available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

# Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions – General approach and initial results for physical functioning

*Morten Aa. Petersen* [a,*], *Mogens Groenvold* [a,b], *Neil K. Aaronson* [c], *Wei-Chu Chie* [d], *Thierry Conroy* [e], *Anna Costantini* [f], *Peter Fayers* [g,h], *Jorunn Helbostad* [i], *Bernhard Holzner* [j], *Stein Kaasa* [k], *Susanne Singer* [l], *Galina Velikova* [m], *Teresa Young* [n], *on behalf of the EORTC Quality of Life Group*

[a] *The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark*
[b] *Institute of Public Health, University of Copenhagen, Copenhagen, Denmark*
[c] *Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands*
[d] *Graduate Institute of Preventive Medicine and Department of Public Health, College of Public Health, National Taiwan University, Taiwan*[1]
[e] *Medical Oncology Department, Centre Alexis Vautrin, France*
[f] *Psychoncology Unit, Sant'Andrea Hospital, 2nd Faculty of Medicine, Sapienza University of Rome, Rome, Italy*
[g] *Department of Public Health, University of Aberdeen, UK*
[h] *Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway*
[i] *Department of Neuroscience, Norwegian University of Science and Technology and St. Olav University Hospital, Trondheim, Norway*
[j] *Department of Psychiatry, Medical University Innsbruck, Innsbruck, Austria*
[k] *Palliative Medicine Unit, University Hospital of Trondheim, Trondheim, Norway*
[l] *Department of Medical Psychology and Medical Sociology, University of Leipzig, Leipzig, Germany*
[m] *Cancer Research UK Centre, University of Leeds, Leeds, UK*
[n] *Lynda Jackson Macmillan Centre, Mount Vernon Hospital, Middlesex, UK*

ARTICLE INFO

*Article history:*
Received 30 November 2009
Accepted 8 February 2010
Available online 16 March 2010

*Keywords:*
Computerised adaptive test
EORTC QLQ-C30
Item response theory
Item banking
Quality of life

ABSTRACT

*Background:* Health-related quality of life (HRQOL) questionnaires should ideally be adapted to the individual patient and at the same time scores should be directly comparable across patients. This is achievable using a computerised adaptive test (CAT). Basing the CAT on an existing instrument enables measurement within an established HRQOL framework and allows backward-compatibility with studies using the original instrument. Because of these advantages the EORTC Quality of Life Group (QLG) has initiated a project to develop a CAT version of the widely used EORTC QLQ-C30.

*Methods:* We present the EORTC QLG's strategy for developing a CAT. For each dimension of the EORTC QLQ-C30 our approach includes literature search and conceptualisation, formulation of new items, expert and patient evaluations, field-testing, and psychometric analyses of the items. The strategy is illustrated with the initial results of the development of CAT for physical functioning (PF).

*Results:* We identified 975 PF items in the literature. Of these, 407 items were deemed relevant, i.e. measured one of the PF aspects measured by the QLQ-C30. Based on these items

* *Corresponding author:* Address: The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, Bispebjerg Bakke 23, 2400 Copenhagen NV, Denmark. Tel.: +45 3531 2025; fax: +45 3531 2071.
E-mail address: map01@bbh.regionh.dk (M.Aa. Petersen).

[1] Grant National Science Council, Taiwan, No. 95-2314-B-002-266-MY2, 97-2314-B-002-020-MY3.


doi:10.1016/j.ejca.2010.02.011

we developed 86 new items. Review by the EORTC CAT-project group reduced this to 66 items. Based on expert and patient evaluations several items were revised and the list was further reduced to 51 items.

*Conclusions:* Based on the findings for PF, we believe that our approach will generate item pools that are relevant and appropriate for cancer patients. These will form the basis for a backward-compatible CAT assessing the HRQOL dimensions of the EORTC QLQ-C30.



## 1. Introduction

The primary source of information about patients' health-related quality of life (HRQOL) is self-report questionnaires, also termed patient-reported outcomes (PROs). These instruments have typically been developed and implemented using classical methods like sum scoring (adding the scores to individual items into scales). However, these classical methods have some limitations. For example, all patients have to answer the same set of items for sum scores to be comparable. This means that PRO instruments often constitute a compromise between optimal measurement (requiring longer instruments) and reasonable response burden (requiring shorter instruments).

In recent years there has been an increasing interest in methods based on item response theory (IRT)[1,2] for measuring HRQOL (see e.g.[3–10]). IRT is a statistical framework for assessing the characteristics of the items in multi-item scales. IRT models estimate among other things the 'difficulty' of each item. For example, taking a long walk is more demanding/difficult than taking a short walk. Hence, asking patients with good physical functioning about taking a long walk may be highly informative while for patients with poor functioning an item about taking a short walk may be more informative. IRT-based methods take this into account.

One of the major advantages of IRT methods compared to classical methods, and one of the reasons for the interest in these methods, is that when a set of items has been calibrated (estimated) to an IRT model all scores based on *any* subset of the items are on the same metric. That is, even if two patients answer different subsets of items from the same item pool, their scores are directly comparable. This unique feature means that a questionnaire can be adapted to the individual to optimise the measurement properties, yet at the same time comparability of scores are maintained across patients. This possibility for adapting the instrument is fully utilised in a computerised adaptive test (CAT)[11] to construct individualised instruments: based on the responses to the preceding items, a computer programme evaluates which item should be asked next to obtain maximal information. The additional items are administered until a predefined level of information (i.e. precision) has been reached or until a predefined number of items has been administered. Fig. 1 shows two simple examples of how a CAT measuring physical functioning could proceed. If the patient reports few problems on an item, the next item will concern a more demanding task, while if severe problems are reported the following item will concern a less demanding task. In this way the questionnaire is individualised, using the most informative items for each patient. In contrast, traditional questionnaires ask the same items to all, and hence, generally need more items to obtain the same level of precision.

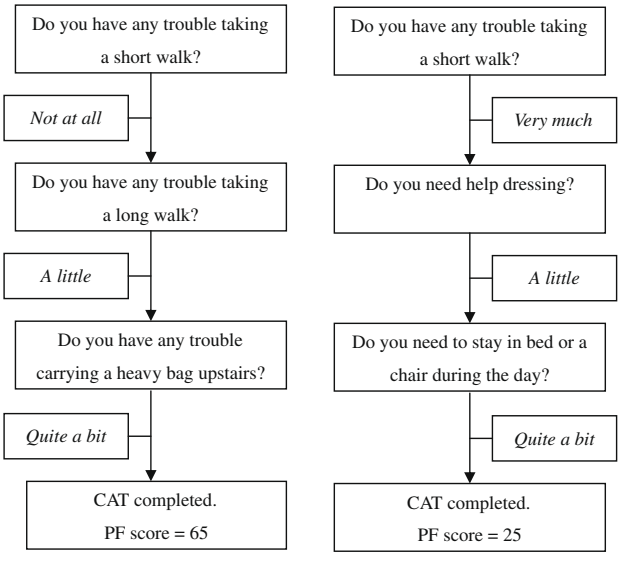


**Fig. 1 – Two examples of how a CAT measurement of physical functioning could proceed.**

CAT measurement has several advantages compared to traditional questionnaires including: increased measurement precision and/or reduced response burden, increased flexibility as the questionnaire can be adapted to the individual study or patient, avoidance of asking uninformative questions, and immediate calculation and presentation of results.

Because of the clear advantages of CAT measurement a number of research groups are developing CAT's for measurement of HRQOL, including the Patient-Reported Outcomes Measurement Information System (PROMIS)[12], QualityMetric[13], and the European Palliative Care Research Collaborative.[14]

The European Organisation for Research and Treatment of Cancer Quality of Life Group (EORTC QLG) has initiated a CAT-project for the HRQOL dimensions measured by the EORTC Quality of Life Questionnaire (EORTC QLQ-C30).[15] The EORTC QLQ-C30 is one of the most widely used quality of life questionnaires in cancer research.[16,17] It consists of 30 items measuring 15 aspects of HRQOL: five functional measures, nine symptom measures and one measure of overall health/quality of life.[18] At present the EORTC QLQ-C30 exists only in a traditional version where all patients are asked the same 30 items.

The aim of the EORTC CAT-project is to measure the same 15 HRQOL dimensions as measured with the QLQ-C30, but

with higher efficiency and precision. This requires new items supplementing the existing items, filling 'gaps', e.g. new items enabling assessment of patients in very poor condition. Initial investigations indicated that it might be difficult to develop enough relevant but different items for 'overall quality of life'. Therefore, the project focuses on developing CAT for the other 14 HRQOL dimensions of the QLQ-C30. Basing the CAT on the QLQ-C30 ensures backward-compatibility with the substantial literature of studies using the QLQ-C30. The CAT instrument aims to measure the same well-validated and well-known HRQOL dimensions with significantly improved precision. Most other CATs do not have such a direct link to an existing instrument. Instead they are developed 'from scratch' and are unconstrained by existing conceptual frameworks, resulting in a completely new instrument. This gives greater freedom in the development process, but may also result in unfamiliar measures.

The purpose of the current paper is to provide an overview of the methods used to develop the new EORTC QLQ-C30 CAT instrument, illustrated by the development of CAT for physical functioning (PF).

## 2. Methods

The development of the EORTC CAT is conducted by members of the EORTC QLG. This international CAT-project group includes oncologists, psychologists, statisticians, and others with considerable experience in developing HRQOL instruments.

The development of the item pools forming the basis for the CAT can be divided into four phases: (1) literature search, (2) operationalisation (selection and formulation of items), (3) pre-testing (patient interviews), and (4) field-testing (data collection and psychometric analyses).

### 2.1. *Phase 1: literature search*

For each dimension the literature search has two overall goals:

1. to gain general knowledge about the dimension,
2. to identify existing instruments and items used to measure the dimension.

The literature search elucidates how the dimension has been conceptualised/defined in the literature and how the QLQ-C30 items fit into such definitions; in particular, whether the dimension seems to be unidimensional or whether it is composed of several subdimensions. If the dimension consists of several subdimensions it should be clarified to which subdimension(s) the QLQ-C30 items belong. As the new items should only measure the same subdimensions, subsequent item development focuses on these subdimensions. The items identified from existing instruments used to measure the dimension form the basis for developing new items covering the relevant subdimensions and levels (e.g. from poor to good physical functioning) of the dimension.

The literature search is primarily conducted on MEDLINE, but other databases and sources (e.g. textbooks, monographs) are also searched when deemed relevant. In addition, the PROQOLID database (http://www.proqolid.org) and the EORTC QLG Item Bank[19] are searched for relevant items.

It should be noted that the intent is not necessarily to carry out a fully comprehensive literature review for each dimension, but rather to identify the most important conceptual frameworks for the dimension and to identify existing, validated items that can form the basis for developing the item pools.

### 2.2. *Phase 2: operationalisation*

Based on the conceptualisation from the previous phase, the list of identified items is trimmed: items measuring subdimensions not covered by the QLQ-C30 items and redundant (closely similar) items are deleted. Items that cannot be reformulated to fit the 'QLQ-C30 item style' (i.e. the response categories 'not at all', 'a little', 'quite a bit' and 'very much'; a one week time frame, etc.) are also deleted.

The 'shortlist' of items forms the basis for formulating new items measuring the relevant subdimensions and fitting the item style. If items for patients at certain levels of the dimension (e.g. patients with severe symptoms) are lacking, new items particularly relevant for these patients are formulated.

This item selection and formulation are carried out independently by (at least) two members of the project group. After each step, possible differences are discussed and a consensus is reached.

The list of developed items is evaluated by at least 10 experts coming from three or more countries. The experts judge whether the items measure the relevant aspects of the dimension, are appropriate, clear and well-formulated, and whether items for some of the relevant aspects are lacking.

### 2.3. *Phase 3: pre-testing*

Next, the revised list of items is evaluated by cancer patients (the primary target group). At least 10 patients from each of at least three countries are interviewed about the items. The interview procedure follows the EORTC QLG guidelines for pre-testing of items.[20] The patient sample should represent the diversity of the target group: different levels of the dimension, sites, age groups, both genders, etc. In order to conduct these interviews, the candidate items are translated into the relevant languages using rigorous forward–backward translation procedures developed by the EORTC QLG.[21] The interviews are intended to elucidate whether patients find some of the items difficult to answer, confusing, annoying, upsetting, intrusive, etc. and hence need to be rephrased or deleted. The interviews also explore the need for additional items to cover important topics which may have been missed.

### 2.4. *Phase 4: field-testing*

The revised list of items is field-tested in at least three countries representing different regions of Europe. To ensure stable calibration of the IRT models and to be able to compare item characteristics across countries, responses from a minimum of 100 (and preferably 200) patients from each country,

with a total of at least 1000 responses will be collected.[22] For the results to be generally applicable to cancer patients, the sample must be as heterogeneous as possible in terms of disease sites, stages, age groups, etc. The patients are asked to complete the new items together with the QLQ-C30. They also complete 'debriefing items' covering similar issues as in the interviews (e.g. whether certain items are too difficult, inappropriate, ambiguous, etc.).

The resulting data set forms the basis for the final evaluation and selection of items and the calibration of the items to the IRT model used for the CAT. The following is a short summary of the methods used for the psychometric evaluation:

1. Descriptive statistics including item mean scores, response frequencies, percent missing responses and correlations with the relevant QLQ-C30 items.
2. Factor analysis for ordinal variables to explore the dimensionality of the set of items.[23,24] Local independence (a requirement for standard IRT models[2]) is investigated in the factor analyses using residual correlations.[2,3,25]
3. Differential item functioning (DIF) analysis using ordinal logistic regression.[26,27] If, for example, men and women answer an item differently even though they are at the same level of the dimension (e.g. have the same level of physical functioning), the item is said to show gender DIF. If an item shows gender DIF it may be problematic to compare scores for men and women. Hence, to be able to compare scores across all groups of patients the items should not exhibit substantial DIF.
4. Calibration of the IRT model. We primarily use the generalised partial credit model (GPCM)[28], but other IRT models may be explored as well; for example, if the GPCM has poor fit or if a more restrictive and robust model (like the partial credit model) seems to fit. Model fit is investigated using Muraki's test for item fit[28] and comparisons of observed and expected item responses.

The results of these analyses are used to identify and discard items that exhibit unacceptable levels of DIF, do not fit the IRT model, etc.

## 3. Results

We illustrate the methodology using the results of the first three phases of the development of the item pool for PF. The field-testing is in progress and the results of this and the psychometric analyses (phase 4) will be described in detail in a separate paper.

### 3.1. *Phase 1: literature search*

The literature search revealed few thoroughly developed conceptual frameworks for PF. We based our conceptualisation of the QLQ-C30 PF on the framework of Stewart and Kamberg who defined PF as '... the performance of or capability to perform a variety of physical activities ...' such as 'bathing, dressing, walking, bending, climbing stairs, and running.'[29] and on the World Health Organization's International Classification of Functioning, Disability and Health (ICF).[30] Using the ICF subdimensions (categories) we classified the QLQ-C30 PF items as shown in Table 1. We concluded that QLQ-C30 PF was composed of the four subdimensions described in the table and therefore, focused our item selection and development on these four subdimensions.

In collaboration with the European Palliative Care Research Collaborative, we included an early version of the list of items they had identified in the literature when developing a PF CAT for use in palliative care.[31] This list consisted of 946 items. A review of the PROQOLID database identified 24 additional PF items and five additional items were identified from the EORTC QLG Item Bank. In all 975 PF items were identified.

### 3.3. *Phase 2: operationalisation*

We classified the identified items into one of the ICF categories listed in Table 1 or an 'other' category if the item was judged to measure something other than one of the four relevant subdimensions. The two reviewers agreed that 454 items measured one of the relevant subdimensions, while 521 items measured something else and were discarded. Of the 454 items kept, 47 were deleted because of redundancy

**Table 1 – QLQ-C30 PF items and our classification of the items (based on WHO's ICF[32]).**

| Item text | ICF category | Description of category |
|---|---|---|
| Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase? | Lifting and carrying objects | Raising up an object or taking something from one place to another |
| Do you have any trouble taking a *long* walk? | Walking and moving | Moving the whole body from one place to another like walking, running, and climbing stairs |
| Do you have any trouble taking a *short* walk outside of the house? | Walking and moving | |
| Do you need to stay in bed or a chair during the day? | Mobility, unspecified | Limitations in mobility in general, unspecific terms |
| Do you need help with eating, dressing, washing yourself or using the toilet? | Self-care | Caring for oneself, washing and drying oneself, caring for one's body and body parts, dressing, eating and drinking |

*Note:* all five items have the response categories: 'not at all', 'a little', 'quite a bit', and 'very much'.

or because they could not be reformulated into the QLQ-C30 item style. Although differing in wording, many of the remaining 407 items covered the same activity, e.g. walking, running, climbing stairs or dressing. We grouped the items according to these activities. We used the items in each activity-group as inspiration for formulating new items covering the activity and complying with the QLQ-C30 item style. To begin with we formulated 86 new items. After evaluations by the members of the project group the list of items was reduced to 66. These items were judged to cover the relevant subdimensions and the different levels of PF (from poor to good PF) satisfactorily.

The 66 items were evaluated by 10 experts from Denmark, Germany, The Netherlands, and the UK. The expert evaluations resulted in rewording of 12 items and deletion of 11 items: four because of redundancy, three were judged difficult to understand/answer, two could not be translated into an equivalent wording in some languages, and two items were judged to measure something other than the relevant subdimensions of PF.

### 3.4. *Phase 3: pre-testing*

A total of 43 patients were interviewed about the 55 remaining candidate items and the five original QLQ-C30 PF items. Clinical characteristics of the patients are shown in Table 2. Based on the interviews we changed the wording of 12 items to make them clearer, unambiguous, and more consistent. Four items were deleted: two because several patients found them difficult to understand/answer, one because of redundancy, and one because the item text did not fit well with the response options and it could not be reformulated to do so without being too similar to some of the other items. Deleting the four items did not compromise the content coverage. The patients did not suggest any additional items regarding the relevant subdimensions.

After the first three phases 51 new candidate items remained for the PF item pool (see Appendix 1) together with the five QLQ-C30 PF items. These will be field-tested and the final item selection and the IRT calibration will be conducted.

**Table 2 – Clinical characteristics of the 43 patients interviewed in phase 3.**

| | N/mean |
|---|---|
| Gender | |
| Men | 19 (44%) |
| Women | 24 (56%) |
| Country | |
| Denmark | 11 (26%) |
| France | 12 (28%) |
| Germany | 10 (23%) |
| UK | 10 (23%) |
| Age (mean years) | 58 (range 27–88) |
| Cancer stage | |
| I–II | 5 (12%) |
| III–IV | 31 (72%) |
| Unknown | 7 (16%) |
| Cancer site | |
| Breast | 10 (23%) |
| Gastrointestinal | 6 (14%) |
| Urogenital | 5 (12%) |
| Gynaecological | 5 (12%) |
| Head and neck | 2 (5%) |
| Prostate | 2 (5%) |
| Other | 5 (12%) |

## 4. Discussion

CAT measurement has several advantages compared to traditional questionnaire measurement. In particular, it can improve measurement precision without increasing the response burden for the patients. To utilise these new methods to improve the measurement of cancer patients' HRQOL the EORTC QLG has initiated a project to develop CAT for the widely used EORTC QLQ-C30 questionnaire.

The EORTC CAT development can be divided into four phases: literature search, operationalisation, pre-testing, and field-testing. These phases are closely related to the phases of EORTC QLG module development.[20] Here we have reported the results of the first three phases of the item pool development for physical functioning. Our development procedure was feasible and useful for developing items for assessing PF. The literature search yielded valuable insights into how PF has been defined, conceptualised and measured, and how the EORTC measurement fits into these definitions. This was a useful inspiration for formulating items measuring PF in the 'EORTC QLQ-C30 way'. The expert and patient evaluations were invaluable in identifying problematic items, and in optimising item wording. We believe that our developmental procedure will generate item pools that are relevant and appropriate for use with cancer patients.

We are working in parallel on the development of item pools for all QLQ-C30 dimensions (except overall quality of life). As each of the item pools for the various QLQ-C30 dimensions is completed, we will develop and implement the appropriate CAT, ultimately resulting in a complete CAT version of the original questionnaire.

By explicitly and systematically basing the CAT development on a single, existing instrument our approach differs from that of most other research groups developing CAT for HRQOL measurement. Basing the CAT on an established HRQOL instrument as the QLQ-C30 has several advantages including backward-compatibility with a huge literature and measurement of well-validated and well-known HRQOL dimensions. It also simplifies the conceptual part of the development because we know what to measure (the same as the existing instrument) and do not have to establish a whole new framework of measurement. However, as a consequence the new items are required to conform to a similar layout, response options, timeframe, etc. and have to measure the same aspects of HRQOL as the existing items do. In some cases this limits the number of applicable items it is possible to devise. Still, for an instrument such as the QLQ-C30 with few items for each dimension, CAT may significantly improve the measurement. Furthermore, such a CAT may serve as a platform for developing further CAT instruments which in addition to being compatible with the QLQ-C30, may cover other HRQOL dimensions, other time frames, utilise

other innovative item designs, etc. A platform it is feasible to develop within a foreseeable timeframe and which in itself will improve the precision and efficiency of the assessment of the HRQOL of cancer patients significantly.

## Sources of support

The study was funded by grants from the EORTC Quality of Life Group.

## Conflict of interest statement

None declared.

## Acknowledgement

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

## Appendix 1

The 51 candidate items for the EORTC PF CAT item pool after the pre-testing (phase 3). Note that this is not the final list of items, but the list of items for the field-testing (phase 4).

Do you have any trouble lifting a full cup or glass to your mouth?
Do you need help to walk about indoors (e.g. a walking stick or someone to support you)?
Do you have any trouble walking 100 m?
Do you need help caring for your feet (e.g. cutting your toenails)?
Do you have any trouble walking around indoors?
Do you have any trouble walking up a flight of stairs?
Do you need help with grooming (e.g. cleaning nails, brushing teeth, combing your hair)?
Do you have any trouble walking down a flight of stairs?
Do you have any trouble walking over uneven ground such as grass or gravel?
Do you have any trouble walking for 30 min?
Do you have any trouble walking 100 m on level ground?
Do you need help putting on or taking off trousers or a skirt?
Do you need help taking a shower?
Do you have any trouble lifting a box weighing about 10 kg?
Do you have any trouble hiking 3 km on uneven surfaces?
Do you have any trouble participating in strenuous sports, like running 10 km, cycling 25 km, or a similar activity?
Do you need help dressing?
Do you need help putting on your shoes?
Do you have any trouble taking a long, brisk walk (more than 30 min)?
Do you have any trouble walking outdoors on flat ground?
Do you need help brushing your teeth?
Do you have any trouble walking 500 m?
Do you have any trouble climbing three flights of stairs?
Do you have any trouble walking a few steps?
Do you have any trouble lifting a box weighing about 10 kg and carrying it for 1 min?
Do you need help pulling on a sweater?
Do you need help putting on a shirt?
Do you have any trouble walking 100 m while carrying a heavy shopping bag in each hand?
Do you have any trouble reaching and getting down an object weighing about 2 kg (such as a bag of flour) from just above your head?
Do you have any trouble bending over to pick up a light object from the floor?
Do you have any trouble running 1 km?
Do you have any trouble running fast?
Do you have any trouble walking 10 m inside?
Do you have any trouble taking a brisk walk?
Do you have any trouble walking without losing your balance?
Do you have any trouble walking 1 km?
Do you have any trouble carrying a heavy bag upstairs?
Do you need help washing your face and hands?
Do you need help to go to the lavatory?
Do you have any trouble walking 10 m outside your home?
Do you have any trouble running 100 m?
Do you have any trouble taking a long walk carrying a heavy pack on your back (e.g. a filled rucksack)?

**Appendix 1** (*continued*)

| |
|---|
| Do you have any trouble running up three flights of stairs without a rest? |
| Do you have any trouble carrying something weighing about 5 kg? |
| Do you need help to walk about outside (e.g. a walking stick or someone to support you)? |
| Do you need help washing and drying your whole body? |
| Do you need help undressing? |
| Do you have any trouble carrying something in both hands (e.g. shopping bags) while climbing a flight of stairs? |
| Do you need help eating? |
| Do you have any trouble lifting a full teapot/coffeepot? |
| Do you have any trouble running a short distance, such as to catch the bus? |

## REFERENCES


1. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Newbury Park: Sage Publications, Inc.; 1991.
2. van der Linden WJ, Hambleton RK. *Handbook of modern item response theory*. Berlin: Springer-Verlag; 1997.
3. Bjorner JB, Kosinski M, Ware Jr JE. 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* 2003;**12**(8):913–33.
4. Bjorner JB, Petersen MAa, Groenvold M, et al. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional function scale. *Qual Life Res* 2004;**13**(10):1683–97.
5. Hahn EA, Cella D, Bode RK, Gershon R, Lai JS. Item banks and their potential applications to health status assessment in diverse populations. *Med Care* 2006;**44**(11 Suppl. 3):S189–97.
6. Kopec JA, Sayre EC, Davis AM, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. *Health Qual Life Outcomes* 2006;**4**(1):33.
7. Petersen MAa, Groenvold M, Aaronson N, et al. Item response theory was used to shorten EORTC QLQ-C30 scales for use in palliative care. *J Clin Epidemiol* 2006;**59**:36–44.
8. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH roadmap cooperative group during its first two years. *Med Care* 2007;**45**(5 Suppl. 1):S3–S11.
9. Cook KF, Teal CR, Bjorner JB, et al. IRT health outcomes data analysis project: an overview and summary. *Qual Life Res* 2007.
10. Orlando-Edelen M, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res* 2007;**16**(Suppl. 1):5–18.
11. Wainer H. *Computerized adaptive testing: a primer*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2000.
12. The Patient-Reported Outcomes Measurement Information System (PROMIS). <http://www.nihpromis.org> [accessed June 2009].
13. Dynamic Computer Adaptive (DYNHA®) System. <http://www.qualitymetric.com/products/dynhadetails.aspx> [accessed June 2009].
14. Kaasa S, Loge JH, Fayers P, et al. Symptom assessment in palliative care: a need for international collaboration. *J Clin Oncol* 2008;**26**(23):3867–73.
15. Aaronson NK, Ahmedzai S, Bergman B, et al. 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* 1993;**85**(5):365–76.
16. Fayers P, Bottomley A. Quality of life research within the EORTC – the EORTC QLQ-C30. European Organisation for Research and Treatment of Cancer. *Eur J Cancer* 2002;**38**(Suppl. 4):S125–33.
17. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ* 2002;**324**(7351):1417–9.
18. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A. *The EORTC QLQ-C30 scoring manual*. Brussels: European Organisation for Research and Treatment of Cancer; 2001.
19. EORTC Item Bank Guidelines. <http://groups.eortc.be/qol/downloads/200104itembank_guidelines.pdf> [accessed June 2009].
20. Blazeby JM, Sprangers MA, Cull A, Groenvold M, Bottomley A. *EORTC quality of life group – guidelines for developing questionnaire modules*. Brussels: European Organisation for Research and Treatment of Cancer; 2002.
21. Dewolf L, koller M, Velikova G, Johnson C, Scott N, Bottomley A. *EORTC quality of life group translation procedure*. Brussels: European Organization for Research and Treatment of Cancer; 2009.
22. Muraki E, Bock RD. *PARSCALE – IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago: Scientific Software International, Inc.; 1996.
23. Muthen B. A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika* 1984;**49**(1):115–32.
24. Muthen LK, Muthen BO. *Mplus user's guide*. Los Angeles, CA: Muthen & Muthen; 2002.
25. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). *Qual Life Res* 2005;**14**(10):2277–91.
26. French AW, Miller TR. Logistic regression and its use in detecting differential item functioning in polytomous items. *J Educ Meas* 1996;**33**(3):315–32.
27. Petersen MAa, Groenvold M, Bjorner JB, et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Qual Life Res* 2003;**12**(4):373–85.
28. Muraki E. A generalized partial credit model. In: van der Linden WJ, Hambleton RK, editors. *Handbook of modern item response theory*. Berlin: Springer; 1997. p. 153–68.
29. Stewart AL, Kamberg CJ. Physical functioning measures. In: Stewart AL, Ware JE, editors. *Measuring functioning and well-being*. London: Duke University Press; 1992. p. 86–101.
30. International Classification of Functioning, Disability and Health (ICF). Retrieved from <http://www.who.int/classifications/icf/en/>; July 2009.
31. Helbostad JL, Holen JC, Jordhoy MS, Ringdal GI, Oldervoll L, Kaasa S. A first step in the development of an international self-report instrument for physical functioning in palliative cancer care: a systematic literature review and an expert opinion evaluation study. *J Pain Symptom Manage* 2009;**37**(2):196–205.
32. ICF browser. Retrieved from <http://apps.who.int/classifications/icfbrowser/>; July 2009.