The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency

Petersen, Morten Aa; Aaronson, Neil K; Arraras, Juan I; Chie, Wei-Chu; Conroy, Thierry; Costantini, Anna; Giesinger, Johannes M; Holzner, Bernhard; King, Madeleine T; Singer, Susanne; Velikova, Galina; Verdonck-de Leeuw, Irma M; Young, Teresa; Grønvold, Mogens; EORTC Quality of Life Group

Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2012.09.010

Publication date:
2013

Document version
Early version, also known as pre-print

Journal of Clinical Epidemiology 66 (2013) 330−339

Journal of Clinical Epidemiology

The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency

Morten Aa. Petersen[a,*], Neil K. Aaronson[b], Juan I. Arraras[c], Wei-Chu Chie[d], Thierry Conroy[e], Anna Costantini[f], Johannes M. Giesinger[g], Bernhard Holzner[g], Madeleine T. King[h], Susanne Singer[i], Galina Velikova[j], Irma M. Verdonck-de Leeuw[k], Teresa Young[l], Mogens Groenvold[a,m], on behalf of the EORTC Quality of Life Group

[a]*The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, Bispebjerg Bakke 23, 2400 Copenhagen NV, Denmark*
[b]*Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands*
[c]*Department of Medical Oncology, Hospital of Navarre, C/Irunlarrea 3, ES-31008, Pamplona, Spain*
[d]*Graduate Institute of Preventive Medicine and Department of Public Health, College of Public Health, National Taiwan University, Roosevelt Road, Taipei, Taiwan (R.O.C.)*
[e]*Department of Medical Oncology, Centre Alexis Vautrin, 6 Avenue de Bourgogne, F-54500 Vandoeuvre-lès-Nancy, France*
[f]*Department of Oncological Sciences, 2nd Faculty of Medicine, Sant'Andrea Hospital, University of Rome, Via di Grottarossa 1035, Rome, Italy*
[g]*Department of Psychiatry and Psychotherapy, Innsbruck Medical University, Anichstr. 35, A-6020 Innsbruck, Austria*
[h]*Quality of Life Office, Psycho-oncology Co-operative Research Group, School of Psychology, Brennan MacCallum Building (A18), University of Sydney, Sydney, Australia*
[i]*Department of Epidemiology and Health Care Research, Johannes Gutenberg University Mainz, Saarstr. 21, D 55122 Mainz, Germany*
[j]*Cancer Research UK Centre, University of Leeds, Woodhouse Lane, Leeds, UK*
[k]*Clinical Psychology, VU University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands*
[l]*Lynda Jackson Macmillan Centre, Mount Vernon Hospital, Rickmansworth Road, Northwood, Middlesex, UK*
[m]*Institute of Public Health, University of Copenhagen, CSS, Oester Farimagsgade 5, Copenhagen, Denmark*

Accepted 10 September 2012

Abstract

Objectives: The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group is developing a computer-adaptive test (CAT) version of the EORTC Quality of Life Questionnaire (QLQ-C30). We evaluated the measurement properties of the CAT versions of physical functioning (PF) and fatigue (FA) and compared these with the corresponding QLQ-C30 scales.

Study Design and Setting: Based on international samples of more than 1,000 cancer patients, we simulated CAT administration of varying numbers of items and compared the resulting scores with those based on all items in the respective item pools. Furthermore, the relative validity (RV) of CATs was compared with that of the QLQ-C30 scales using known groups validity.

Results: For both dimensions, CATs of all lengths resulted in unbiased score estimates. CATs consisting of five or more items had reliability >0.90, correlated ≥ 0.97 with the full scale, and had root mean square error <0.25. The average RVs for these CATs ranged 1.02−1.33, indicating possible savings in sample size requirements of 3−42% using CAT.

Conclusion: The CAT versions of PF and FA exhibited high levels of measurement precision and efficiency. The potential savings in sample size requirements using CATs compared with those using the original QLQ-C30 scales were typically 20% or more.

Keywords: Computer-adaptive test; EORTC QLQ-C30; Fatigue; Physical functioning; Quality of life; Relative validity

1. Introduction

With the widespread access to and use of computers, tablets, smartphones, and the Internet, the assessment of patient-reported outcomes (PROs) is increasingly carried out electronically. Computer-adaptive testing (CAT) is a sophisticated method for assessing PROs electronically [1,2]. CAT tailors the item set to the individual patient. This is achieved by repeatedly estimating the patient's symptom

Funding: The study was funded by grants from the EORTC Quality of Life Group. The work of J.M.G. was funded by a grant from the Austrian Science Fund #502. The work of W.-C.C. was supported by grant National Science Council, Taiwan, No. 95-2314-B-002-266-MY2, 97-2314-B-002-020-MY3.

Conflict of interest statement: There were no financial relationships, personal relationships, academic competition, intellectual commitments, or other conflicts of interest that might have biased the work.

[*] Corresponding author. Tel.: +45-3531-2025; fax: +45-3531-2071.
E-mail address: mpet0009@bbh.regionh.dk (M.Aa. Petersen).

0895-4356/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jclinepi.2012.09.010

What is new?

- The European Organisation for Research and Treatment of Cancer (EORTC) computer-adaptive test (CAT) instrument being developed measures physical functioning (PF) and fatigue (FA) with high levels of measurement precision and efficiency
- The potential savings in sample size requirements in health-related quality of life studies using CAT measures compared with using the original EORTC Quality of Life Questionnaire scales were typically 20% or more
- The CAT instrument may improve the EORTC measurement of PF and FA

or functional level based on responses to previous questions and then selecting and presenting the most appropriate item for that symptom/functional level. CAT has several theoretical advantages including higher measurement precision and/or reduced response burden compared with traditional fixed-length measures requiring that all patients respond to the same set of questions.

The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group is currently developing a CAT version of the EORTC Quality of Life Questionnaire (QLQ-C30) [3], one of the most widely used health-related quality of life (HRQOL) questionnaires in cancer research [4,5]. The aim was to construct a more precise, efficient, and flexible instrument that will allow for the precise measurement of individuals, adaptation to different patient populations, and so forth. Once this developmental work is completed, the resulting CAT version can be used as an alternative to QLQ-C30 during a transition period. In the long run, the CAT version may preempt the original QLQ-C30 as the primary core EORTC quality of life instrument. Note that, as the QLQ-C30 scales are short (mostly just one or two items), for most QLQ-C30 dimensions, we do not expect that the new instrument will result in shorter scales rather in better and more precise measurement. The first two EORTC CAT item banks that have been developed cover physical functioning (PF) and fatigue (FA) [6–9].

Although, theoretically, CAT has clearly superior measurement properties compared with traditional measures such as the fixed-length sum scales of QLQ-C30, it will differ across instruments, dimensions, and patient populations how these superior measurement properties translate into practical advantages in conducting PRO research. Using CAT measurement, the number of items (the length of the questionnaire) is selected for each study. This choice usually involves a trade-off between speed (few items) and precision (many items). Hence, information about measurement efficiency and precision (i.e., how reliable and valid the CAT is with a given number of items) is vital to be able to optimize CAT for a specific study. Furthermore, it may be of particular interest for many users of QLQ-C30 to know whether they can expect savings in study time and/or expenses using the CAT version rather than the existing and familiar fixed-length and fixed-format versions. Evaluation based on the CAT item banks for PF and FA may give valuable information about the measurement quality of the EORTC QLQ-C30 CAT and what may be gained from using CAT.

The aims of the present study were to assess the (1) measurement precision/efficiency of the CAT versions of the EORTC QLQ-C30 PF and FA scales, the first two CATs that have been developed for this questionnaire and (2) potential reduction in sample size requirements using various CAT versions compared with using the original QLQ-C30 PF and FA scales.

2. Methods

2.1. Development of the EORTC CAT item pools

The aim of EORTC CATs is to measure the same HRQOL dimensions as measured with QLQ-C30 but with higher efficiency and precision. For each dimension, the item pool development can be divided into four phases: (1) literature search to gain knowledge about the dimension and identify existing items used to measure the dimension; (2) based on (1) to formulate new items measuring the relevant aspects of the dimension and following the item style of QLQ-C30; (3) interviewing cancer patients from at least three countries to evaluate the content, formulation, and so forth of the items; and (4) finally, collecting responses to the candidate items from at least 1,000 patients which will form the basis for the psychometric analyses and final selection and item response theory (IRT) calibration of items for the pools. Using these developmental steps, we have constructed a PF item pool of 31 items and an FA pool of 34 items. For further details on the development, please see refs. [6–9].

2.2. Sample

For the development of the PF and FA item pools, we had collected responses to the candidate items from 1,176 and 1,321 cancer patients coming from six and eight countries, respectively [7,9] (see Table 1 for details). These two samples form the basis for the analyses reported in the present article.

2.3. Item pools

The PF item pool consists of 31 items (including the five QLQ-C30 PF items), and the FA pool consists of 34 items (including the three QLQ-C30 FA items). All items use

Table 1. Sociodemographic and clinical characteristics of the two analytic samples (N_{PF} = 1,176 and N_{FA} = 1,321)

Characteristics	PF sample, *N*/mean	FA sample, *N*/mean
Age (yr), mean (range)	58 (18–91)	59 (18–99)
Gender (%)		
Male	524 (45)	537 (41)
Female	648 (55)	778 (59)
Country (%)		
Australia	—	122 (9)
Austria	—	183 (14)
Denmark	412 (35)	340 (26)
France	314 (27)	209 (16)
Germany	163 (14)	100 (8)
Italy	87 (7)	—
The Netherlands	—	98 (7)
Spain	—	85 (6)
Taiwan	100 (9)	—
UK	100 (9)	184 (14)
Education, yr (%)		
0–10	315 (27)	243 (18)
11–13	265 (23)	403 (31)
14–16	280 (24)	334 (25)
>16	281 (24)	307 (23)
Work (%)		
Working	389 (33)	418 (32)
Retired	557 (47)	624 (47)
Other	212 (18)	250 (19)
Cohabitation (%)		
Living with a partner	844 (72)	931 (71)
Living alone	305 (26)	369 (28)
Cancer stage (%)		
I–II	399 (34)	612 (46)
III–IV	583 (50)	538 (41)
Cancer site (%)		
Breast	150 (13)	299 (23)
Gastrointestinal	135 (11)	191 (15)
Gynecological	180 (15)	167 (13)
Head and neck	163 (14)	113 (9)
Lung	52 (4)	87 (7)
Urogenital	181 (15)	150 (11)
Other	124 (11)	306 (23)
Current treatment (%)		
Chemotherapy	443 (38)	558 (42)
Other treatment	97 (8)	248 (19)
No current treatment	605 (52)	511 (39)

Abbreviations: PF, physical functioning; FA, fatigue.

a 4-point response scale: "not at all," "a little," "quite a bit," and "very much." The PF items do not refer to a specific time frame but ask about performing a task generally, whereas the FA items use a "during the past week" recall period. Based on the information functions, both item pools were found to provide highly reliable measurement for wide ranges of PF and FA, about 3.0–3.5 SD units [7,9].

2.4. Evaluation of measurement precision

The evaluation of measurement precision was based on the observed responses of the patients from the two samples. From each patient's responses to the items in an item pool, we simulated how a CAT administration would have proceeded, assuming that the patients would have answered the questions in the same way had they responded to a CAT version of the questionnaire. Close agreement between responses to computerized and conventional paper questionnaires has been found [10].

For each of the PF and FA item pools, we evaluated the measurement precision of all possible "fixed-length" CATs, varying from 1 item to all −1 items. Other stopping rules than the number of items asked (e.g., fixed information) may be used in CAT measurement. Here, we have focused on fixed-length CATs as these allow for the simplest and most direct comparison of the measurement properties of CATs and the standard fixed-length QLQ-C30 scales. The θ estimates based on these CATs were compared with those using all items in each item pool. That is, we used the full-length θ as the "gold standard." The first item used for each CAT version was the item that provided the most information at the prior mean, that is, at 0. At each step of the CAT procedure, the item with the maximum information at the current θ estimate was selected. The θ was estimated using expected a posteriori (EAP) estimation [11].

2.5. Evaluation of statistical power

We used the method of known groups comparison [12] to evaluate the statistical power of CATs in detecting group differences compared with that of the two original QLQ-C30 scales. These comparisons yielded information on the potential savings in sample size requirements if one were to use the CAT measures instead of the original QLQ-C30 PF and FA scales. We used two-sample *t*-test sizes to calculate the relative validity (RV) [12] of the CAT measures compared with the original QLQ-C30 scales.

We conducted two types of known groups comparisons: one based on the observed data and one based on the simulated data. For the analyses based on the observed data, we posed a priori hypotheses based on differences (definitely) expected in PF and FA as a function of various patient characteristics (Table 1). Specifically, we hypothesized that (1) there would be significant differences in PF and FA across age (i.e., older patients would have significantly worse PF and FA than younger patients, with age divided at the median of 60 years); (2) patients with stage III or IV disease would have significantly worse PF and FA than those with stage I or II disease; (3) patients who were not employed would have significantly worse PF than those who were employed; and (4) finally, patients undergoing chemotherapy at the time of questionnaire completion would be significantly more fatigued than those not receiving chemotherapy.

RV was calculated when at least one measure (the CAT version or original QLQ-C30 scale) yielded statistically significant group differences ($P < 0.05$), as hypothesized. Based on these RVs, we estimated the potential savings in sample size requirements based on the CAT versions, to detect an effect size (ES) of 0.5, with power of 0.80, and α set at 0.05. The expected savings for any combination of power and ES will be similar to those presented here, except in very extreme cases with very low power or high ES.

Although we expected that the aforementioned groups would differ, we did not know whether this would be the case or size of such differences. Therefore, in addition to the known group comparisons based on actual data, we also evaluated known groups validity based on simulated data. Specifically, we simulated responses to the items in the two-item pools based on θ's sampled from $N(\varepsilon,1)$, with different mean values ε. From these simulated responses, we derived fixed-length CAT measurement and calculated the QLQ-C30 scale scores. As was the case with the real data, we compared the ability of these simulated CAT and QLQ-C30 scale scores to detect group differences using *t*-tests and RVs.

Simulating the responses using the IRT models forming the basis for CATs might favor CATs. Therefore, to make the comparisons "fair," we divided the full-length θ estimates in the observed samples into groups of approximately 50 patients, and in each θ group, we calculated the distribution of responses to each item. For each simulated θ, we then randomly generated item responses based on the distribution of responses in the θ group to which it belonged. In this way, the simulated responses depended on the IRT model only through the use of θ estimates to construct the relevant response distributions from which to simulate.

We compared groups of size $N_1 = N_2 = 25$, 50, and 100, respectively, and true ESs of 0.2, 0.5, and 0.8, respectively. For each of these $3 \times 3 = 9$ possible settings, we ran 2,000 simulations. In addition, we evaluated the type I error rate by sampling groups from the same distribution, that is, $ES = 0$. For each setting, we calculated the percentage simulations with $P < 0.05$ (power) and the average RVs across the 2,000 simulations. From the resulting RVs, we estimated the potential savings in sample size requirements using CAT. In the interest of space and clarity, we report here only the results based on CATs of length 3, 5, 10, and maximum (i.e., all) items.

The simulations of CAT administration were performed using the Firestar program [13]. All the other statistical analyses were performed using SAS v. 9.1.3 (SAS Institute Inc., Cary, NC, USA) [14].

3. Results

3.1. Evaluation of measurement precision

Fig. 1 shows the median and percentiles for the differences between θ's estimated with CATs of 1,2, …, all −1 items, respectively, and the full-length θ. For both PF and FA, the median differences were very close to 0 for all CAT lengths. The percentiles indicated, however, that for very short CATs, there were some deviations for most patients. For example, when only two items were used, the CAT scores deviated about 0.2 or more for 50% of the patients. However, with five or more PF items, the CAT scores deviated less than 0.1 for 50% of the patients. For FA, a similar level of precision required about 10 items. The θ estimates ranged about five points for both PF and FA, indicating that a deviation of 0.1 is about 2% of the possible score range.

Correlations and root mean square errors (RMSEs) of the θ's estimated using CAT and the full-length θ are shown in Fig. 2. For both PF and FA, all CAT lengths correlated $\geq$0.85 with the full-length θ; using three or more items yielded correlations $\geq$0.95. The RMSEs ranged from 0.00 to 0.52. With five or more items, the RMSEs were $<$0.25 (corresponding to 5% of the total θ range).

Fig. 3 shows the average reliability of CATs in the two samples (calculated as $1 - \text{mean (SEM)}^2$ [15]). When using one item only, the reliabilities were low ($<$0.50); but with just two items, the reliabilities were $\geq$0.75; and with five or more items, the reliabilities of CATs were $\geq$0.88. The reliabilities of the full-length θ estimates were 0.94 (PF) and 0.96 (FA). Thus, a CAT of five items had a reliability of more than 90% of the reliability obtained when using all items.

3.2. Evaluation of statistical power

Fig. 4 shows the average RVs and relative sample size requirements using CAT compared with those using the original QLQ-C30 scales based on the observed data. Contrary to expectations, we did not observe a significant difference between younger and older patients for any of the FA measures. Therefore, the results for FA were based on the comparisons of cancer stage and chemotherapy use only. For PF, the results indicated that there might be a considerable increase in the validity/reduction in sample size requirements when using CAT. With just two items, PF CAT was estimated to have the same power as the original QLQ-C30 scale with only about 70% of the sample size. Using five items, the sample size requirements could be reduced to about 60%. Using more than five items in CAT did not seem to result in further reductions in sample size requirements.

For FA, the estimated savings using CAT were considerably smaller. CATs of four or fewer items were estimated to require larger sample sizes to obtain the same power as the original QLQ-C30 scale. With six or more items, sample size requirements were estimated to be 12−22% lower using CAT than using the original QLQ-C30 scale. Increased savings in sample size requirements were noted with longer CATs, up to CATs of about 20 items.

Based on the simulated responses, PF CATs and the original QLQ-C30 scale had estimated type I error rates ranging between 0.045 and 0.054, that is, very close to the expected 0.05. For FA, the type I error rates ranged from 0.040 to 0.047, that is, slightly below the expected 0.05 (detailed results not presented).

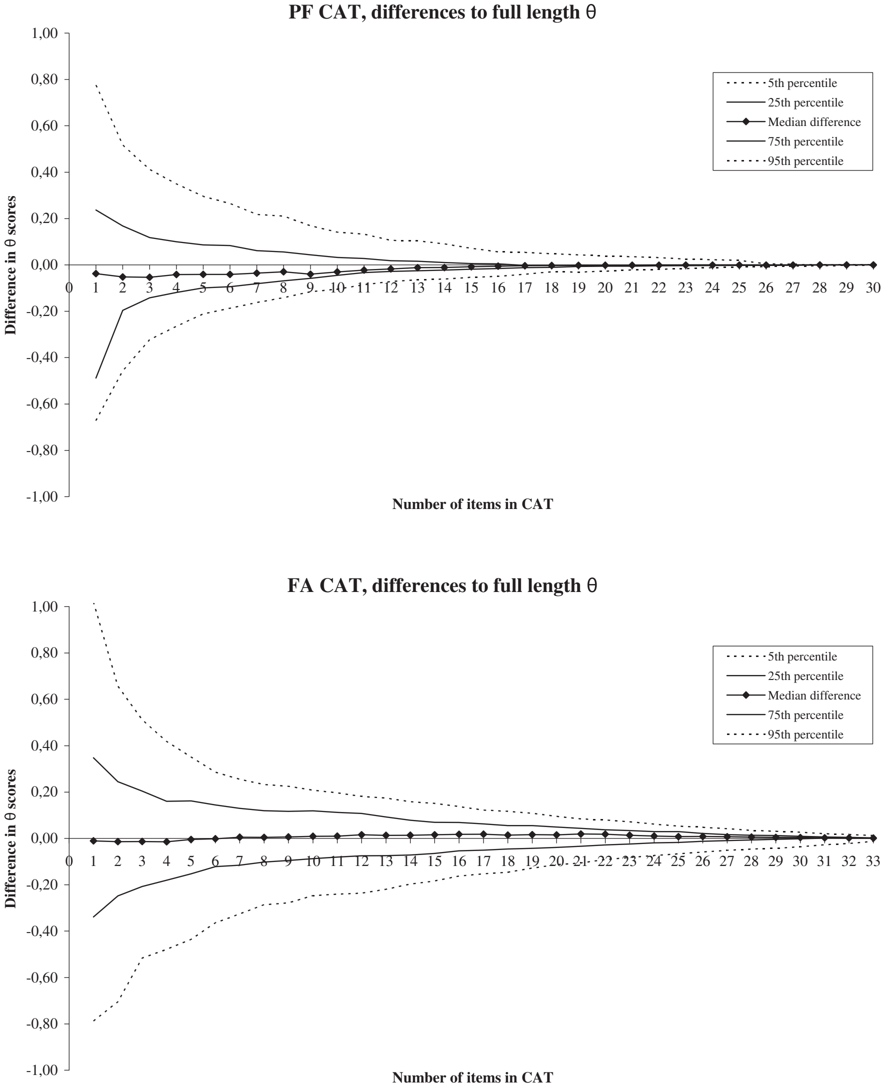

Fig. 1. Median and percentiles for differences between θ based on fixed-length CATs and full-length θ for PF and FA, respectively. CAT, computer-adaptive test; PF, physical functioning; FA, fatigue.

Table 2 summarizes the known groups validity testing based on the simulations. Contrary to the results based on the observed responses, the results for these simulations were very similar for PF and FA (usually within a few percent). Therefore, the table shows the average results across the two dimensions. The table displays the power (i.e., the magnitude of group differences detected) obtained using the QLQ-C30 scales and CATs with 3, 5, 10, and all items, as well as the respective RVs and sample size requirements. An average of the results across the various settings is shown at the bottom of the table. Across all combinations of ES and sample size, the simulations estimated an increase in power using CAT of any length compared with those using the QLQ-C30 scales. The power generally increased with the length of CAT. However, there were only small gains from using more than five items. All the measures had low power to detect an ES of 0.2 for the studied sample sizes. This was also the case when using the full-length θ's, in which groups of 100 patients resulted in a power of 35% to detect ES = 0.2. At the other end of the spectrum, all measures had power ≥95% to detect ES = 0.8 with groups of 50 or more patients. The increased power using CAT resulted in estimated savings in sample size requirements of 12–28%. On average, the savings ranged from 15% using CATs with three items to 22% using CATs with 10 or more items. The increased power using longer CATs was relatively small. On average, obtaining a given power required less than 3% larger samples using a five-item CAT instead of using all items.

4. Discussion

One of the most important rationales advocating the use of CAT is that by tailoring the item set to the individual patient, more precise estimates of the patient's symptom

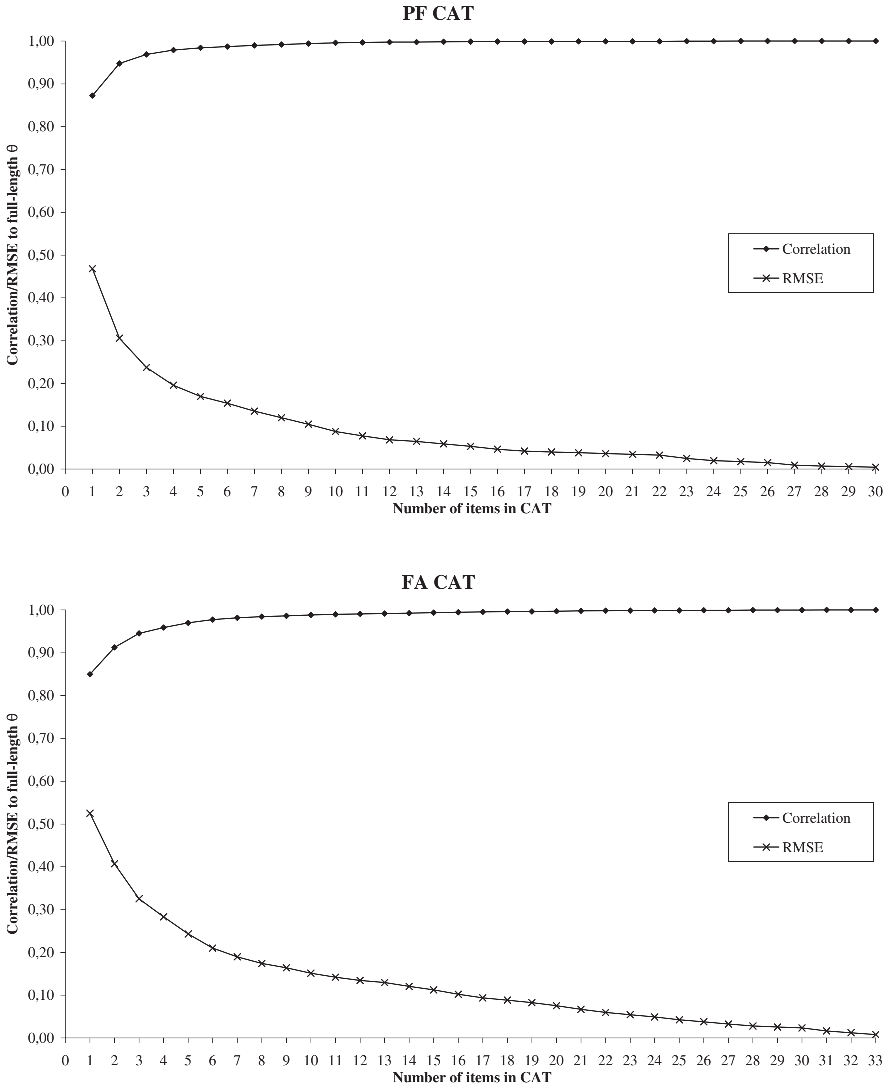

Fig. 2. Correlations and RMSEs of θ's based on fixed-length CATs to full-length θ for PF and FA, respectively. CAT, computer-adaptive test; PF, physical functioning; FA, fatigue; RMSE, root mean square error.

burden, functional capacity, or health status can be obtained. To evaluate the precision and efficiency of the EORTC CAT measures of PF and FA, we compared scores obtained using CATs of varying lengths with the full-length scores based on all items. These evaluations confirmed that the CAT measures can be highly efficient and reliable: with just three items ($<$10% of the item pools), the CAT scores correlated $\geq$0.95 with the full-length θ's; with five items (about 15% of the item pools), the reliabilities of CATs were $\geq$0.88, corresponding to more than 90% of the reliability obtained when administering all items. However, the results also indicated that, for some patients, more items might be required to obtain a precise estimate. For example, using five items to measure FA, 10% of the patients obtained a θ estimate deviating more than 0.35 (about 7% of the score range) from the full-length θ. About half of these patients had relatively low information (high measurement error) based on the five-item CAT. If a fixed information rather than a fixed-length stopping rule had been used in CAT, these patients may have been asked more items, and the deviations would likely have been reduced. The remaining half may have had unexpected responses to one or more items. Some of these unexpected responses may be errors, but likely some patients have specific problems resulting in answers that differ from what we would predict from the model. For example, a patient may generally have a good PF and have no trouble taking a walk or climbing stairs, but she may recently have broken her wrist so that she find dressing and washing herself difficult. Such specific problems may make it difficult to predict a patient's score. But this is probably a universal problem applying to most instruments and populations. Still, the overall picture derived from this set of analyses is that most patients' scores can be estimated with a high degree of precision, using only three to five items. This means that the CAT measures will also be suitable for measurement at the individual patient level. Standard instruments will

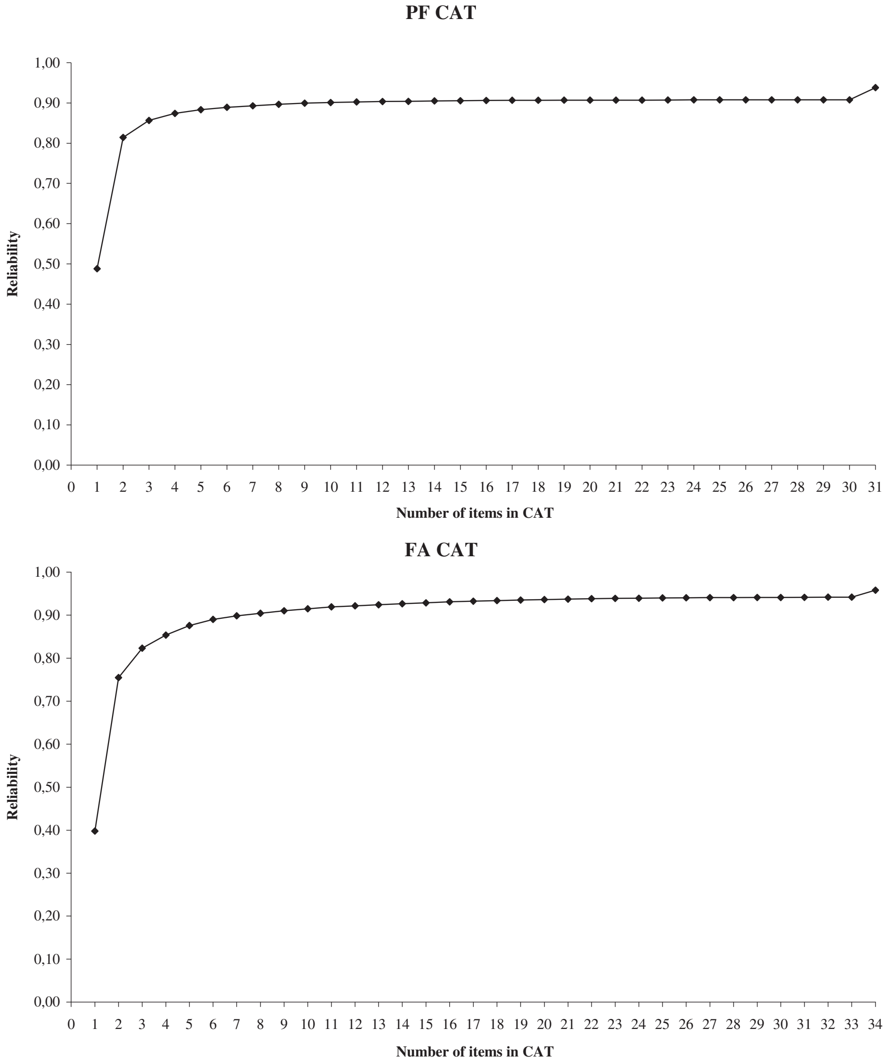

Fig. 3. The average reliability of CATs at the θ estimates in the PF and FA sample, respectively. CAT, computer-adaptive test; PF, physical functioning; FA, fatigue.

typically be either too imprecise or too lengthy/time consuming for precise and efficient measurement at the individual level, but because of their high precision with only a few items, these CAT measures may be highly useful, for example, for daily monitoring of the HRQOL of individual patients in a clinical setting.

The known group comparisons for FA of the original QLQ-C30 scale and CAT-based observed data indicated that with four or fewer items, CAT may have lower power than the original scale. However, these comparisons were based on two comparisons only and were not confirmed by the more rigorous simulations, which showed increased power also for these short FA CATs. Hence, these findings may have been caused by coincidences favoring the sum scale. In the same way, the findings for PF based on the observed data that using CAT with just two items may reduce the sample size requirements with 30%, may be overly optimistic, and may have been caused by similar coincidences here just favoring CAT.

All analyses indicated a possible increase in power/reduction in sample size requirements using the CAT measures with five or more items compared with those using the QLQ-C30 PF and FA scales. The simulations indicated potential savings in sample size requirements using CATs of five or more items of about 20%. This was generally consistent with the findings from the analyses based on observed data for FA, whereas the results based on observed data indicated a potential for even greater savings for PF.

Except for the emotional functioning scale, which consists of four items, the remaining dimensions assessed by QLQ-C30 consist of only one or two items. The current results suggest that using just one item may result in relatively low precision but that there may be considerable gain in reliability and validity when using CATs with only a few additional items. Hence, for the one- and two-item QLQ-C30 scales, there may be an even greater potential for gain in measurement precision and sample size savings from using CATs of just a few items (e.g., three to five

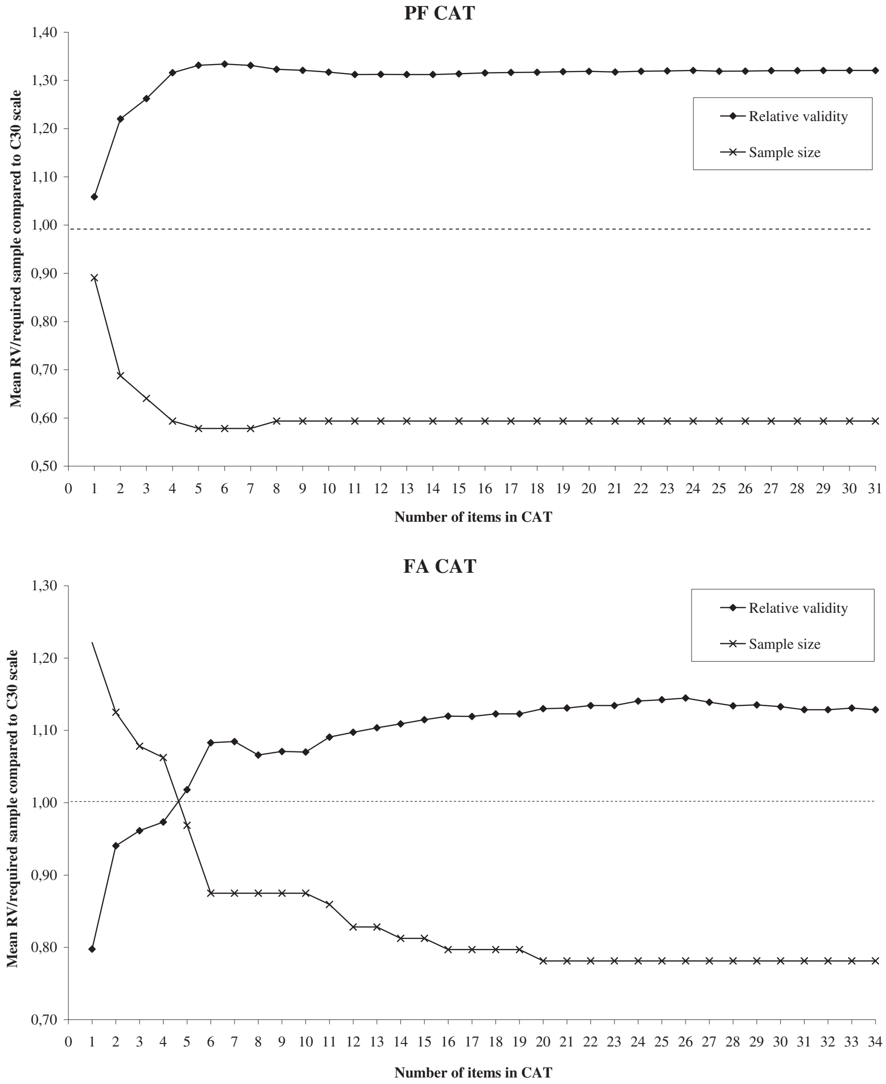

Fig. 4. The average RV and relative required sample size using CAT measurement compared with those using the QLQ-C30 sum scale based on the observed data. CAT, computer-adaptive test; QLQ-C30, Quality of Life Questionnaire; RV, relative validity.

items) than found here. We intend to investigate this in future studies.

Clearly, there will be differences in sample size savings depending on the characteristics of a study. The results presented in Table 2 for the different settings evaluated in the simulations may be a good indication of the possible savings in a particular study. All in all, our results suggest that a CAT of about five items may be a sensible choice in most situations; this will yield precise score estimates for most patients and potential sample size savings of about 20%.

It should be noted that the potential reduction in sample size requirements is not necessarily proportional with increase in measurement precision. The main reason for this is probably that the total variation measured in a sample can be subdivided into that reflecting variation between subjects (true variance) and that reflecting measurement error. Clearly, only the latter source of variance can (and should) be reduced via CAT.

Some limitations of the present study should be noted. First, the analyses were based on the same data as was used to calibrate the IRT models. This may have resulted in an overestimation of precision. In the future, we intend to conduct additional analyses using independent data sets.

Second, the variables available to conduct the observed data known groups comparisons were limited, and these were not ‘‘gold standards.’’ Hence, we do not know whether there were true differences between the groups compared, and therefore, in principle, we do not know whether the measure with the largest *t*-test size is actually the best. Nevertheless, we believe that the grouping variables used for our study were reasonable choices for investigating the known groups validity of the PF and FA measures.

Third, the simulated responses were, in part, based on the calibrated IRT models. This may have favored the CAT versions over the original QLQ-C30 scales. However, this dependence was probably only minor because this

Table 2. Summary of the simulations of the power of CAT measurement compared with those using the original QLQ-C30 PF and FA scale

		CAT			
Simulations	C30 sum scales	Three items	Five items	10 items	All items
ES = 0.2, N_1 = N_2 = 25					
Power (%)[a]	10.3	12.4	12.5	12.8	12.8
RV[b]	1.00	1.13	1.17	1.18	1.16
Sample requirement (%)[c]	100	78	73	72	74
ES = 0.2, N_1 = N_2 = 50					
Power (%)[a]	15.4	18.3	19.5	20.4	19.9
RV[b]	1.00	1.11	1.15	1.17	1.16
Sample requirement (%)[c]	100	81	76	73	74
ES = 0.2, N_1 = N_2 = 100					
Power (%)[a]	27.2	32.3	33.5	35.5	34.7
RV[b]	1.00	1.10	1.14	1.17	1.16
Sample requirement (%)[c]	100	83	77	73	74
ES = 0.5, N_1 = N_2 = 25					
Power (%)[a]	36.8	39.7	40.9	42.4	42.7
RV[b]	1.00	1.07	1.09	1.11	1.12
Sample requirement (%)[c]	100	88	84	81	81
ES = 0.5, N_1 = N_2 = 50					
Power (%)[a]	61.4	68.1	69.6	70.9	71.6
RV[b]	1.00	1.07	1.09	1.11	1.12
Sample requirement (%)[c]	100	88	84	81	81
ES = 0.5, N_1 = N_2 = 100					
Power (%)[a]	89.7	92.1	93.2	94.5	94.6
RV[b]	1.00	1.07	1.09	1.11	1.12
Sample requirement (%)[c]	100	88	84	81	81
ES = 0.8, N_1 = N_2 = 25					
Power (%)[a]	71.2	77.6	79.3	80.2	80.3
RV[b]	1.00	1.10	1.11	1.12	1.13
Sample requirement (%)[c]	100	85	81	81	77
ES = 0.8, N_1 = N_2 = 50					
Power (%)[a]	94.8	97.6	97.9	98.3	98.5
RV[b]	1.00	1.09	1.11	1.12	1.13
Sample requirement (%)[c]	100	85	81	81	77
ES = 0.8, N_1 = N_2 = 100					
Power (%)[a]	99.9	100.0	100.0	100.0	100.0
RV[b]	1.00	1.09	1.11	1.12	1.12
Sample requirement (%)[c]	100	85	81	81	81
Total average					
Power (%)[a]	56.3	59.8	60.7	61.7	61.7
RV[b]	1.00	1.09	1.12	1.13	1.14
Sample requirement (%)[c]	100	85	80	78	78

Abbreviations: CAT, computer-adaptive test; QLQ-C30, Quality of Life Questionnaire; PF, physical functioning; FA, fatigue; RV, relative validity; ES, effect size.

[a] Percent simulations across PF and FA resulting in a *t*-test with $P < 0.05$.

[b] The average RV across PF and FA compared with that using the QLQ-C30 sum scale.

[c] Sample size requirements compared with those using the original QLQ-C30 scales to obtain a power of 80% at $\alpha = 0.05$ to detect an ES of 0.2, 0.5, or 0.8, respectively.

dependence was only through the use of response patterns in groups formed from the full-length θ's. The simulated responses depended primarily on the observed item responses, which do not favor the CAT versions in particular; in fact, the opposite could also have been the case.

Finally, we have focused on fixed-length CATs using maximum information in θ for item selection and EAP for θ estimation. Although these are commonly used settings, many other settings are possible in CAT, for example, using a fixed information stopping rule or (weighted) maximum likelihood estimation of θ. Varying all these settings may affect the measurement properties of CAT, that is, other settings may result in different findings than we have observed here. Including evaluations of other settings would be too extensive here but would clearly be relevant in future studies.

Despite these limitations, we believe that the analyses reported here provide useful information about and insight into the performance of the CAT versions of the EORTC QLQ-C30 PF and FA scales. We found these two CAT measures to be precise and efficient, even with only a few items, making them suitable for measurement at the individual patient level. The potential savings in sample size requirements using these CAT measures rather than the existing QLQ-C30 scales varied as a function of the settings, but for most settings, the savings seemed to be about 20% or more. Even in light of the limitations noted previously, these findings confirm our expectation that the CAT versions of EORTC QLQ-C30 scales provide a precise and efficient means of assessing patients' HRQOL. Future studies will seek to replicate the positive results reported here for PF and FA, not only for CATs simulated from the full item pools in independent data sets but also for data generated by CAT administration in clinical populations and to determine if they also hold for other dimensions of the QLQ-C30. If so, the CAT version of the QLQ-C30 will facilitate both precise HRQOL measurement at the individual patient level and more power and thus more modest sample size requirements at the group level.

References

[1] van der Linden WJ, Glas CAW. Computerized adaptive testing: theory and practice. The Netherlands: Kluwer Academic Publishers; 2000.

[2] Wainer H. Computerized adaptive testing: a primer. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 2000.

[3] Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, 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;85:365–76.

[4] 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.

[5] Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002;324:1417–9.

[6] Petersen MAa, Groenvold M, Aaronson NK, Chie WC, Conroy T, Costantini A, et al. Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions—general approach and initial results for physical functioning. Eur J Cancer 2010; 46:1352–8.

[7] Petersen MAa, Groenvold M, Aaronson NK, Chie WC, Conroy T, Costantini A, et al. Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. Qual Life Res 2011;20:479–90.

[8] Giesinger JM, Petersen MAa, Groenvold M, Aaronson NK, Arraras JI, Conroy T, et al. Cross-cultural development of an item list for computer-adaptive testing of fatigue in oncological patients. Health Qual Life Outcomes 2011;9:19.

[9] Petersen MAa, Giesinger JM, Holzner B, Arraras JI, Conroy T, Gamper EM, et al. Psychometric evaluation of the EORTC computerized adaptive test (CAT) fatigue item pool. Submitted for publication 2011.

[10] Mead AD, Drasgow F. Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. Psychol Bull 1993;114(3):449–58.

[11] Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 1982;6(4):431–44.

[12] Fayers PM, Machin D. Quality of life. Assessment, analysis and interpretation. 2nd ed. Chichester, England: John Wiley & Sons Ltd; 2007.

[13] Choi SW. Firestar: computerized adaptive testing (CAT) simulation program for polytomous IRT models. Appl Psychol Meas 2009; 33(8):644–5.

[14] SAS Institute Inc. SAS/STAT® 9.1 User's Guide. Cary, NC: SAS Institute Inc.; 2004.

[15] Wright BD, Masters GN. Rating scale analysis—Rasch measurement. Chicago, IL: MESA PRESS; 1982.