Article:
Petersen, MA, Aaronson, NK, Chie, WC et al. (9 more authors) (2016) Development of an item bank for computerized adaptive test (CAT) measurement of pain. Quality of Life Research, 25 (1). pp. 1-11. ISSN 0962-9343

https://doi.org/10.1007/s11136-015-1069-5

European Journal of Pain

Development of an item bank for computerized adaptive test (CAT) measurement of pain

--Manuscript Draft--

Manuscript Number:	
Article Type:	Original Manuscript
Corresponding Author:	Morten Aagaard Petersen Copenhagen, DENMARK
First Author:	Morten Aagaard Petersen
Order of Authors:	Morten Aagaard Petersen
	Neil K. Aaronson
	Wei-Chu Chie
	Thierry Conroy
	Anna Costantini
	Eva Hammerlid
	Marianne J. Hjermstad
	Stein Kaasa
	Jon H. Loge
	Galina Velikova
	Teresa Young
	Mogens Groenvold
Abstract:	Background: Patient-reported outcomes should ideally be adapted to the individual patient while maintaining comparability of scores across patients. This is achievable using computerized adaptive testing (CAT). The aim here was to develop an item bank for CAT measurement of the pain domain as measured by the EORTC QLQ-C30 questionnaire. Methods: The development process consisted of four steps. 1) Literature search. 2) Formulation of new items. 3) Pre-testing. 4) Field-testing and psychometric analyses for the final selection of items. Results: In step 1) we identified 337 pain items from the literature. 2) Twenty-nine new items fitting the QLQ-C30 item style were formulated. Expert evaluations reduced this to 26 items. 3) Based on patient interviews (N=31) the list was further reduced to 21 items. 4) We obtained responses from 1,103 cancer patients from five countries. Psychometric evaluations showed that 16 items could be retained in a unidimensional item bank. Evaluations indicated that use of the CAT measure may reduce sample size requirements with 15-25% compared to using the QLQ-C30 pain scale. Conclusions: We have established an item bank of 16 items suitable for CAT measurement of pain. We recommend initiating CAT measurement by screening for pain using the two original QLQ-C30 pain items.
Suggested Reviewers:	Dennis Revicki Dennis.Revicki@unitedbiosource.com
	Mathias Rose rose@charite.de
	Bryce Reeve reeveb@mail.nih.gov
	Dagmar Amtmann dagmara@u.washington.edu

2014-01-31

Dept. of Palliative Medicine

+45 3531 2025
+45 3531 2071
e-mail • Mpet0009@bbh.regionh.dk

European Journal of Pain

Submission of article manuscript

We would like to submit an original article manuscript for possible publication in the European Journal of Pain. The manuscript has not been published or submitted for publication elsewhere.

Manuscript title:
Development of an item bank for computerized adaptive test (CAT) measurement of pain

Short summary of the study:
An item bank of 16 items allowing for CAT measurement of pain was developed. This is expected to clearly improve EORTC measurement of pain.

Author contributions:
Morten Aa. Petersen: conception and design; analysis and interpretation of data; drafting the article; final approval of the version to be published.
Neil K. Aaronson: conception and design; interpretation of data; revising the article critically for important intellectual content; final approval of the version to be published.
Wei-Chu Chie: acquisition of data; revising the article critically for important intellectual content; final approval of the version to be published.
Thierry Conroy: acquisition of data; revising the article critically for important intellectual content; final approval of the version to be published.
Anna Costantini: acquisition of data; revising the article critically for important intellectual content; final approval of the version to be published.
Eva Hammerlid: acquisition of data; revising the article critically for important intellectual content; final approval of the version to be published.

Marianne J. Hjermstad: Acquisition of data (literature review); revising the article critically for important intellectual content; final approval of the version to be published.
Stein Kaasa: Acquisition of data (literature review); revising the article critically for important intellectual content; final approval of the version to be published.
Jon H. Loge: Acquisition of data (literature review); revising the article critically for important intellectual content; final approval of the version to be published.
Galina Velikova: conception and design; acquisition of data; revising the article critically for important intellectual content; final approval of the version to be published.
Teresa Young: acquisition of data; revising the article critically for important intellectual content; final approval of the version to be published.
Mogens Groenvold: conception and design; interpretation of data; revising the article critically for important intellectual content; final approval of the version to be published.

As possible reviewers for this manuscript we suggest:

- Dennis A. Revicki: Center for Health Outcomes Research, United BioSource Corporation, 7101 Wisconsin Ave., Suite 600, Bethesda, MD 20814, USA. Dennis.Revicki@unitedbiosource.com
- Mathias Rose: Psychosomatische Medizin und Psychotherapie, Charité-Universitätsmedizin Berlin, Berlin, Germany. rose@charite.de
- Bryce B. Reeve: Outcomes Research Branch, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, EPN 4088, 6130 Executive Blvd., MSC 7344, Bethesda, MD 20892-7344, USA. reeveb@mail.nih.gov
- Dagmar Amtmann, Ph.D., Research Assistant Professor, University of Washington, Department of Rehabilitation Medicine, Box 357920, Seattle, WA 98195-7920, (206) 543-4741 V, (206) 685-9224 FAX, dagmara@u.washington.edu

Conflict of Interest Statement

The study was funded by a grant from the EORTC Quality of Life Group. There were no financial relationships, personal relationships, academic competition, intellectual commitments, or other conflicts of interest that might have biased the work.

Yours sincerely on behalf of the authors,

Morten Aa. Petersen
The Research Unit, Department of Palliative Medicine,
Bispebjerg Hospital,
Bispebjerg bakke 23,
2400 Copenhagen NV, Denmark.
Telephone: (+45) 3531 2025. Fax: (+45) 3531 2071. Email: mpet0009@bbh.regionh.dk

Background: Patient-reported outcomes should ideally be adapted to the individual patient while maintaining comparability of scores across patients. This is achievable using computerized adaptive testing (CAT). The aim here was to develop an item bank for CAT measurement of the pain domain as measured by the EORTC QLQ-C30 questionnaire.

Methods: The development process consisted of four steps. 1) Literature search. 2) Formulation of new items. 3) Pre-testing. 4) Field-testing and psychometric analyses for the final selection of items.

Results: In step 1) we identified 337 pain items from the literature. 2) Twenty-nine new items fitting the QLQ-C30 item style were formulated. Expert evaluations reduced this to 26 items. 3) Based on patient interviews (N=31) the list was further reduced to 21 items. 4) We obtained responses from 1,103 cancer patients from five countries. Psychometric evaluations showed that 16 items could be retained in a unidimensional item bank. Evaluations indicated that use of the CAT measure may reduce sample size requirements with 15-25% compared to using the QLQ-C30 pain scale.

Conclusions: We have established an item bank of 16 items suitable for CAT measurement of pain. We recommend initiating CAT measurement by screening for pain using the two original QLQ-C30 pain items.

Development of an item bank for computerized adaptive test (CAT) measurement of pain

M. Aa. Petersen[1,*], N. K. Aaronson[2], W.-C. Chie[3], T. Conroy[4], A. Costantini[5], E. Hammerlid[6], M. J. Hjermstad[7,8], S. Kaasa[9], J. H. Loge[8,10], G. Velikova[11], T. Young[12] & M. Groenvold[1,13] on behalf of the EORTC Quality of Life Group

[1] The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, University of Copenhagen , Copenhagen, Denmark

[2] Division of Psychosocial Research & Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[3] Institute of Epidemiology and Preventive Medicine and Department of Public Health, College of Public Health, National Taiwan University, Taiwan

[4] Medical Oncology Department, Institut de cancérologie de Lorraine, Vandoeuvre-lès-Nancy, France

[5] Psychoncology Unit, Sant'Andrea Hospital, Faculty of Medicine and Psychology Sapienza University, Rome, Italy

[6] Dept of Otolaryngology Head and Neck Surgery, Sahlgrenska University Hospital, Göteborg University, Göteborg, Sweden

[7] Regional Centre for Excellence in Palliative Care, Department of Oncology, Oslo University Hospital, Oslo, Norway

[8] European Palliative Care Research Centre, Faculty of Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[9] Palliative Medicine Unit, University Hospital of Trondheim, Trondheim, Norway

[10] National Resource Centre for Late Effects after Cancer Treatment, Oslo University Hospital, Oslo, Norway

[11] Cancer Research UK Centre, University of Leeds, Leeds, UK

[12] Lynda Jackson Macmillan Centre, Mount Vernon Hospital, Middlesex, UK

[13] Institute of Public Health, University of Copenhagen, Copenhagen, Denmark

[*] Corresponding author: The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, Bispebjerg bakke 23, 2400 Copenhagen NV, Denmark. Telephone: (+45) 3531 2025. Fax: (+45) 3531 2071. Email: mpet0009@bbh.regionh.dk.

Running head: Development of a pain item bank

Key words: Computerized adaptive test; EORTC QLQ-C30; item response theory; item development; item banking; pain; patient-reported outcome.

Manuscript submitted as an original article.

Funding Sources: This study was funded by grants from the EORTC Quality of Life Group.

Conflicts of interest disclosures: None of the authors have conflicts of interest with respect to this work.

What's already known about this topic?

- The EORTC QLQ-C30 is a well-validated and widely used questionnaire for measuring health-related quality of life including pain.
- Computer adaptive testing (CAT) adapts the questionnaire to the individual, thereby optimising measurement efficiency and reliability.

What does this study add?

- An item bank of 16 items for CAT measurement of pain was developed.
- The new items measure the same pain aspects as the QLQ-C30 pain scale.
- The new EORTC CAT instrument improves the measurement of pain compared to the QLQ-C30.

Abstract:

Background: Patient-reported outcomes should ideally be adapted to the individual patient while maintaining comparability of scores across patients. This is achievable using computerized adaptive testing (CAT). The aim here was to develop an item bank for CAT measurement of the pain domain as measured by the EORTC QLQ-C30 questionnaire.

Methods: The development process consisted of four steps. 1) Literature search. 2) Formulation of new items. 3) Pre-testing. 4) Field-testing and psychometric analyses for the final selection of items.

Results: In step 1) we identified 337 pain items from the literature. 2) Twenty-nine new items fitting the QLQ-C30 item style were formulated. Expert evaluations reduced this to 26 items. 3) Based on patient interviews (N=31) the list was further reduced to 21 items. 4) We obtained responses from 1,103 cancer patients from five countries. Psychometric evaluations showed that 16 items could be retained in a unidimensional item bank. Evaluations indicated that use of the CAT measure may reduce sample size requirements with 15-25% compared to using the QLQ-C30 pain scale.

Conclusions: We have established an item bank of 16 items suitable for CAT measurement of pain. We recommend initiating CAT measurement by screening for pain using the two original QLQ-C30 pain items.

Key words: Computerized adaptive test; EORTC QLQ-C30; item response theory; item development; item banking; pain assessment; patient-reported outcome.

Introduction

Adequate pain management requires reliable and precise assessment. As pain is a subjective symptom, assessment should be based on the patients' own perception of their pain (Green et al., 2010; Noble et al., 2005). This can be achieved using self-report questionnaires, also termed patient-reported outcomes (PROs). PROs have typically been developed using classical methods like sum scoring of items. However, classical methods have some limitations. For example, all patients have to answer the same set of items for scores to be comparable. This means that patients often have to answer items that are not relevant for their level of pain and/or relevant items are missing since the total number of items has to be limited to keep the respondent burden at a reasonable level.

Using item response theory (IRT) (Hambleton et al., 1991; van der Linden and Hambleton, 1997) for developing and scoring PROs overcomes some of these limitations. In particular, when a set of items has been calibrated (estimated) to an IRT model all scores based on any subset of the items are comparable. This unique feature means that a questionnaire can be adapted to the individual without compromising comparability across patients. This is utilized in computer adaptive testing (CAT): (Wainer, 2000) based on the responses to the preceding items, a computer program evaluates which item should be asked next to obtain maximal information. In this way the questionnaire is adapted to the individual, using the most informative items for each patient, thereby optimising both the efficiency and the reliability of the assessment.

The European Organisation for Research and Treatment of Cancer Quality of Life Group (EORTC QLG) (http://groups.eortc.be/qol/) is carrying out a large scale project with the overall aim to improve the measurement of the health-related quality of life (HRQOL)

domains included in the EORTC Quality of Life Questionnaire, the QLQ-C30 (Aaronson et al., 1993; Giesinger et al., 2011; Petersen et al., 2012; Petersen et al., 2013; Petersen and Groenvold, 2013; Petersen et al., 2010; Petersen et al., 2011). This is achieved by developing a CAT measure for each domain in the EORTC QLQ-C30. This CAT instrument will be more precise, efficient, and flexible than the EORTC QLQ-C30.

The QLQ-C30 is one of the most widely used HRQOL questionnaires in cancer research (Fayers and Bottomley, 2002; Garratt et al., 2002). It consists of 30 items measuring 15 aspects of HRQOL. In the QLQ-C30 pain is measured with two items, one about pain intensity ("Have you had pain?") and one about pain interference ("Did pain interfere with your daily activities?") combined into a single overall score (Fayers et al., 2001). These are generic items relevant for all patients. However, if more items assessing different levels of pain intensity/interference were available, more precise pain measurement could be obtained. With CAT this can be achieved without imposing an unreasonable response burden on the patients.

The aim of the study was to develop a collection of items (a so-called "item bank") for CAT measurement of the pain domain as measured with the QLQ-C30. The intention is to supplement the two pain-items of the QLQ-C30 with new items thereby increasing measurement precision and extend the range of pain intensity/interference that can be assessed. The current paper describes the development and initial validation of this EORTC pain item bank.

Materials and Methods

The general steps in the development of the item banks for the CAT version of the EORTC QLQ-C30 have been described in detail previously (Giesinger et al., 2011; Petersen et al., 2012; Petersen et al., 2013; Petersen et al., 2010; Petersen et al., 2011). The following summarises each step of the development of the pain item bank.

1. Literature search

A literature search was conducted to identify existing instruments and items used to measure pain. This was not a systematic review, but rather aimed at acquiring sufficient information about pain measurement to form the basis for formulating new, relevant items. In the current case, the literature search was primarily based on an early version of the review of pain assessment tools conducted by the European Palliative Care Research Collaborative (Holen et al., 2006). This list of items was supplemented with searches for additional pain items in the PROQOLID database (http://www.proqolid.org) and the EORTC QLG Item Bank.

2. Formulation of items and expert evaluations

First, the list of items identified in step 1 was trimmed: items assessing aspects of pain other than intensity or interference and items with content that did not fit the "QLQ-C30 item style" (i.e., the response categories, time frame, etc.) were deleted. The resulting "shortlist" of items was used as inspiration for formulating new items measuring pain intensity and interference. The new items should have the same item style as the two QLQ-C30 pain items, i.e. they should fit the timeframe "during the past week" and the response options "not at all", "a little", "quite a bit" and "very much". The intension was that the new items should extend the range of pain that could be assessed (from minimal to very high levels of pain). Therefore, we aimed at formulating items relevant for different levels of pain. The item selection and formulation was carried out independently by two members of the project group. After each

step, possible differences were discussed and a consensus was reached. The list of developed items was evaluated by international experts in pain measurement. The item list was revised based on these experts' evaluations.

3. Pre-testing

The revised list of items was evaluated by a mixed, international sample of cancer patients. Before the interviews, the items were translated into the relevant languages by the Translation Office of the EORTC Quality of Life Department according to rigorous and well-established guidelines developed by the EORTC (Dewolf et al., 2009; Koller et al., 2007). The interviews followed the EORTC QLG guidelines for pre-testing of items (Johnson et al., 2011) and elucidated whether patients found some of the items difficult to answer, confusing, annoying, upsetting, intrusive, etc.

4. Field-testing and psychometric analyses

The items were field-tested in an international and heterogeneous sample of cancer patients. The sample included patients having different levels of pain, from "no" pain to "severe" pain. To ensure stable calibration of the IRT model, we aimed to collect at least 1,000 responses (Muraki and Bock, 1996). The patients completed the new pain items together with the QLQ-C30. They also completed sociodemographic items and "debriefing items" to clarify whether certain items were inappropriate, ambiguous, etc.

The resulting dataset formed the basis for the final psychometric evaluations. These evaluations included:

a. Descriptive and basic statistical analyses. This included calculation of item mean scores, response frequencies, percent missing responses, and correlations with the QLQ-C30 pain scale.

b. Evaluation of dimensionality and local dependence. As the QLQ-C30 assesses pain using a unidimensional scale, the aim was a unidimensional CAT measure of pain. We used factor analysis for ordinal variables to explore the dimensionality of the item set (Muthen, 1984; Muthen and Muthen, 2002). This included evaluations of dimensionality based on eigenvalues (including scree plot (Cattell, 1966)) and the following fit indices: root mean square error of approximation (RMSEA), the Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI). Local independence (i.e. whether item responses are independent when controlling for the overall level of pain (van der Linden and Hambleton, 1997)) was investigated using residual correlations (Bjorner et al., 2003; Fliege et al., 2005; van der Linden and Hambleton, 1997).

c. Calibration of the IRT model and evaluation of item fit. We used the generalized partial credit model (GPCM) (Muraki, 1997) as the IRT model forming the basis for the CAT. This was calibrated using Parscale (Muraki and Bock, 1996). In the GPCM, each item has a slope parameter describing the item's ability to discriminate between subjects with different levels of pain, and a set of threshold parameters describing how likely it is to report problems on the item. Item fit was examined using Muraki's test (Muraki, 1997), bias estimates (average difference between expected and observed item responses) and the infit and outfit statistics, which are mean square residuals often used in Rasch fit analysis (Bond and Fox, 2007; Petersen et al., 2013; Wright and Linacre, 1994; Wright and Masters, 1982). Infit and outfit values between 0.7 and 1.3 are often regarded as acceptable (Wright and Linacre, 1994).

d. Test for differential item functioning (DIF). DIF analysis explores whether items function differently for different groups of patients (Holland and Wainer, 1993). Using ordinal logistic regression methods (French and Miller, 1996; Petersen et al., 2003) we tested for DIF with regard to gender, age, country, cancer site, cancer stage, current treatment, education, work, and cohabitation. Significant DIF findings (Bjorner et al., 1998; Petersen et al., 2003) were evaluated for their impact on the estimation of pain, i.e. whether the DIF findings seemed to have practical consequences for pain estimation (Hart et al., 2009; Petersen et al., 2013; Petersen et al., 2011).

e. Evaluation of measurement properties. We evaluated the measurement precision of the resulting CAT pain measure using simulations of CAT administration based on the collected responses. We simulated CATs asking 1, 2,… up to all but 1 item, respectively, estimated the pain score based on these CATs, and compared them with the pain score based on all items. Using two-sample t-test sizes we evaluated the relative validity (RV) of these CATs as compared to the QLQ-C30 pain scale in detecting expected group differences (Fayers and Machin, 2007). We hypothesized that patients currently on treatment (chemotherapy or other cancer related treatment) would have significantly more pain than patients not on treatment, and that patients with stage III or IV disease would have more pain than patients with stage I or II disease. In addition to these evaluations based on the observed data, we also evaluated the RV of the CATs based on simulated data. We simulated responses to the items based on pain scores sampled from normal distributions with different means. We compared groups of size $N_1=N_2=25$, 50, and 100, respectively and true effect sizes (ESs) of 0.2, 0.5, and 0.8, respectively. For each of these 3x3=9 possible settings, we ran 2,000 simulations. For further details please see Petersen el al. (Petersen et al., 2012) From the RVs we estimated the approximate savings in sample sizes using the CATs compared to the QLQ-C30 scale. The RVs can

also be expressed as ratios of SDs, e.g. a RV=1.1 corresponds to a ratio of SDs of 1.1^{-1} = 0.9 (when the mean difference is fixed) reflecting that a more precise measure results in less noise (smaller SD). From these ratios of SDs approximate sample size reductions can be estimated (Petersen et al., 2012).

The study was approved by the local ethical committees in the participating countries. Informed consent was obtained from each participating patient.

Results

1. Literature search

The review of pain assessment tools (Holen et al., 2006) resulted in a list of 231 items. This was supplemented with 65 additional pain items identified in PROQOLID and with 41 items identified in the EORTC QLG Item Bank, resulting in a total of 337 pain items.

2. Formulation of items and expert evaluations

We classified the identified items as measuring either pain intensity, interference or something else. Only items judged to measure pain intensity or interference were retained. As the two QLQ-C30 items ask about pain, in general, without reference to specific body parts, the items were further required to ask about pain in general or to be able to be reformulated to do so. In all, 140 items complied with these requirements and were therefore retained. Next, we deleted redundant items and items that could not be reformulated into the QLQ-C30 item style. This resulted in the deletion of 113 items, leaving a list of 27 items. For example, “How OFTEN did you have it (pain)?” did not fit the response categories in the QLQ-C30 and “Over the past 3 days, have you been affected by pain?” was judged to be too close in content to the QLQ-C30 item “Have you had pain?”. The remaining 27 items were

used as inspiration for formulating new, unique items complying with the QLQ-C30 item style. This resulted in the formulation of 29 new candidate pain items.

The 29 items were evaluated by 11 experts from Denmark, Germany, Italy, the Netherlands, and the UK. The expert evaluations resulted in rewording of three items and deletion of four items: two because of redundancy, one because of ambiguity and one because of poor fit to the response options. The expert evaluations resulted in the addition of one new item. Hence, after these evaluations, the list consisted of 26 items.

3. Pre-testing

A total of 31 patients were interviewed about the 26 candidate items and the two QLQ-C30 pain items. The patients came from Denmark, France and the UK and included both genders and 11 different cancer sites. Based on the interviews we changed the wording of four items to make them clearer and seven items were deleted: five because of redundancy and two because several patients found them ambiguous/unclear. Hence, after these interviews the list consisted of 19 candidate items plus the two QLQ-C30 items, in all 21 pain items. Of these, 15 measured pain interference and six pain intensity.

4. Field-testing and psychometric analyses

We obtained responses from 1,103 cancer patients. Patient characteristics are reported in Table 1.

a. Descriptive and basic statistical analyses. The response rate per item was generally high (98.1%-99.1%). The average response across the items ranged from 0.33 to 0.86 on a 0-3 scale (with 0=”not at all”), indicating generally low pain levels in the sample, although about 250 patients (23%) reported “quite a bit” or “very much” pain (item 21). Polychoric

correlations between the new items and the QLQ-C30 pain scale ranged from 0.79 to 0.92.

(Table 1 about here)

b. Evaluation of dimensionality and local dependence. Exploratory eigenvalues analysis indicated that the first factor explained 85% of the total variation. Subsequent factors all explained <4% of the variation and all had eigenvalues<1. A unidimensional solution with the 21 items had RMSEA=0.147, CFI=0.977, TLI=0.995. The RMSEA was somewhat large, but all other indices indicated that all 21 items could be retained in a unidimensional model. All residual correlations between the 21 items were <0.10, i.e. no indications of local dependence. Therefore, all 21 items were retained.

c. Calibration of the IRT model and evaluation of item fit. Our initial attempts to fit an IRT model to the 21 items failed; either the estimation procedure could not converge or it resulted in unreliable, extreme estimates of some parameters. Inspections of item crosstabs and correlations revealed that, although residual correlations did not indicate local dependence, several items were highly correlated (polychoric correlations>0.9). The main reason was that 38% of the sample had responded "not at all" (i.e. no pain) to all items. The responses from these patients are clearly mutually highly predictable/locally dependent. Among patients responding "not at all" to the two original QLQ-C30 pain items (44% of the sample), 85% had responded "not at all" to all items, and their average score was 0.7 on a 0-100 scored sum scale consisting of the 21 items. Hence, asking these "no pain" patients several pain items would have very little relevance from either a clinical or a measurement perspective. Therefore, we decided to exclude the patients responding "not at all" to the two QLQ-C30 pain items from the IRT analyses. Doing this, it was possible to fit an IRT model. Evaluation

of item fit indicated that five items had poor fit to the model and/or were locally dependent with some of the other items and were therefore deleted. Fit statistics for the remaining 16 items are shown in Table 2. All item fit tests had $p>0.04$ (and except for item 11 all >0.10). Bias estimates were all very close to 0 (<0.1). The infit statistics ranged 0.76-1.07 and the outfit statistics 0.71-1.03; all within the acceptable range. Hence, the fit of these 16 items was deemed acceptable.

(Table 2 about here)

d. Test for DIF. There was no significant DIF with regard to age, gender, cancer stage, education, or cohabitation. Items 4, 6, 7, and 21 showed significant DIF between countries and item 7 also showed DIF between patients on and off treatment. There was significant DIF between cancer sites for item 16 and between working and retired patients for item 18. We evaluated the possible impact of these DIF findings for the estimation of pain. These evaluations indicated that the possible DIF would have only negligible impact on the estimation of pain, i.e. the possible DIF did not result in biased pain scores for any groups of patients. Therefore, no items were deleted because of DIF.

Parameter estimates of the final 16 items are shown in Table 2. The slopes indicate how well the items discriminate between different levels of pain, while the locations (the average of the item's threshold parameters) indicate the level of pain for which the items are most informative. Except for item 21 all locations were >0 indicating that the items were mainly relevant for patients with at least a little pain. This is also evident from Fig. 1 which shows the total information when using all 16 items and the two QLQ-C30 items only, respectively. Considering information=10 (corresponding to a reliability of 0.90) as a threshold for reliable

measurement, the total item bank provided reliable measurement from about –1.0 to 2.5 (3.5 standard deviation units). Asking the two QLQ-C30 items only provide markedly less information for all levels of pain except for patients with "no pain" (lower extreme). Hence, the two QLQ-C30 items may be particularly useful for screening for patients with "no pain"; 88% of those answering "not at all" to these two items, answer "not at all" (i.e. "no pain") to all 16 items.

(Fig. 1 about here)

(Fig. 2 about here)

e. Evaluation of measurement properties.

Fig. 2 shows that for CATs of all lengths the median pain score was very close to the median score based on all items (all deviations<0.06). For about 50% of the patients the scores obtained using only one item deviated >0.4 from the score based on all items, while when asking five items only about 10% had scores deviating >0.4. As the score based on all items ranged from -2.2 to 3.0, 0.4 is less than 8% of the possible score range, i.e. similar to 8 points on a 0-100 scale. Scores based on one item correlated 0.77 with the scores based on all items, with two items the correlation was 0.88, while using three or more items the correlations were >0.92 (results not shown).

Fig. 3 summarizes the results of the known-groups comparisons. The comparisons based on the observed data confirmed the hypothesized group differences and indicated that CAT measurement asking four or more items reduced the sample size requirements by about 20-25% without loss of power compared to using the original QLQ-C30 pain scale. The estimated reduction in required sample size/increased power was somewhat lower based on

the simulated data. These simulations indicated that, regardless of the length of the CAT, sample sizes may be reduced by less than 15%, and at least seven items may be required for a reduction >10%.

(Fig. 3 about here)

Discussion and conclusions

The EORTC QLG is developing a CAT instrument for assessing the HRQOL domains included in the widely used EORTC QLQ-C30 questionnaire. The EORTC CAT development process can be divided into four phases: literature search, item construction, pre-testing, and field-testing. These phases are closely related to the phases of EORTC QLG module development (Johnson et al., 2011). Here we have reported the results of the development of the pain item bank.

The literature search yielded valuable insights into how pain may be measured. This was a useful inspiration for formulating items measuring pain in the "EORTC QLQ-C30 way". Note that the literature search was not, and was not intended to be, an exhaustive review of all available pain items, but was intended to identify the different ways pain items may be formulated to cover different aspects and levels of pain. With the identification of over 300 pain items we feel confident that all relevant item variants were covered sufficiently. From the literature search and item formulation it was apparent that, with our requirements for item formatting (response options etc.), it was difficult to construct a large number of relevant and distinct items, particularly about pain intensity. Hence, only four of the 16 items in the final item bank ask directly about pain intensity. The remaining items measure pain interference. But clearly, these items also provide valuable insight into the level of pain (the more pain

interfered, the more severe it was likely to have been). The psychometric analysis also indicated that the intensity and interference items together formed a unidimensional construct. Still, it may be preferable to avoid asking only interference items. In this case the CAT can be programmed to always include both intensity and interference items.

Basing the CAT on an established HRQOL instrument, the QLQ-C30, has several advantages including backward compatibility with a very large literature, measurement of well-validated HRQOL domains, and simplified conceptual work as we did not have to establish a whole new framework of measurement. However, this approach limited the number of relevant items that could be constructed. Still, we have extended the QLQ-C30 pain scale from two to 16 items, a considerable expansion.

The expert and patient evaluations were invaluable in identifying problematic items, and in optimizing item formulation. The patient perspective was particularly important in ensuring that the items are appropriate and comprehensible for cancer patients. We deleted seven of the candidate items because of redundancy or other problems pointed out by patients during the pre-testing.

The psychometric analyses indicated that 16 items could be retained in a unidimensional item bank. IRT calibration and evaluations generally showed good fit of these items. We found some indications of DIF. However, evaluations of the DIF indicated that these had no significant impact on the estimation of pain. Hence, we consider the items to be appropriate for general use in cancer patients, regardless of gender, age, cancer site etc. Further, as the items have been developed, calibrated and evaluated in an international setting, the CAT

measure will be appropriate for international use. As with the QLQ-C30, the CAT measure is intended to be applicable and available in a multitude of languages.

To estimate the IRT model we had to exclude a substantial proportion of patients (44% of our sample) who had responded "not at all" to the two QLQ-C30 pain items. The responses from these patients reflected that they generally had no or very little pain and, hence, detailed questioning about their level of pain would have little relevance from either a clinical or a measurement perspective. Further, as is apparent from the item location parameters which all are > 0 except for item 21 (Table 2) and the information function (Fig. 1), the items are primarily relevant for patients who have at least a little pain. Therefore, it may be most appropriate to first screen patients for pain using the two QLQ-C30 pain items. Only those patients who report at least "a little" pain on one of these two items would then go on to complete the CAT items (the number of which would depend on the desired level of measurement precision).

In general, the CAT pain items appeared to be efficient and precise. For example, pain scores based on three items correlated 0.93 with the scores based on all 16 items. However, as the evaluations were based on the data used to calibrate the IRT model, the efficiency and precision may have been overestimated. We intend to conduct additional studies to examine these issues in independent data.

The results of the known-groups comparisons based on observed data indicated that, if four or more items are asked, sample sizes may be reduced by 20-25% without loss of power as compared to the QLQ-C30. The simulations on the other hand indicated that this reduction would be at most 10-15%. Hence, both analyses indicate increased power using the new

measure, but it is inconclusive what the actual gain may be and this will probably vary across studies. Additional, detailed evaluations of the power of the CAT measure in independent data are needed and are planned.

The EORTC CAT pain measure includes the two QLQ-C30 pain items, and all new items have been constructed and selected to measure the same pain aspects as the C30 items and to have the same item format. The purpose of this was to obtain a homogenous and user-friendly measure measuring the same concept as the QLQ-C30 pain scale. Based on the strong associations between the original and new items it seems reasonable to assume that the new measure is also valid. However, it would still be interesting and relevant to compare the EORTC CAT pain instrument with external, validated pain measures. This will elucidate the validity and importantly, will also allow the construction of linking (Chen et al., 2009; Dorans, 2007) (also called cross-walking (Noonan et al., 2012)) between the EORTC pain measure and other established pain measures.

Although the development of the EORTC CAT has not yet been completed, the current version may be used for "experimental" purposes. By "experimental" is meant that until the final, validated version of the EORTC CAT is released, it should be used in parallel with the EORTC QLQ-C30. The EORTC CAT item banks may also be used to construct so called (paper) short forms. For example, if a trial is comparing a new analgesic to the standard treatment, then to increase the study power (without increasing sample size) it may be advantageous to supplement the QLQ-C30 with a pain short form of e.g. five additional pain items, targeted to the study population. As the items are selected from the calibrated item bank, scores based on such short forms are directly comparable to scores based on the CAT

measure. For more information on this preliminary use of the EORTC CAT and short forms please visit http://groups.eortc.be/qol/eortc-cat.

In conclusion, we have developed an item bank of 16 items for CAT measurement of pain. This CAT measure will be backward compatible with the QLQ-C30 and hence with the many studies that have used this questionnaire. The item bank showed good psychometric properties and high measurement precision for patients with some degree of pain. It is for these patients that more detailed information would be particularly useful. Evaluations of power were somewhat ambiguous, but indicated that sample sizes may be reduced up to 25% without loss of power, compared to the QLQ-C30. However, these measurement properties should be validated with new data before drawing any final conclusions. Even though the item bank is targeted cancer patients, the items are formulated in general, non-cancer specific terms and hence may be applied in other patient populations (and the general population) as well.

Acknowledgements

The authors would like to thank the participating experts and patients and our collaborators for helping with the collection of the essential patient data.

References

EORTC Item Bank Guidelines. Available at http://groups.eortc.be/qol/downloads/200104itembank_guidelines.pdf Accessed June, 2009.

Aaronson,N.K., Ahmedzai,S., Bergman,B., Bullinger,M., Cull,A., Duez,N.J., Filiberti,A., Flechtner,H., Fleishman,S.B. & de Haes,J.C. (1993) The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst, **85**, 365-376.

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

Bjorner,J.B., Kreiner,S., Ware,J.E., Damsgaard,M.T. & Bech,P. (1998) Differential item functioning in the Danish translation of the SF-36. J Clin Epidemiol, **51**, 1189-1202.

Bond,T.G. & Fox,C.M. (2007) Applying the Rasch model: Fundamental measurement in the human sciences Lawrence Erlbaum Associates, Inc., New Jersey.

Cattell,R.B. (1966) Scree Test for Number of Factors. Multivariate Behavioral Research, **1**, 245-276.

Chen,W.H., Revicki,D.A., Lai,J.S., Cook,K.F. & Amtmann,D. (2009) Linking pain items from two studies onto a common scale using item response theory. J.Pain Symptom.Manage., **38**, 615-628.

Dewolf,L., Koller,M., Velikova,G., Johnson,C., Scott,N. & Bottomley,A. (2009) EORTC Quality of Life Group Translation Procedure European Organization for Research and Treatment of Cancer, Brussels.

Dorans,N.J. (2007) Linking scores from multiple health outcome instruments. Qual Life Res, **16**, 85-94.

Fayers,P. & Bottomley,A. (2002) Quality of life research within the EORTC-the EORTC QLQ - C30. European Organisation for Research and Treatment of Cancer. Eur J Cancer, **38 Suppl 4**, S125-S133.

Fayers,P.M., Aaronson,N.K., Bjordal,K., Groenvold,M., Curran,D. & Bottomley,A. (2001) The EORTC QLQ-C30 Scoring Manual European Organisation for Research and Treatment of Cancer, Brussels.

Fayers,P.M. & Machin,D. (2007) Quality of Life. The assessment, analysis and Interpretation of patient-reported outcomes John Wiley & Sons Ltd, Chichester.

Fliege,H., Becker,J., Walter,O.B., Bjorner,J.B., Klapp,B.F. & Rose,M. (2005) Development of a computer-adaptive test for depression (D-CAT). Qual Life Res, **14**, 2277-2291.

French,A.W. & Miller,T.R. (1996) Logistic regression and its use in detecting differential item functioning in polytomous items. J Educ Meas, **33**, 315-332.

Garratt,A., Schmidt,L., Mackintosh,A. & Fitzpatrick,R. (2002) Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ, **324**, 1417-1419.

Giesinger,J.M., Petersen,M.Aa., Groenvold,M., Aaronson,N.K., Arraras,J.I., Conroy,T., Gamper,E.M., Kemmler,G., King,M.T., Oberguggenberger,A.S., Velikova,G. & Young,T. (2011) Cross-cultural development of an item list for computer-adaptive testing of fatigue in oncological patients. Health Qual Life Outcomes, **9**.

Green,E., Zwaal,C., Beals,C., Fitzgerald,B., Harle,I., Jones,J., Tsui,J., Volpe,J., Yoshimoto,D. & Wiernikowski,J. (2010) Cancer-related pain management: a report of evidence-based recommendations to guide practice. Clin J Pain, **26**, 449-462.

Hambleton,R.K., Swaminathan,H. & Rogers,H.J. (1991) Fundamentals of Item Response Theory Sage Publications, Inc, Newbury Park.

Hart,D.L., Deutscher,D., Crane,P.K. & Wang,Y.C. (2009) Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. Qual Life Res, **18**, 1067-1083.

Holen,J.C., Hjermstad,M.J., Loge,J.H., Fayers,P.M., Caraceni,A., De Conno,F., Forbes,K., Furst,C.J., Radbruch,L. & Kaasa,S. (2006) Pain assessment tools: is the content appropriate for use in palliative care? J Pain Symptom Manage, **32**, 567-580.

Holland,P.W. & Wainer,H. (1993) Differential Item Functioning Lawrence Erlbaum Associates, Hilsdale, NJ.

Johnson,C., Aaronson,N., Blazeby,J.M., Bottomley,A., Fayers,P., Koller,M., Kulis,D., Ramage,J., Sprangers,M., Velikova,G. & Young,T. (2011) EORTC Quality of Life Group - Guidelines for Developing Questionnaire Modules European Organisation for Research and Treatment of Cancer, Brussels.

Koller,M., Aaronson,N.K., Blazeby,J., Bottomley,A., Dewolf,L., Fayers,P., Johnson,C., Ramage,J., Scott,N. & West,K. (2007) Translation procedures for standardised quality of life questionnaires: The European Organisation for Research and Treatment of Cancer (EORTC) approach. Eur J Cancer, **43**, 1810-1820.

Muraki,E. (1997) A Generalized Partial Credit Model. In van der Linden,W.J. & Hambleton,R.K. (eds), Handbook of Modern Item Response Theory. Springer, Berlin, pp. 153-168.

Muraki,E. & Bock,R.D. (1996) PARSCALE - IRT based Test Scoring and Item Analysis for Graded Open-ended Exercises and Performance Tasks Scientific Software International, Inc., Chicago.

Muthen,B. (1984) A General Structual Equation Model With Dichotomous, Ordered Catagorical and Continuous Latent Variable Indicators. Psychometrika, **49**, 115-132.

Muthen,L.K. & Muthen,B.O. (2002) Mplus User's Guide Muthen & Muthen, Los Angeles, CA.

Noble,B., Clark,D., Meldrum,M., ten,H.H., Seymour,J., Winslow,M. & Paz,S. (2005) The measurement of pain, 1945-2000. J Pain Symptom.Manage., **29**, 14-21.

Noonan,V.K., Cook,K.F., Bamer,A.M., Choi,S.W., Kim,J. & Amtmann,D. (2012) Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the Modified Fatigue Impact Scale and the PROMIS Fatigue Short Form. Qual.Life Res., **21**, 1123-1133.

Petersen,M.Aa., Aaronson,N.K., Arraras,J.I., Chie,W.-C., Conroy,T., Costantini,A., Giesinger,J.M., Holzner,B., King,M.T., Singer,S., Velikova,G., Young,T. & Groenvold,M. (2012) The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency. J Clin Epidemiol, **66**, 330-339.

Petersen,M.Aa., Giesinger,J.M., Holzner,B., Arraras,J.I., Conroy,T., Gamper,E.M., King,M.T., Verdonck-de Leeuw,I.M., Young,T. & roenvold,M. (2013) Psychometric evaluation of the EORTC computerized adaptive test (CAT) fatigue item pool. Qual Life Res, **22**, 2443-2454.

Petersen,M.Aa. & Groenvold,M. (2013) Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30. MAPI PRO Newsletter, **Spring 2013**.

Petersen,M.Aa., Groenvold,M., Aaronson,N.K., Chie,W.-C., Conroy,T., Costantini,A., Fayers,P., Helbostad,J., Holzner,B., Kaasa,S., Singer,S., Velikova,G. & Young,T. (2010) Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions – General approach and initial results for physical functioning. Eur J Cancer, **46**, 1352-1358.

Petersen,M.Aa., Groenvold,M., Aaronson,N.K., Chie,W.-C., Conroy,T., Costantini,A., Fayers,P., Helbostad,J., Holzner,B., Kaasa,S., Singer,S., Velikova,G. & Young,T. (2011) Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. Qual Life Res, **20**, 479-490.

Petersen,M.Aa., Groenvold,M., Bjorner,J.B., Aaronson,N.K., Conroy,T., Cull,A., Fayers,P.M., Hjermstad,M., Sprangers,M.A. & Sullivan,M. (2003) Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Qual Life Res, **12**, 373-385.

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

Wainer,H. (2000) Computerized Adaptive testing: A Primer Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.

Wright,B.D. & Linacre,J.M. (1994) Reasonable mean-square fit values. Rasch Measurement Transactions, **8**, 370.

Wright,B.D. & Masters,G.N. (1982) Rating scale analysis - Rasch measurement MESA PRESS, Chicago.

Figure legends

Fig. 1. Test information function for the 16 items in the final model and for the two EORTC QLQ-C30 pain items, respectively.

Footnote to Fig. 1: The pain scores obtained if answering “not at all”, “a little”, “quite a bit”, or “very much”, respectively to all 16 items are indicated at the horizontal axis.

Fig. 2. Median and percentiles for differences between pain scores based on fixed length CATs and scores based on all items.

Fig. 3. The average relative validity and relative required sample size using CAT measurement compared to using the QLQ-C30 sum scale based on observed and simulated data, respectively.

Table 1. Sociodemographic and clinical characteristics of the field study sample (N=1,103).

		N/mean
Age (mean years)		60 (range 19-90)
Gender	Male	484 (44%)
	Female	619 (56%)
Country	Denmark	435 (39%)
	Italy	81 (7%)
	Sweden	220 (20%)
	Taiwan	103 (9%)
	UK	264 (24%)
Education	0-10 years	335 (30%)
	11-13 years	275 (25%)
	14-16 years	236 (21%)
	>16 years	232 (21%)
Work	Working	391 (30%)
	Retired	516 (47%)
	Other	172 (16%)
Cohabitation	Living with a partner	817 (74%)
	Living alone	266 (24%)
Cancer stage	I-II	536 (49%)
	III-IV	518 (47%)
Cancer site	Breast	199 (18%)
	Gastrointestinal	131 (12%)
	Gynaecological	179 (16%)
	Head and neck	165 (15%)
	Lung	33 (3%)
	Other	191 (17%)
Current treatment	Chemotherapy	249 (23%)
	Other treatment	271 (24%)
	No current treatment	577 (52%)

Table 2. Parameter estimates (slope and location) and fit statistics for the 16 items in the final IRT model.

Item (Heading for all items: during the past week)	**Slope**	**Location**	**Item fit, p-value**	**Bias**	**Infit**	**Outfit**
1. Have you had any trouble falling asleep because of pain?[a]	1.59	1.06	0.109	0.01	0.98	0.93
2. Has pain made it difficult for you to do the jobs that you usually do around the house?[a]	2.25	0.62	0.122	-0.04	0.92	0.89
3. Have you had extreme pain?[b]	2.45	0.87	0.425	-0.04	0.96	1.03
4. Has pain made it difficult for you to stand for more than a few minutes?[a]	1.49	1.35	0.558	-0.00	1.03	0.88
6. Have you had pain you could not ignore?[b]	1.70	0.56	0.546	-0.03	0.86	0.85
7. Have you had to stay in bed during the day because of pain?[a]	1.87	1.52	0.323	-0.02	1.07	0.84
8. Has pain limited your ability to concentrate on work or other daily activities?[a]	2.53	0.81	0.205	-0.01	1.00	0.87
10. Have you had any trouble sleeping because of pain?[a]	2.10	0.88	0.190	0.02	0.94	0.83
11. Has pain interfered with your leisure activities (e.g. sports and hobbies)?[a]	1.99	0.39	0.043	-0.07	0.91	0.88
13. Has pain interfered with your social activities?[a]	2.72	0.69	0.307	0.00	0.94	0.81
14. Have you had severe pain?[b]	2.43	0.78	0.603	-0.02	0.89	0.94

15. Have you woken up earlier than you wanted to because of pain?[a]	2.06	0.84	0.431	0.03	0.93	0.82
16. Has pain made it difficult for you to sit for more than 1 hour?[a]	1.56	1.29	0.586	-0.01	0.96	0.88
17. Did pain interfere with your daily activities? (QLQ-C30 item 19)[a]	2.64	0.59	0.124	0.01	0.76	0.71
18. Have you had any trouble taking a walk because of pain?[a]	1.70	0.64	0.166	0.01	1.04	0.95
21. Have you had pain? (QLQ-C30 item 9)[b]	2.26	-0.28	0.400	-0.04	0.80	0.84

[a]: Item measures primarily pain interference.

[b]: Item measures primarily pain intensity.

Fig 1

Click here to download high resolution image

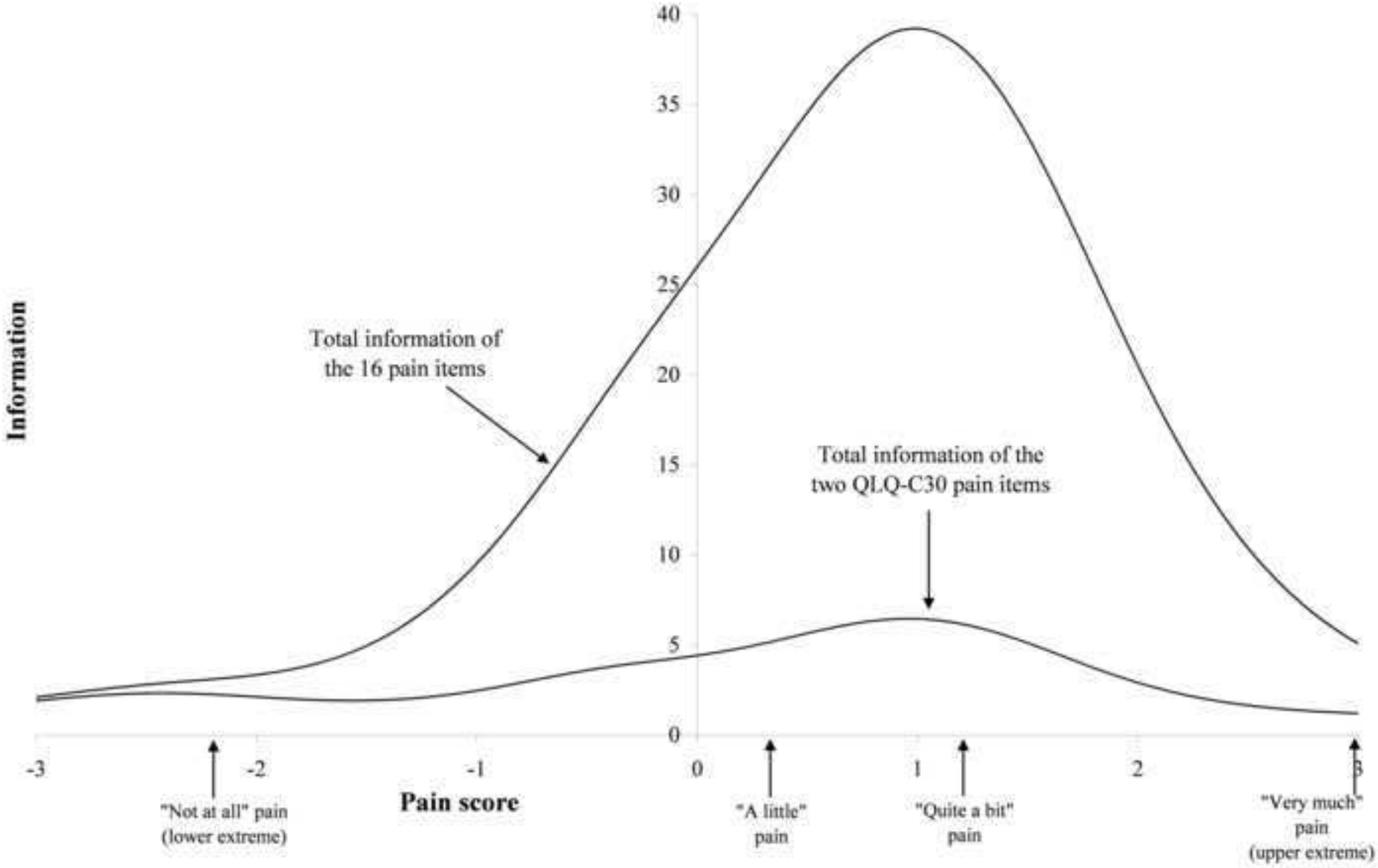

Fig 2

Click here to download high resolution image

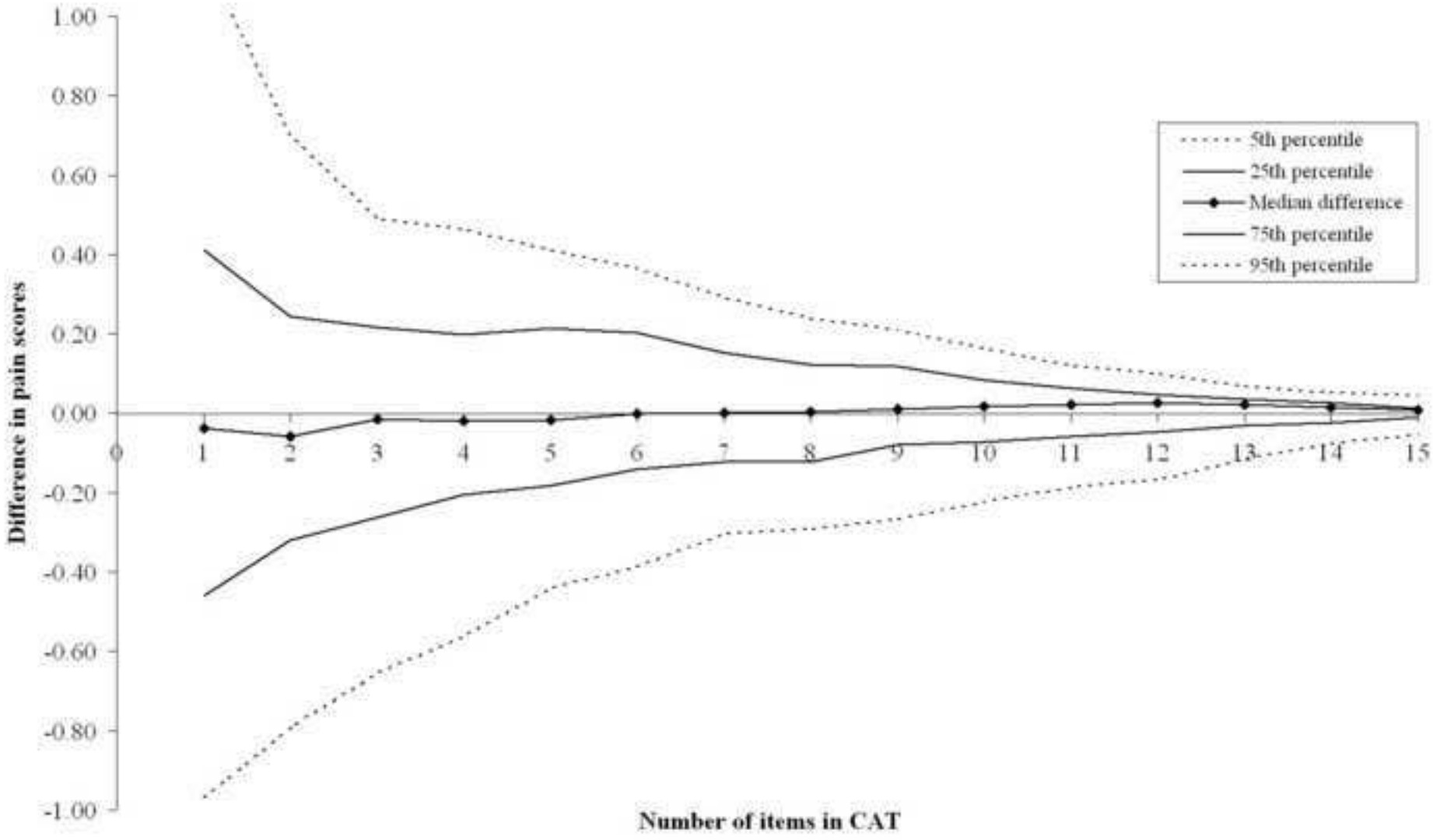

Fig 3
Click here to download high resolution image