSQ) Version II [7], the HDI [10, 11], the Headache Impact Questionnaire (HIMQ) [12], the Migraine Disability Assessment Score (MIDAS) [13–15] (53 items in total, see examples in Table 1).

The MSQ version 2 was revised in several ways from version 1 (discussed in [17]): the response choices, the wordings of some items, and the ordering of items [7]. For one item, we included both version 1 and version 2 to study the effect of the different response choices. Based on the results from analyses of version 1 we excluded one item on recovery from migraine attacks (MSQ12 in this version). Further, we changed the disease attribution from migraine (which is originally used in the MSQ) to headache because we wanted the item pool to measure the broader concept of headache impact. One advantage of this approach is that many patients with migraine as defined by research criteria are not aware of their diagnosis. Also, the other questionnaires in the pool all use headache as the disease attribution.

The HDI consists of 25 items (Table 1) scored to produce two scales of headache impact: (1) emotional function; and (2) role function [10, 11]. The original instrument includes three response categories with the following values: (1) *Yes* = 4 points; (2) *Sometimes* = 2 points; and (3) *No* = 0 points. For the HIT item pool, the response choice categories of HDI items were modified to include five categories on a true/false continuum.

From the HIMQ [12] we used eight items that originally were scored as a single index of headache impact. The items concern average pain intensity for headaches and lost time in work outside the home, in household work, and in non-work activities in the last 3 months. We modified some HIMQ answer types (0–10 for pain intensity and 0–100% for lost time) to five category rating scales. In addition, the recall period was modified to 30 days.

The MIDAS consists of five items that capture information on lost time from work for pay, housework or chores, and leisure activities over the last 3 months due to headaches [13, 14]. The original items, which require the respondent to write the number of days (from 0 to 90) were included along with two modified MIDAS item sets. The first modified MIDAS item set used a five category rating scale on a continuum from *all of the time* to *none of the time*; the second set used a five category rating scale on a continuum from *always* to *never*. For both modified item sets, the recall period was changed to the previous 30 days. The two modified item sets were randomly administered to one-half of the sample to avoid overexposure to items of the same content.

We performed a content analysis of the HIT item pool using the overall conceptual framework of the Medical Outcomes Study [24]. As illustrated in Table 2, the HIT item pool covers a broader spectrum of health outcomes (ranging from pain to emotional distress) than any of the original scales by themselves. The questions do not cover an overall evaluation of quality of life. In addition to the HIT item pool the interview was comprised of 13 headache screener items (see [25]), 14 general disease screener items, the SF-8, and six items on basic sociodemographic information. The interview and informed consent statement were approved by Essex Institutional Review Board. The interview was pilot-tested (n = 50) before implementation.

Analysis

Our analyses were performed in seven steps, to examine the following research problems (see [23] for discussion of the approach):

1. *Basic descriptive analyses*: proportion of missing, frequency distribution, skewness etc.

2. *Test dimensionality*: whether the items are measuring one underlying dimension or several separate dimensions. We used factor analysis to investigate this issue. We used methods for factor analysis of categorical data [26] because traditional factor analysis might overestimate the number of factors and underestimate the factor loadings when analyzing skewed categorical data [27]. We used weighted least squares estimation with robust standard errors and mean- and variance-adjusted χ^2 statistics as implemented in the Mplus software [26]. We evaluated model fit using the root mean square error of approximation (RMSEA) as implemented for categorical data [26]. For continous data, a RMSEA value below 0.05 is usually taken as an indication of good fit. We did eigenvalue analyses and tested the latent structure hypothesized by the questionnaire developers using confirmatory factor analysis. Eigenvalues were evaluated by Scree plots (see e.g. [28]).

Table 1. Sample of items used in the Headache Impact Test (HIT)

Source Instrument	Scale	Item stem	Response choices
Migraine Specific Questionnaire (MSQ)	Role restrictive	In the *past 4 weeks*, how often have you had *difficulty* in performing work or other daily activities because of headache symptoms?	None of the time (1) – All of the time (5)
Migraine Specific Questionnaire (MSQ)	Role preventive	In the *past 4 weeks*, how often were you *not able to go to* social events such as parties, dinner with friends, because you had a headache?	None of the time (1) – All of the time (5)
Migraine Specific Questionnaire (MSQ)	Emotional function	In the *past 4 weeks*, how often have you *felt* fed up or frustrated because of your headaches?	None of the time (1) – All of the time (5)
Headache Disability Index (HDI)	Emotional function	My headaches make me feel confused. Would you say this statement is…	Definitely true (1) – Definitely false (5)[a]
Headache Disability Index (HDI)	Role function	I avoid traveling because of my headaches. Would you say this statement is …	Definitely true (1) – Definitely false (5)[a]
Headache Impact Questionnaire (HIMQ)[b]		When you have headaches, how often is the pain severe?	Never (1) – Always (5)
Migraine Disability Assessment Score (MIDAS)[c]		On how many days in the *past 3 months* did you miss work or school because of your headaches?	# of days
MIDAS categorical		How much of the time in *the past 4 weeks* did you miss work or school because of your headaches?	None of the time (1) – All of the time (5) *or* Never (1) – Very often (5)[d]

[a] Two response choices added to the choices of the original instrument.
[b] Original instrument used VAS scales with five descriptive anchors. We used a five category scoring with categories defined by the original descriptions.
[c] We studied both the original period of three months and a recall period of four weeks (used in the adaptive tests reported here).
[d] Respondents randomized to two different formats of categorical responses.

Table 2. Content of questions from widely used static questionnaires and HIT

Dimension	MSQ	HDI	HIMQ	MIDAS	HIT
Pain			X		X
Role functioning	X	X	X	X	X
Social functioning	X	X	X	X	X
Energy/fatigue	X				X
Cognitive function	X	X			X
Mental health	X	X			X

3. Analyze the option characteristic curves: Modern measurement theory assumes that each response choice option has a characteristic relation to the latent scale so that each response choice has maximum probability of being selected over a unique interval of the scale (see Figure 1 for illustration). We examined this using non-parametric analyses of the option characteristic curves with the program Testgraf [29, 30]. In case the analyses indicated that two categories were not distinct for a particular item, we collapsed them before fitting the item response model.

4. Fit an item-response model and examine the item properties: We used the generalized partial credit IRT model [31] and the maximum marginal likelihood estimation procedure [32] and the Parscale software [33] (see [34] for a discussion of IRT model choice). Item information functions were calculated using SAS V8.01. We examined item fit as described by Muraki [31]: divided the IRT scores into 20 groups, compared the expected and observed item frequency distribution within each score group, and calculated an overall fit statistic. Since many (53) tests were performed, we run the risk of significant results due to multiple testing. However, since the tests are not independent, a strict Bonferroni correction would be too conservative. As a compromise, we chose a cut-of value of 0.01 as indicating significant misfit.

Fitting and testing the IRT model took place in two steps. To take advantage of the full sample, we first fitted and tested a model that excluded the MIDAS categorical items. Then we fitted and tested the MIDAS categorical items in the subsample where they had been administered, fixing the item parameters for the other items to the values they achieved in the total sample, thus producing MIDAS parameter estimates on the same scale as the other items. We did not attempt to include the original MIDAS items in an IRT-like analysis (see [35] for such an analysis and discussion).

5. Test for differential item functioning (*DIF*): The IRT model assumes that the IRT parameters pertain to all subgroups in the population, or in other words, people with the same level of headache impact should have the same probability of answering an item in a certain way, regardless of what group they belong to. If one item functions differently for different groups (DIF [36]) the item parameters for that item will differ between groups. This may happen if the item has a special meaning for some groups. It is possible to test directly for group differences in IRT item parameters [37, 38] but these tests can be cumbersome to use in standard item pool development. A practical alternative is to use logistic regression methods [39, 40]. Here, the simple sum score of the items is used as a proxy for headache impact. DIF is tested by testing for associations between each item and subgroup membership, while conditioning on the sum score. The approach can handle both differences in threshold parameters (uniform DIF) and differences in slope parameters (non-uniform DIF) [40]. In the logistic regression approach the magnitude of DIF can be quantified by a pseudo-R^2 difference measure [40] that expresses the increase in explained item variance by including the variable for group membership. No DIF implies an expected R^2-difference of zero.

We used the logistic regression approach to test for DIF against the variables gender, age, education, and employment (employed vs. not employed). For each item we performed one overall test for uniform and non-uniform DIF together. Since multiple tests were performed, we applied a double criterion for considering an item to exhibit DIF: statistical significance (p-value below 0.05) and magnitude of DIF (R^2 difference (Δ-R^2) of at least 2% using Nagelkerke R^2 [41]). Other authors have suggested higher cut-off levels [40] (meaning that we consider cases of DIF that they would dismiss as non-important), but these are based on other R^2 approaches [40]. In our experience, 2% is a relevant level.

Together, analyses 1 through 5 also examine whether the modifications made to some items (as

described above) work well. Thus, these analyses provide a revalidation of the item pool.

6. Develop the computerized adaptive testing algorithms: including procedures for item selection and IRT score estimation: We based item selection on the principle of maximum information and IRT scoring on the expected a posteriori (EAP) approach [21, 42] (see also [17]). To avoid giving some persons negative scores in the feedback, we rescaled the final IRT score so that the average headache sufferer has a score of 50 and the standard deviation is 10 for the headache population ($y = 10x + 50$, i.e. transformations were $-2 \rightarrow 30$, $-1 \rightarrow 40$, $0 \rightarrow 50$, $1 \rightarrow 60$, $2 \rightarrow 70$). We evaluated the algorithms by running simulations of a CAT on the data sets already collected. These algorithms implement the steps shown in Figure 2. The total set of responses from the persons in the study was used as an input file, but in the simulation the computer only read the responses that corresponded to the questions that would have been asked during a real CAT (this approach is sometimes called 'real simulation' to distinguish it from situations where the responses are simulated). These simulations were programmed in SAS V8.01.

7. Test whether unfortunate choice of items could introduce bias/multidimensionality: Even though analyses 1–5 have evaluated dimensionality of the total item pool (and identified a pool of unidimensional items) the possibility remains that an unfortunate choice of items could lead to a biased result – e.g. if all items chosen belonged to some subdimension of the pool, that could not be clearly identified in the overall analyses ('random multidimensionality'). To test this possibility, we performed additional sets of analyses:

1. We conducted an additional factor analysis among the 20 items that were picked in simulation studies (based on [34]).

2. We compared IRT scores from different cases of worst-case assessments to the IRT score based on all the items. The unidimensionality assumption implies that all of the short-forms should provide unbiased measures of headache impact, that is, in plots of long-form scores against short-form scores the long-form IRT scores should vary around the identity line. We used the following assessments:

- Five item CAT selecting the least informative items.
- Five item fixed assessment selecting the items that had the lowest (non-parametric) correlation with the long-form item score.
- Five item CAT selecting the most informative items within the limited subset of one of the original seven scales (seven separate analyses).

Factor analyses were performed using Mplus [26], initial analyses of option characteristic curves used Testgraf [29, 30], item parameters were estimated using Parscale [33], and for all other analyses we used SAS 8.01.

Results

1. Descriptive analyses: Table 3 shows basic descriptive information on the HIT item pool. Most items yielded high response rates, but we found high frequencies of missing responses to one HIMQ item (HIMQ14: *When you have a headache while at work (or school) how much is your ability to work reduced?*) and two MIDAS items (MIDAS1: *On how many days in the past 4 weeks did you miss work or school because of your headaches?* and MIDAS2: *How many days in the past 4 weeks was your productivity at work or school reduced by half or more because of your headaches?*). High frequencies of missing responses on these items were particularly seen among those who were not working at a paying job.

All items had a right skew and many items had a high percentage of respondents at the floor (little migraine impact). The original MIDAS items had the most pronounced skew of all items in the pool and the percentage of respondents at the floor was high for all MIDAS items (56–80%). The categorical versions of the MIDAS items were less skewed than the original version but still showed a large floor phenomenon (54–82%).

2. Test of dimensionality: The factor analysis focused initially on the 801 persons with complete scores on the categorical items from the MSQ, the HDI, the HIMQ and on the original items from the MIDAS. To handle the skewness of the MIDAS items, the responses were grouped into eight categories and analyzed as categorical responses. The eigenvalue analysis showed that approximately 60% of the total variation was explained by one component (Table 4). Although five components had eigenvalues above 1, analysis of the

Table 3. Descriptive statistics for items in the HIT item pool

Scale	Item	Percent missing	Percent floor	Skew
MSQ-role restrictive	MSQ1	0.1	33.2	0.7
	MSQ2	0.4	32.9	0.7
	MSQ3	0.2	50.6	1.1
	MSQ4	0.1	44.2	1.0
	MSQ5	0.5	32.6	0.6
	MSQ6	0.3	46.0	0.9
	MSQ7	0.6	28.4	0.6
	MSQ13 (V1)	0.5	33.7	0.7
MSQ-role Preventive	MSQ8	0.6	70.7	1.8
	MSQ9	0.4	71.8	1.9
	MSQ10	0.6	53.9	1.3
	MSQ11	0.8	72.1	2.2
MSQ-emotional function	MSQ13	0.4	62.4	1.5
	MSQ14	0.3	39.3	0.7
	MSQ15	0.4	73.4	2.0
	MSQ16	0.4	66.4	1.6
HDI-emotional function	HDI1E	0.4	66.2	1.6
	HDI2E	2.3	53.8	1.0
	HDI3E	0.3	52.4	0.9
	HDI4E	0.6	68.6	1.6
	HDI5E	1.6	56.9	1.1
	HDI6E	0.3	77.1	2.1
	HDI7E	0.4	65.6	1.5
	HDI8E	0.4	76.0	1.8
	HDI9E	0.5	79.5	2.4
	HDI10E	0.3	58.7	1.2
	HDI11E	0.4	27.5	0.1
	HDI12E	0.5	65.6	1.5
	HDI13E	0.4	38.0	0.4
HDI-role function	HDI1F	0.2	47.4	0.8
	HDI2F	0.5	53.5	1.0
	HDI3F	0.7	55.8	1.0
	HDI4F	0.3	72.6	1.9
	HDI5F	0.6	43.6	0.6
	HDI6F	0.5	71.2	1.8
	HDI7F	0.5	47.1	0.9
	HDI8F	0.6	35.9	0.4
	HDI9F	0.8	49.6	0.8
	HDI10F	0.8	66.5	1.5
	HDI11F	0.7	32.3	0.2
	HDI12F	0.7	37.8	0.5
HIMQ	HIMQ3	0.6	1.7	0.3
	HIMQ4	0.5	13.1	0.1
	HIMQ5	0.3	14.5	0.2
	HIMQ8	1.6	29.9	0.7
	HIMQ10	1.1	34.9	0.8
	HIMQ12	2.3	55.1	1.4
	HIMQ14	8.7	32.3	0.9
MIDAS	MIDAS1	6.7	79.5	11.4
	MIDAS2	9.4	73.8	6.3
	MIDAS3	4.3	55.6	6.2
	MIDAS4	6.2	63.0	8.5
	MIDAS5	2.6	72.7	2.9
MIDAS-Categorical 1	MIDAS1	8.0	81.7	1.9
	MIDAS2	8.8	71.5	1.1
	MIDAS3	2.9	57.4	1.1
	MIDAS4	2.7	55.7	1.9
	MIDAS5	2.1	72.7	2.4
MIDAS-Categorical 2	MIDAS1	9.3	77.6	1.7
	MIDAS2	10.4	66.9	1.1
	MIDAS3	2.4	53.5	1.2
	MIDAS4	2.6	55.7	1.8
	MIDAS5	2.0	68.5	8.5

stepwise increments in eigenvalues (by Scree plots) showed that the increments were substantial only for the first two components.

Based on these results, we investigated three factor models in confirmatory analyses: a one-factor, a two-factor, and a seven-factor model. The two-factor model had an emotional function factor (MSQ emotional function and HDI emotional function items) and a role function factor (all other items). This model was based on results of a previous study [17]. The seven-factor model simply corresponded to the seven original scales.

Comparison of factor loadings (Table 5) in the one- and two-factor models showed a small increase ($\approx$0.03) in loadings for the emotional functioning items and almost no increase in loadings for other items. In the two-factor model, the factor correlation was 0.92. Compared to the one-factor model, factor loadings in the seven-factor model increased by $\approx$0.05. The largest increase was seen for the MIDAS items ($\approx$0.09). In the seven-factor model, factor correlations were high between all factors – although all were significantly different from 1 (Table 6). The lowest factor correlation was between the MIDAS factor and HDI emotional functioning factor, but in general, the correlations between the MIDAS factors and the other factors were of the same magnitude as the other factor correlations.

In terms of standard fit statistics, all three factor models showed poor fit (RMSEA values were: one-factor model 0.114; two-factor model 0.107; and seven-factor model 0.078). Analysis of residual correlations revealed that some item pairs had high residual correlations in all three models.

Table 4. Eigenvalue analysis of the HIT item pool (53 items)

Component	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Eigenvalues	31.20	2.53	1.46	1.36	1.18	1.11	0.93	0.77	0.73	0.69	0.66	0.58	0.54	0.50
Difference in eigenvalues	28.66	1.07	0.10	0.18	0.07	0.19	0.16	0.04	0.04	0.04	0.08	0.04	0.04	0.02
% variation explained	58.9	4.8	2.8	2.6	2.2	2.1	1.8	1.5	1.4	1.3	1.2	1.1	1.0	1.0
Cummulative explained variance	58.9	63.7	66.4	69.0	71.2	73.3	75.1	76.5	77.9	79.2	80.4	81.5	82.5	83.5

Table 5. Comparison of item factor loadings and model fit

Questionnaire	Scale	Loadings one factor		Loadings two factors		Loadings seven factors	
		Min	Max	Min	Max	Min	Max
MSQ-V2	Role function – restrictive	0.74	0.83	0.75	0.83	0.79	0.88
	Role function – preventive	0.77	0.81	0.77	0.81	0.84	0.88
	Emotional function	0.74	0.80	0.78	0.81	0.80	0.86
HDI	Emotional function	0.65	0.84	0.68	0.88	0.69	0.88
	Role function	0.68	0.89	0.69	0.90	0.71	0.92
HIMQ		0.62	0.82	0.63	0.82	0.67	0.88
MIDAS		0.67	0.80	0.68	0.81	0.76	0.90
RMSEA		0.114		0.107		0.078	

Table 6. Factor correlations – HIT item pool (53 items) (factors based on original scales)

Questionnaire	Scale (number)		1	2	3	4	5	6	7
MSQ V2	Role function – restrictive (1)	Corr	1						
	Role function – preventive (2)	Corr	0.89	1					
	Emotional function (3)	Corr	0.89	0.91	1				
HDI	Emotional function (4)	Corr	0.81	0.76	0.86	1			
	Role function (5)	Corr	0.84	0.82	0.87	0.94	1		
HIMQ*	HIMQ (6)	Corr	0.84	0.90	0.82	0.80	0.85	1	
MIDAS	MIDAS (7)	Corr	0.86	0.85	0.80	0.75	0.78	0.85	1

SD of correlations ranged from 0.012 to 0.022.
* Seven items from HIMQ.

These item pairs are presented in Table 7, which shows the largest residual correlations for the one-factor model. There were no negative residual correlations of such magnitude. The item pairs are characterized by pairwise content similarities and by similarities in wording. A few of the item pairs were adjacent in the questionnaire.

We achieved some improvement in model fit by setting these six residual correlations free. This had very little effect on the other parameters in the model (all parameter changes were below 0.02). However, model fit statistics were still not satisfactory based on standard criteria for fit of structural equation models.

We ran additional models to test the MIDAS categorical items with the rest of the item pool (excluding the original MIDAS items). These models were run separately for the two random halves of the sample that received different versions of the MIDAS categorical items. The results mimicked the results presented above except that the version that used the *All of the time–None of the time* response choices had somewhat higher factor loadings (suggesting that the respondents

Table 7. Item wording and residual correlations (one-factor model)

Item1	Wording	Item 2	Wording	Residual correlation
HIMQ03	How would you rate the pain from your headaches?	HIMQ04	When you have headaches, how often is the pain severe?	0.303
HDI12E	My headaches make me feel confused	HDI07F	I am unable to think clearly because of my headaches	0.299
HDI02E	No one understands the effect that my headaches have on my life	HDI05E	My spouse or family and friends have no idea what I am going through because of my headaches	0.276
MIDAS3	On how many days in the *past 3 months* did you not do household work because of your headaches?	MIDAS4	How many days in the *past 3 months* was your productivity in household work reduced by half or more because of your headaches?	0.220
HIMQ12	When you have headache, how often do you miss work, school or other daily activities for all or part of the day?	MIDAS1	On how many days in the *past 3 months* did you miss work or school because of your headaches?	0.216
MSQ15	In the *past 4 weeks*, how often have you *felt* like you were a burden on others because of your headaches?	MSQ16	In the *past 4 weeks*, how often have you been *afraid* of letting others down because of your headaches?	0.205

found these answer categories more distinct). Based on the high factor correlations in multifactor models and the high factor loadings we found it justified fitting a unidimensional IRT model to the total item pool, including items from the MIDAS content domain with the response choices *All of the time–None of the time*.

3. Option characteristic curves: Evaluation of option characteristic curves indicated the need to collapse response options for several items (see Table 8 for number of categories after collapsing). For the HDI items we had tested a five-choice response format, but the results favored collapsing responses to achieve only three categories for the HDI items.

4. IRT model fitting: Table 8 summarizes the IRT results. Note that since the scaling of the IRT model generally is defined by fixing the mean and standard deviation of the population investigated, the present results cannot be directly compared to the results of IRT studies in a migraine population [17]. The slope estimates ranged from 0.92 to 3.55. The lowest slopes were seen for the HIMQ items on illness behavior (HIMQ05) and frequency of headaches (HIMQ04).

The highest slopes were seen for the HDI items on restrictions in daily (HDI01F) and in recreational activities (HDI02F). Threshold means were positive for all items except HIMQ03 *How would you rate the pain*…. This indicates that the item pool is directed primarily against the more than average headache impact. Low mean thresholds were seen for HIMQ03–HIMQ05 and for some items regarding emotional reactions to headaches: MSQ14 … *fed up*…, HDI03E …*angry*…, HDI11E …*irritable*…, HDI13E …*frustrated*…, HDI08F …*tense*…. Thus, for people with minor headache impact these items will be the most informative. High mean thresholds were seen for items on the need to cancel or miss work or school (MSQ08 and MIDAS1) or social activities (MSQ11) or the need for help in routine tasks (MSQ09). These items are most informative for people with severe headache impact.

Tests of item fit were in the acceptable range. No *p*-value was below our cut-off limit of 0.01. Worst fit was seen for MSQ16 …*afraid of letting others down*…, HDI05E *My spouse … have no idea what I am going through* …, and HIMQ12 … *miss work … for all or part of the day?* However, even for these items plots of predicted versus observed item scores looked reasonable.

Figure 3 summarizes the total information provided by the item pool and the standard error of measurement of the IRT score for a single person at various levels of the scale when all the HIT items are used. The scale is defined so that the mean of the population is zero and the standard deviation is one. The score distribution estimated during item calibration is also shown. The figure

Table 8. Summary of IRT results for the HIT item pool (53 items, n = 852)

Item	Wording	No. categories	Slope		Mean threshold		Item fit *p*-values
			Est	SE	Est	SE	
MSQ01	…interfered with…how well you dealt with family….	5	1.41	0.08	1.37	0.04	0.493
MSQ02	…interfered with…leisure time activities	5	1.32	0.08	1.19	0.04	0.106
MSQ03	…difficulty in performing work or other daily activities…	4	1.70	0.11	1.29	0.04	0.184
MSQ04	…keep you from getting as much done	5	1.56	0.10	1.60	0.04	0.923
MSQ05	…limit your ability to concentrate	5	1.62	0.09	1.19	0.03	0.856
MSQ06	…left you to tired to do work or daily activities	5	1.80	0.11	1.29	0.03	0.637
MSQ07	…limited the number of days you have felt energetic	5	1.43	0.08	1.08	0.04	0.808
MSQ08	…had to cancel work or daily activities…	4	1.90	0.17	2.27	0.08	0.704
MSQ09	…need help in handling routine tasks…	4	2.07	0.18	2.15	0.07	0.223
MSQ10	…stop work or other activities…	4	1.54	0.10	1.21	0.04	0.754
MSQ11	…not able to go to social activities…	4	2.05	0.18	2.07	0.07	0.341
MSQ13	…felt you should avoid social or family activities…	5	1.52	0.10	1.64	0.04	0.046
MSQ14	…felt fed up or frustrated…	5	1.12	0.07	0.84	0.04	0.397
MSQ15	…felt like you were a burden on others…	3	1.71	0.13	1.56	0.06	0.811
MSQ16	…afraid of letting others down…	3	1.85	0.15	1.69	0.07	0.017
MSQ13V1	… limited the number of days you have felt full of pep	5	1.14	0.07	1.57	0.05	0.146
HDI01E	…I feel handicapped	3	2.30	0.18	1.45	0.05	0.069
HDI02E	No one understands the effect …on my life	3	1.48	0.10	1.14	0.05	0.156
HDI03E	My headaches make me angry	3	1.63	0.11	0.96	0.05	0.675
HDI04E	Sometimes I feel that I am going to lose control…	3	1.98	0.14	1.42	0.05	0.162
HDI05E	My spouse…have no idea what I am going through…	3	1.32	0.09	1.21	0.06	0.027
HDI06E	…so bad that I feel I am going to go insane	3	2.29	0.18	1.62	0.05	0.768
HDI07E	My outlook on the world is affected by…	3	1.76	0.12	1.43	0.05	0.396
HDI08E	I am afraid to go outside…	3	2.12	0.16	1.40	0.05	0.812
HDI09E	I feel desperate because of…	3	2.52	0.20	1.74	0.05	0.845
HDI10E	… place stress on my relationships…	3	2.29	0.15	1.19	0.04	0.872
HDI11E	I feel irritable because of…	3	1.61	0.09	0.23	0.04	0.294
HDI12E	… make me feel confused	3	1.10	0.09	1.49	0.08	0.220
HDI13E	… make me feel frustrated	3	2.21	0.13	0.52	0.03	0.803
HDI01F	…I feel restricted in performing…daily activities	3	3.55	0.21	0.82	0.03	0.079
HDI02F	I restrict my recreational activities…	3	2.65	0.16	1.01	0.03	0.432
HDI03F	…I am less likely to socialize	3	2.51	0.16	1.06	0.04	0.891
HDI04F	… concerned that I am paying penalties …	3	2.65	0.19	1.42	0.04	0.159
HDI05F	I avoid being around people when I have a headache	3	2.02	0.12	0.70	0.04	0.172
HDI06F	…difficult for me to achieve my goals in life	3	2.65	0.19	1.41	0.04	0.800
HDI07F	I am unable to think clearly…	3	1.58	0.10	1.08	0.05	0.117
HDI08F	I get tense…	3	1.83	0.11	0.52	0.04	0.043
HDI09F	I do not enjoy social gatherings…	3	2.14	0.13	0.99	0.04	0.067
HDI10F	I avoid travelling…	3	2.28	0.16	1.32	0.04	0.080
HDI11F	I find it difficult to read…	3	1.33	0.08	0.39	0.05	0.629
HDI12F	…difficult to focus my attention away from…	3	1.92	0.11	0.65	0.04	0.029
HIMQ03	How would you rate the pain…	5	1.11	0.06	−0.27	0.05	0.220
HIMQ04	…how often is the pain severe?	5	0.94	0.05	0.35	0.04	0.054
HIMQ05	… how often do you lie down and rest?	4	0.92	0.05	0.18	0.05	0.579
HIMQ08	…ability to perform … housework … reduced?	5	1.63	0.09	0.99	0.03	0.097
HIMQ10	…ability to engage in non-work activities reduced?	5	1.71	0.10	1.15	0.03	0.097
HIMQ12	…miss work…for all or part of the day?	4	1.10	0.07	1.29	0.06	0.032
HIMQ14	…at work or school, …ability to work reduced	5	1.60	0.09	1.11	0.03	0.601
MIDAS1*	… did you miss work or school …	3	1.31	0.16	2.16	0.16	0.590
MIDAS2*	… productivity … reduced by half or more …	3	1.49	0.16	1.69	0.10	0.091
MIDAS3*	… did you not do household work …	4	2.02	0.17	1.15	0.05	0.563
MIDAS4*	... household work reduced by half or more …	4	2.12	0.18	1.18	0.05	0.351
MIDAS5*	… miss family, social or leisure activities …	3	2.47	0.26	1.63	0.07	0.979

* Estimation and fit tests based on 429 persons.

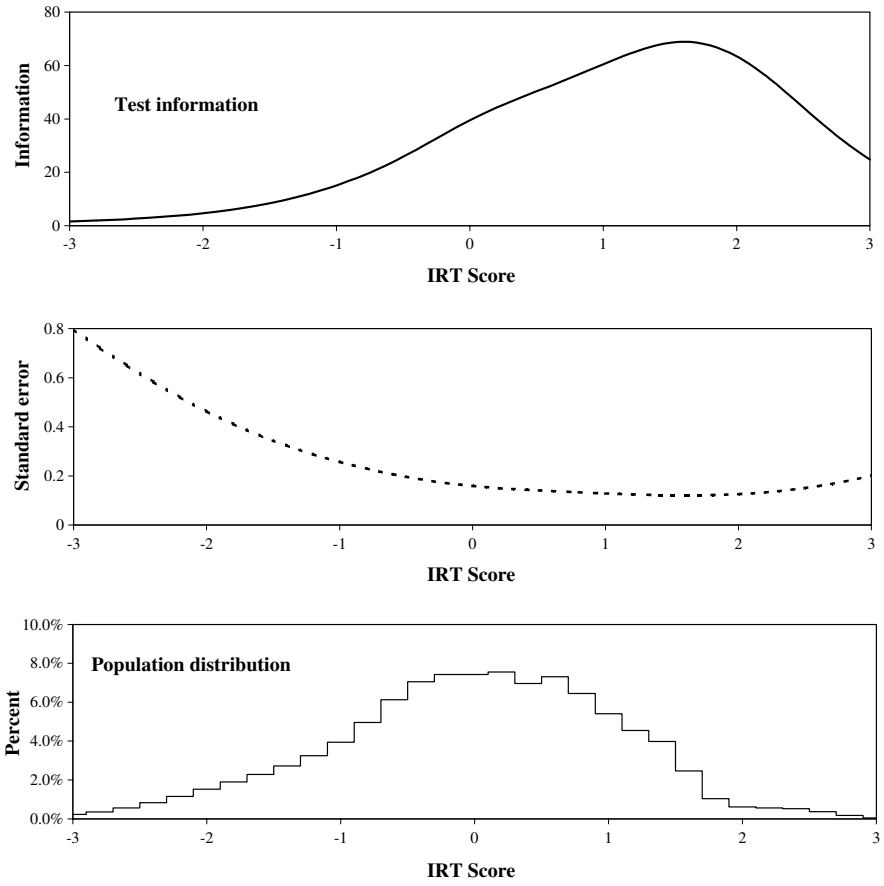

Figure 3. Information function and standard error of measurement for the HIT item pool compared with the population distribution of headache impact.

shows that the item pool provides most information two standard deviations above the mean for headache sufferers, where we find relatively few people. At the range of −0.5 to 3.0 the standard error of measurement is below 0.2, which means that the 95% confidence interval will be below ±0.4 for any person with an IRT-score in that range that answers all items. The measurement precision is less for persons with little headache impact.

5. *DIF analyses*: Although several items had significant DIF (Table 9), no items exceeded the R^2 cutoff limit of a difference of 2%. In Table 9, we have bolded test results with an R^2 difference at 1% or more, these might be regarded as borderline DIF. There were no major signs of DIF against gender; two items, MSQ15 and HIMQ14, showed borderline DIF against age; one item, MSQ01, showed borderline DIF against education; while two items, MSQ08 and HIMQ12, showed borderline DIF against employment. For the MIDAS categorical items we only had data from half the sample. We saw no significant DIF tests, and no R^2 difference above 2% (although a few MIDAS items had R^2 differences above 1%).

6. *CAT simulations*: We selected HIMQ04 as the initial item for our adaptive headache impact test. Although this item has a low slope it covers a wide range of headache impact (lowest threshold −1.6 highest threshold 2.0), which makes it a good candidate for the first item. Further, the wording '*When you have headaches*, *how often is the pain severe*?' seems appropriate for a first item. For the initial evaluations of the adaptive test, we allowed the computer to choose freely among the remaining items and implemented a fixed stopping rule of five

Table 9. Test of DIF in the HIT item pool (n = 829)

Item	Wording	Gender		Age		Education		Employment	
		Δ-R^2	p	Δ-R^2	p	Δ-R^2	p	Δ-R^2	p
MSQ01	…interfered with…how well you dealt with family….	0.004	0.040	0.004	0.069	0.013	0.000	0.001	0.590
MSQ02	…interfered with…leisure time activities	0.001	0.671	0.000	0.877	0.003	0.107	0.005	0.020
MSQ03	…difficulty in performing work or other daily activities…	0.001	0.463	0.000	0.745	0.009	0.002	0.000	0.711
MSQ04	…keep you from getting as much done	0.001	0.336	0.002	0.328	0.008	0.003	0.001	0.473
MSQ05	…limit your ability to concentrate	0.002	0.241	0.003	0.126	0.001	0.532	0.000	0.687
MSQ06	…left you to tired to do work or daily activities	0.004	0.029	0.001	0.630	0.001	0.403	0.005	0.019
MSQ07	…limited the number of days you have felt energetic	0.002	0.186	0.001	0.444	0.002	0.306	0.006	0.012
MSQ08	…had to cancel work or daily activities…	0.004	0.237	0.001	0.731	0.009	0.049	**0.016**	**0.004**
MSQ09	…need help in handling routine tasks…	0.000	0.859	0.002	0.508	0.001	0.638	0.005	0.177
MSQ10	…stop work or other activities…	0.000	0.905	0.001	0.532	0.000	0.943	0.003	0.133
MSQ11	…not able to go to social activities…	0.002	0.533	0.002	0.620	0.006	0.162	0.004	0.261
MSQ13	…felt you should avoid social or family activities…	0.001	0.460	0.008	0.010	0.001	0.390	0.004	0.069
MSQ14	…felt fed up or frustrated…	0.000	0.767	0.000	0.831	0.001	0.502	0.007	0.006
MSQ15	…felt like you were a burden on others…	0.000	0.898	**0.017**	**0.001**	0.005	0.088	0.000	0.974
MSQ16	…afraid of letting others down…	0.002	0.451	0.001	0.705	0.000	0.988	0.003	0.342
MSQ13V1	… limited the number of days you have felt full of pep	0.002	0.231	0.004	0.083	0.002	0.362	0.001	0.499
HDI01E	…I feel handicapped	0.002	0.451	0.007	0.050	0.003	0.274	0.000	0.835
HDI02E	No one understands the effect …on my life	0.000	0.882	0.003	0.271	0.005	0.076	0.002	0.385
HDI03E	My headaches make me angry	0.004	0.088	0.006	0.042	0.003	0.183	0.005	0.061
HDI04E	Sometimes I feel that I am going to lose control…	0.000	0.817	0.001	0.645	0.005	0.075	0.004	0.167
HDI05E	My spouse…have no idea what I am going through…	0.000	0.875	0.003	0.231	0.003	0.208	0.003	0.191
HDI06E	…so bad that I feel I am going to go insane	0.001	0.620	0.002	0.435	0.010	0.013	0.001	0.535
HDI07E	My outlook on the world is affected by…	0.000	0.876	0.001	0.756	0.006	0.056	0.005	0.080
HDI08E	I am afraid to go outside…	0.000	0.928	0.001	0.564	0.001	0.697	0.005	0.108
HDI09E	I feel desperate because of…	0.004	0.207	0.002	0.439	0.002	0.425	0.002	0.492
HDI10E	… place stress on my relationships…	0.002	0.370	0.002	0.397	0.001	0.578	0.001	0.605
HDI11E	I feel irritable because of…	0.006	0.026	0.000	0.898	0.001	0.553	0.004	0.092
HDI12E	… make me feel confused	0.003	0.292	0.007	0.058	0.008	0.024	0.003	0.250
HDI13E	… make me feel frustrated	0.000	0.945	0.002	0.241	0.002	0.235	0.007	0.007
HDI01F	…I feel restricted in performing…daily activities	0.002	0.262	0.003	0.109	0.001	0.430	0.001	0.471
HDI02F	I restrict my recreational activities…	0.001	0.614	0.008	0.009	0.002	0.190	0.000	0.811
HDI03F	…I am less likely to socialize	0.005	0.041	0.001	0.430	0.005	0.039	0.002	0.310
HDI04F	… concerned that I am paying penalties …	0.002	0.272	0.002	0.371	0.000	0.991	0.002	0.373
HDI05F	I avoid being around people when I have a headache	0.006	0.019	0.001	0.416	0.003	0.114	0.000	0.931
HDI06F	…difficult for me to achieve my goals in life	0.002	0.284	0.003	0.252	0.001	0.557	0.001	0.619
HDI07F	I am unable to think clearly…	0.000	0.916	0.008	0.018	0.002	0.292	0.008	0.012
HDI08F	I get tense…	0.000	0.928	0.001	0.730	0.001	0.496	0.000	0.993
HDI09F	I do not enjoy social gatherings…	0.002	0.204	0.002	0.341	0.000	0.763	0.001	0.605
HDI10F	I avoid travelling…	0.005	0.075	0.001	0.512	0.002	0.254	0.001	0.611
HDI11F	I find it difficult to read…	0.005	0.079	0.006	0.052	0.002	0.306	0.004	0.117

HDI12F	…difficult to focus my attention away from…	0.002	0.238	0.001	0.640	0.000	0.901	0.003	0.149
HIMQ03	How would you rate the pain…	0.008	0.005	0.001	0.596	0.001	0.578	0.000	0.771
HIMQ04	…how often is the pain severe?	0.002	0.244	0.005	0.044	0.003	0.108	0.003	0.159
HIMQ05	… how often do you lie down and rest?	0.009	0.007	0.001	0.781	0.002	0.312	0.001	0.728
HIMQ08	…ability to perform … housework … reduced?	0.001	0.284	0.003	0.081	0.004	0.039	0.001	0.578
HIMQ10	…ability to engage in non-work activities reduced?	0.000	0.860	0.003	0.094	0.003	0.068	0.000	0.661
HIMQ12	…miss work…for all or part of the day?	0.001	0.450	0.002	0.419	0.004	0.088	**0.010**	**0.002**
HIMQ14	…at work or school, …ability to work reduced	0.004	0.046	**0.010**	**0.001**	0.008	0.001	0.003	0.061
MIDAS1*	…… did you miss work or school …	0.000	0.936	0.005	0.396	0.013	0.094	0.012	0.110
MIDAS2*	… productivity … reduced by half or more …	0.002	0.644	0.013	0.062	0.001	0.759	0.004	0.421
MIDAS3*	… did you not do household work …	0.003	0.345	0.001	0.664	0.001	0.822	0.003	0.409
MIDAS4*	... household work reduced by half or more …	0.006	0.120	0.005	0.232	0.002	0.431	0.003	0.405
MIDAS5*	… miss family, social or leisure activities …	0.002	0.569	0.004	0.380	0.001	0.685	0.002	0.630

Bold values: Δ-R^2 at least 0.01 and p at most 0.05.

*Test based on 421 persons.

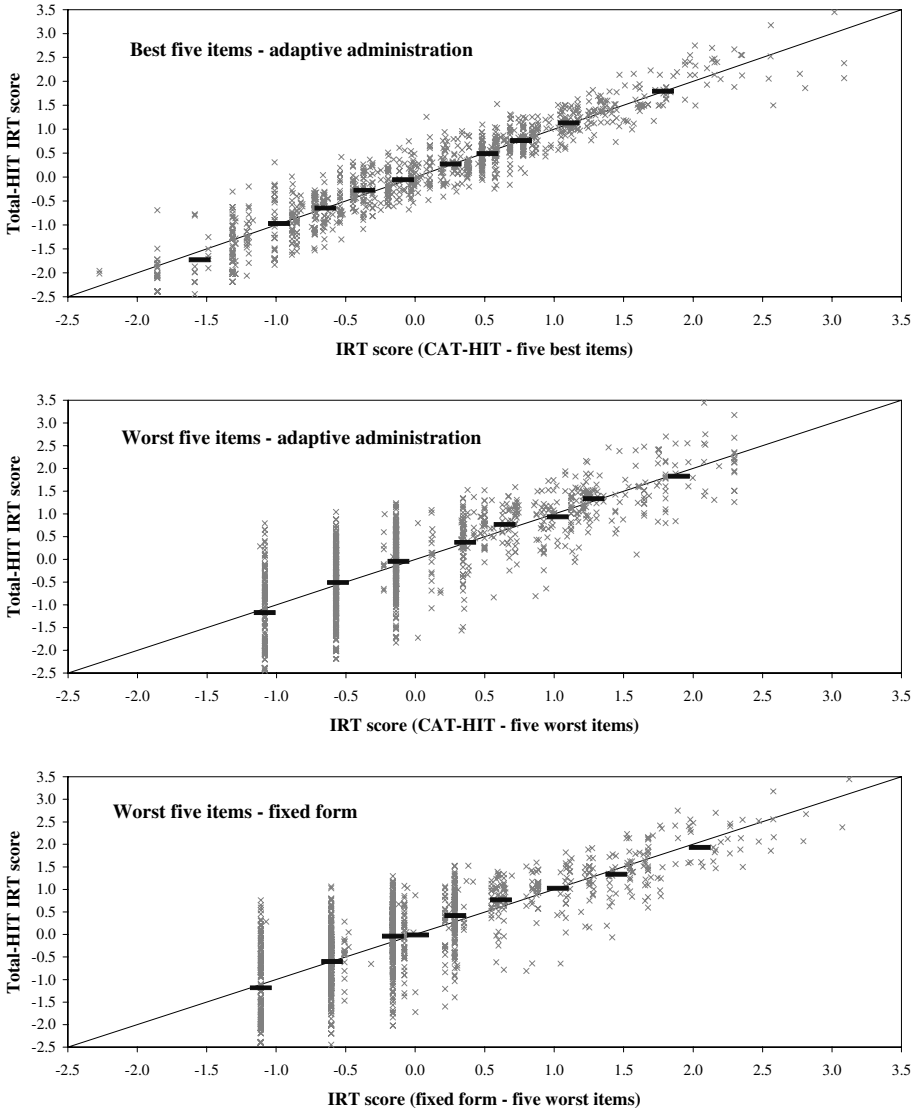

Figure 4. Total item pool IRT scores (total-HIT) against scores from three different five-item tests. Gray x: observations. Black horizontal lines: means.

items. This testing logic was evaluated in simulations of a CAT using the data from the National Survey of Headache Impact. Figure 4 (top part) illustrates the concordance between the results from the five-item adaptive test (CAT-HIT-5) and the results based on the full HIT item pool (total-HIT). The graph and the correlation between the two measures (0.92) shows that the two measures are in very good agreement, although the agreement at the low end of the CAT-HIT-5 scale is less than for the rest of the scale. We grouped the CAT-HIT-5 scores into 10 groups and compared the mean CAT-HIT-5 score and the mean total-HIT score for each group. The results indicated no significant departures from the identity line.

7. *Test of* '*random multidimensionality*': Factor analyses of a reduced set of 20 items (see [34]) basically repeated the results from the analyses of the total item pool. The first factor explained 67% of the item variation, had factor loadings comparable to the loadings from the full item pool, and the RMSEA was 0.13.

In the bottom part of Figure 4 we present results regarding the association between the IRT score based on all items and the IRT score from different kinds of worst-case item selection scenarios. Compared to the best-case scenario (upper graph), the worst case scenarios are characterized by crude measurement, floor effects, and lack of precision (as indicated by the variance of the total-HIT scores around the identity line). However, the long-form scores are spread symmetrically around the identity line, without any sign of bias. We saw the same results when we performed analyses on short-form assessments, resulting from selecting five items from limited item pools representing each of the seven original scales: the mean short-form scores agreed highly with the mean long-form scores, although the short-form measures were more crude (data not shown).

Discussion

The results of this paper are relevant to two main topics: (1) How many dimensions (scales) are necessary for assessment of headache impact? (2) Are IRT-based estimates and adaptive testing of headache impact feasible? We conclude from our results that one dimension is sufficient for the assessment of headache impact and that an IRT-based CAT is feasible.

Our conclusion on dimensionality is to some degree in contrast to the current status of the field. From a content point of view the HIT item pool includes questions on pain, role functioning, social functioning, fatigue, cognition, and mental health. Previous questionnaires that include a broad spectrum of these concepts (HDI and MSQ) are scored as two or three scales, while the questionnaires that are scored as only one scale (HIMQ and MIDAS) do not include items on fatigue, cognition, and mental distress. Our own previous factor analytic studies of the MSQ suggested that a one-factor model was adequate [17]. However, some factor analytic studies have supported at least two factors for the MSQ [43]. Our IRT-based analyses of the MSQ found good fit to a unidimensional model for most items but also some indications of a second dimension reflecting mental health [17].

Our present analyses of eigenvalues suggested that at most two factors were needed. Further, we found very high correlations between the factors hypothesized by the original test constructers, both within and between the original questionnaires. As discussed in [17], the difference in results might partially be due to different analytic techniques. We think that the high factor correlations and the agreement between various short-form scores and the long-form scores justify us in assessing headache impact using only one score. Further support for this approach is found in the fact that the MSQ subscales all showed the same trends in analyses of responsiveness [22]. Thus, we are confident that a single score adequately represents headache impact.

It is not a prerequisite for IRT-based adaptive testing that the item pool is unidimensional; multidimensionality can be handled in several different ways. If the dimensions are fairly independent, the most logical solution is to develop a separate item pool for each dimension and report results separately. Alternatively, both dimension-specific scores and an overall score could be reported. If the dimensions are highly correlated but judged to capture different content domains, item selection rules can be specified to make sure that assessment is never performed using only items from one content domain. Thus, only a single score is provided, but it is based on a balanced item content. We are currently exploring this assessment strategy for the HIT.

We find that the factor analytic results regarding the MIDAS questionnaire are particularly noteworthy. The original MIDAS items differ from the rest of the items in response format (number of days are reported) and recall period (three months) and MIDAS does not include items on fatigue, cognition, nor mental distress. Never the less, the MIDAS factor had high correlation with all the other factors and the MIDAS items had high loadings on a common factor. We interpret these results as a fairly strong indication that the construct measured by MIDAS is well represented in the HIT item pool. The primary differences between MIDAS and HIT is the response format, the scale scoring, and the level of impact the scales are targeted for. In our view the MIDAS has the advantage of being simple and achieving a score (total number of days with one

of five components of headache impact within the last three months) that may be readily understood (although a score greater than 90 can be achieved if several components of impact are encountered simultaneously). On the other hand, by focusing on role and social functioning and setting a high threshold before impact is recorded, the MIDAS is not responsive to change in headache impact for people with minor impact. In contrast, the CAT simulations illustrate that persons with minor headache impact can use the HIT without floor problems.

The high percentages of missing responses on the first two MIDAS items (on missing or having reduced productivity at work or at school) for the non-employed suggest that these items appear less relevant for this group (probably for the same reason, two other work-related items, MSQ08 and HIMQ12, showed borderline DIF for employment status). Further, although the MIDAS in theory enables respondents to make a very fine-graded report on total headache impact (from 0 to 90 days), results from cognitive psychology indicate that respondents are not really able to recall information at that level of precision on any single item [44]. We note that the developers of MIDAS recommend collapsing scores into four categories to provide an interpretation of the severity of migraine [14, 45]. Thus, the less sharply defined response options in the HIT item pool may be more in line with how respondents actually think. For this reason we included a categorical version of the MIDAS items in the pool. We also changed the recall period to 30 days. The question of recall period is complex. In this cross-sectional analysis, the original MIDAS items (with a three month recall period) had a high loading on the common factors once we dealt with the skewness issue. However, this result may not generalize to a study involving short-term changes, such as an intervention study. Here, we anticipate that the MIDAS would respond more slowly to changes in headache impact because of the long recall period. For this reason we prefer to use the same recall period for all the items in our item pool. It could be argued that a recall period of 30 days would be to short and introduce random noise when assessing the impact of an episodic disease like headache. However, since the use of a longer recall period introduces the risk of error due to cognitive problems (e.g. forgetting rate and telescoping [46]) we have chosen to stick with a 30-day recall period.

Our factor analysis model had a poor overall fit in terms of the RMSEA statistic. In the analysis of continuous data a standard rule of thumb is that an RMSEA below 0.05 indicates good overall fit. However, there is still limited experience with the distribution of the fit statistics when analyzing categorical data (this applies to the RMSEA as well as to the χ^2 statistic). For this reason we refrained from extensive model revisions to improve the fit. Although the model fit was not optimal, we think that the magnitude of the factor loadings and factor correlations is still interpretable. We regard the factor analysis as a step towards the final model, which is specified as an IRT model.

The advantages and disadvantages of maintaining links to the original items from widely-used instruments are apparent in our results. The advantage is that all items that fit the model can be calibrated on a common metric. Accordingly, published results can be compared and improved score estimates based on the new model can be compared with results previously published. We published a conversion table for the HDI, HIMQ, MIDAS and MSQ, in relation to HIT [23] and a more thorough analysis is the subject of a companion paper here [35]. The disadvantage is that needed improvements in some items have been delayed. The need for these improvements is evidenced in the large proportion (about half) of items required collapsing of one or more response categories in order to fit the IRT model. Such improvements in the wording of items and response choices should be evaluated in future studies.

Our interpretation of the results concerning the fit of the IRT model is that the model fits well. However, only little research has been performed on the fit statistic for polytomous items [31] and more experience from simulation studies would be helpful. Glas [47] has developed fit indices for polytomous and dichotomous IRT models based on the Lagrange multiplier test. These fit indices are well founded in theory but they require the first- and second-order derivatives of the log likelihood and therefore cannot be computed from the standard output of IRT software.

Also, we see the DIF results as fairly encouraging. Some clinicians have been concerned that

the questions mentioning the word *work* might be seen as irrelevant by people who are not employed. Therefore their answers to these items might not be informative about their headache impact. Although some tendencies are seen in this direction (results for MSQ08 *…had to cancel work or daily activities…* and HIMQ12 *…miss work…for all or part of the day?*) the concern is not supported for the majority of items that contains the word *work*. In educational research, judgments about DIF are based on both statistical tests and expert judgment [48]. For the HIT item pool, we have withdrawn two items (MIDAS1 and MIDAS2) from use in the online HIT test (see [34]), because of criticisms that these items were not appropriate for the non-employed.

By building on previous questionnaires, we have established an item pool that is clinically relevant but shares some of the weaknesses of the original instruments. As illustrated in Figure 3, the item pool provides most information for people with more than average headache impact.

Hambleton et al. suggest that increasing test information beyond 25 leads to comparatively little reduction in standard error of measurement [49]. Based on this criteria, the item pool has more than sufficient information from −0.5 and upward (≈69% of the population). However, the pool provides little information for people with minor headache impact. Thus, the overall precision of the item pool can be improved by developing items aimed at minor headache impact. One potential source of such items is the MSQOL questionnaire [8, 9]. The IRT methodology allows for later inclusion of such items without affecting the metric of the IRT scale. We are currently conducting studies to include additional items suggested by clinical experts.

Finally, by simulating computerized adaptive health test we showed that an adaptive test of only five items had a very high concordance with the score based on all items. We emphasize that a fixed five-item stopping rule is but one of the possible stopping rules. By specifying a stopping rule based on test precision, we could have achieved even higher concordance with the score on the total item pool – at the expense of more items administered. Such stopping rules can easily be modified depending on the purpose of the assessment, without compromising comparability across assessments.

We conclude that we have established a suitable item pool to be used as a basis for a computerized adaptive test of headache impact [23].

Acknowledgements

This work was supported by grants from GlaxoSmithKline. We thank two reviewers for valuable suggestions and comments to a previous version of the paper.

References

1. Breslau N, Rasmussen BK. The impact of migraine: Epidemiology, risk factors, and co-morbidities. Neurology 2001; 56: S4–S12.
2. Schwartz BS, Stewart WF, Simon D, Lipton RB. Epidemiology of tension-type headache. JAMA 1998; 279: 381–383.
3. Stewart WF, Shechter A, Lipton RB. Migraine heterogeneity. Disability, pain intensity, and attack frequency and duration. Neurology 1994; 44: S24–S39.
4. The Subcutaneous Sumatriptan International Study Group. Treatment of migraine attacks with sumatriptan. N Engl J Med 1991; 325: 316–321.
5. Lipton RB, Stewart WF, Goadsby PJ. Headache-related disability in the management of migraine. Neurology 2001; 56: S1–S3.
6. Goadsby PJ, Lipton RB, Ferrari MD. Migraine – current understanding and treatment. N Engl J Med 2002; 346: 257–270.
7. Martin BC, Pathak DS, Sharfman MI, et al. Validity and reliability of the migraine-specific quality of life questionnaire (MSQ Version 2.1). Headache 2000; 40: 204–215.
8. Wagner TH, Patrick DL, Galer BS, Berzon RA. A new instrument to assess the long-term quality of life effects from migraine: Development and psychometric testing of the MSQOL. Headache 1996; 36: 484–492.
9. Patrick DL, Hurst BC, Hughes J. Further development and testing of the migraine-specific quality of life (MSQOL) measure. Headache 2000; 40: 550–560.
10. Jacobson GP, Ramadan NM, Aggarwal SK, Newman CW. The Henry Ford Hospital Headache Disability Inventory (HDI). Neurology 1994; 44: 837–842.
11. Jacobson GP, Ramadan NM, Norris L, Newman CW. Headache disability inventory (HDI): Short-term test–retest reliability and spouse perceptions. Headache 1995; 35: 534–539.
12. Stewart WF, Lipton RB, Simon D, Liberman J, Von Korff M. Validity of an illness severity measure for headache in a population sample of migraine sufferers. Pain 1999; 79: 291–301.
13. Stewart WF, Lipton RB, Dowson AJ, Sawyer J. Development and testing of the Migraine Disability Assessment

(MIDAS) Questionnaire to assess headache-related disability. Neurology 2001; 56: S20–S28.
14. Stewart WF, Lipton RB, Kolodner K, Liberman J, Sawyer J. Reliability of the migraine disability assessment score in a population-based sample of headache sufferers. Cephalalgia 1999; 19: 107–114.
15. Stewart WF, Lipton RB, Kolodner KB, Sawyer J, Lee C, Liberman JN. Validity of the Migraine Disability Assessment (MIDAS) score in comparison to a diary-based measure in a population sample of migraine sufferers. Pain 2000; 88: 41–52.
16. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? Quality Life Res 1995; 4: 293–307.
17. Bjorner JB, Kosinski M, Ware JE Jr. The feasibility of applying item response theory to measures of migraine impact: A re-analysis of three clinical studies. Quality Life Res 2003; 12: 887–902.
18. Muraki E. Information functions of the generalized partial credit model. Appl Psychol Measur 1993; 17: 351–363.
19. Wainer H, Dorans NJ, Eignor D, et al. Computerized Adaptive Testing: A Primer. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
20. Wainer H, Mislevy RJ. Item Response Theory, Item Calibration, and Proficiency Estimation. In: Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ, Steinberg L, Thissen D (eds), Computerized Adaptive Testing: A Primer, Hillsdale, NJ: Lawrence Erlbaum Associates, 2000; 61–101.
21. Thissen D, Orlando M. Item Response Theory for items scored in two categories. In: Thissen D, Wainer H (eds), Test Scoring. Mahwah: Lawrence Earlbaum, 2001; 73–140.
22. Kosinski M, Bjorner JB, Ware JE Jr, Batenhorst A, Cady RK. The responsiveness of headache impact scales scored using 'classical' and 'modern' psychometric methods: A re-analysis of three clinical trials. Quality Life Res 2003; 12: 903–912.
23. Ware JE Jr, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. Med Care 2000; 38: II73–II82.
24. Stewart AL, Ware JE Jr. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. London: Duke University Press, 1992.
25. Stewart WF, Lipton RB, Celentano DD, Reed ML. Prevalence of migraine headache in the United States. Relation to age, income, race, and other sociodemographic factors. JAMA 1992; 267: 64–69.
26. Muthen BO, Muthen L. Mplus User's Guide. Los Angeles: Muthén & Muthén, 2001.
27. Bollen KA, Barb KH. Pearson's r and coarsely categorized measures. Am Sociol Rev 1981; 46: 232–239.
28. Nunnally JC, Bernstein IH. Psychometric Theory. New York: McGraw-Hill, Inc., 1994.
29. Ramsay JO. Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika 1991; 56: 611–630.
30. Ramsay JO. TestGraf – A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data. Montreal: McGill University, 1995.
31. Muraki E. A Generalized Partial Credit Model. In: van der Linden WJ, Hambleton RK (eds), Handbook of Modern Item Response Theory. Berlin: Springer, 1997: 153–164.
32. Bock RD, Aitkin M. marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 1981; 46: 443–459.
33. Muraki E, Bock RD. Parscale – IRT based Test Scoring and Item Analysis for Graded Open-ended Exercises and Performance Tasks. Chicago: Scientific Software Inc., 1996.
34. Ware JE Jr, Kosinski M, Bjorner JB, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Quality Life Res 2003; 12: 935–952.
35. Bjorner JB, Kosinski M, Ware JE Jr. Using item response theory to calibrate the Headache Impact Test (HIT^{TM}) to the metric of traditional headache scales. Quality Life Res 2003; 12: 981–1002.
36. Holland PW, Wainer H. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 1993.
37. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holland PW, Wainer H (eds), Differential Item Functioning. Hillsdale NJ: Lawrence Erlbaum Ass, 1993: 67–113.
38. Muraki E. Stepwise analysis of differential item functioning based on multiple-group partial credit model. Educ Meas 1999; 36: 217–232.
39. Swaminathan H, Rogers JH. Detecting differential item functioning using logistic regression procedures. J Educ Measur 1990; 27: 361–370.
40. Zumbo BD. A handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense, 1999.
41. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika 1991; 78: 691–692.
42. Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 1982; 6: 431–444.
43. Jhingran P, Davis SM, LaVange LM, Miller DW, Helms RW. MSQ: Migraine-Specific Quality-of-Life Questionnaire. Further investigation of the factor structure. Pharmacoeconomics 1998; 13: 707–717.
44. Bradburn NM, Rips LJ, Shevell SK. Answering autobiographical questions: The impact of memory and inference on surveys. Science 1987; 236: 157–161.
45. Lipton RB, Goadsby PJ, Sawyer JPC, Blakeborough P, Stewart WF. Migraine: Diagnosis and assessment of disability. Rev Contemp Pharmacotherapy 2000; 11: 63–73.
46. Jobe JB, Tourangeau R, Smith AF. Contributions of survey research to the understanding of memory. Cognitive Psychol 1993; 7: 567–584.
47. Glas CAW. Modifications indices for the 2-PL and the Nominal Response Model. Psychometrika 1999; 64: 273–294.

48. Zieky M. Practical questions in the use of DIF statistics in test development. In: Holland PW, Wainer H (eds), Differential Item Functioning. Hillsdale (NJ): Lawrence Erlbaum Associates, 1993: 337–347.
49. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. London: Sage Publications, 1991.

Address for correspondence: Jakob Bue Bjorner, QualityMetric Incorporated, 640 George Washington Highway, Lincoln, RI 02895, USA
Phone: +1-401-334-8800 ext. 271; Fax: +1-401-334-8801
E-mail: jbjorner@qualitymetric.com