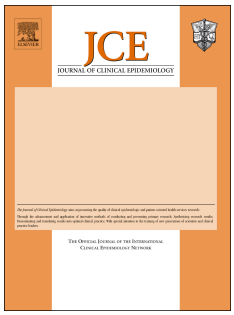
Accepted Manuscript



An emotional functioning item bank of 24 items for computer adaptive testing (CAT) was established

Morten Aa. Petersen, Eva-Maria Gamper, Anna Costantini, Johannes M. Giesinger, Bernhard Holzner, Colin Johnson, Monika Sztankay, Teresa Young, Mogens Groenvold, on behalf of the EORTC Quality of Life Group

PII: S0895-4356(15)00421-7

DOI: 10.1016/j.jclinepi.2015.09.002

Reference: JCE 8975

To appear in: *Journal of Clinical Epidemiology*

Received Date: 15 December 2014

Revised Date: 4 September 2015

Accepted Date: 7 September 2015

Please cite this article as: Petersen MA, Gamper E-M, Costantini A, Giesinger JM, Bernhard Holzner Colin Johnson Sztankay M, Young T, Groenvold M, on behalf of the EORTC Quality of Life Group, An emotional functioning item bank of 24 items for computer adaptive testing (CAT) was established, *Journal of Clinical Epidemiology* (2015), doi: 10.1016/j.jclinepi.2015.09.002.

An emotional functioning item bank of 24 items for computer adaptive testing (CAT) was established

Morten Aa. Petersen[a,*], Eva-Maria Gamper[b], Anna Costantini[c], Johannes M. Giesinger[b], Bernhard Holzner[b], Colin Johnson[d], Monika Sztankay[b], Teresa Young[e] & Mogens Groenvold[a,f] on behalf of the EORTC Quality of Life Group

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

[b] Department of Psychiatry and Psychotherapy, Innsbruck Medical University, Innsbruck, Austria

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

[d] Surgical Unit, University of Southampton, Southampton, UK

[e] Lynda Jackson Macmillan Centre, Mount Vernon Cancer Centre, Northwood, Middx, UK

[f] 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: Morten.Aagaard.Petersen@regionh.dk

Abstract

Objective: To improve measurement precision the EORTC Quality of Life Group is developing an item bank for computerized adaptive testing (CAT) of emotional functioning (EF). The item bank will be within the conceptual framework of the widely used EORTC QLQ-C30 questionnaire.

Study Design and Setting: Based on literature search and evaluations by international samples of experts and cancer patients 38 candidate items were developed. The psychometric properties of the items were evaluated in a large international sample of cancer patients. This included evaluations of dimensionality, IRT model fit, differential item functioning (DIF), and of measurement precision/statistical power.

Results: Responses were obtained from 1,023 cancer patients from four countries. The evaluations showed that 24 items could be included in a unidimensional IRT model. DIF did not seem to have any significant impact on the estimation of EF. Evaluations indicated that the CAT measure may reduce sample size requirements by up to 50% compared to the QLQ-C30 EF scale without reducing power.

Conclusion: Based on thorough psychometric evaluations we have established an EF item bank of 24 items. This will allow for more precise and flexible measurement of EF, while maintaining backward compatibility with the QLQ-C30 EF scale.

Key words: Computer adaptive test; EORTC QLQ-C30; emotional functioning; item response theory; oncology; patient-reported outcome

Running head: An item bank of 24 items for CAT measurement of emotional functioning

Word count: 4873

1. Introduction

Computerized adaptive testing (CAT) is a form of intelligent questionnaire; the basic idea is to maximize the precision by only asking questions relevant for the individual [1-3]. For example, if a patient has reported severe emotional problems to the previous items (questions), the next item will be one relevant for patients with severe problems. In this sense the questionnaire is adapted "on-the-fly" to the individual using previous responses to select the most informative next item. Clearly, such adaptation cannot be done using usual paper questionnaires, but requires the use of computer technology. All items used in a CAT are selected from a collection of items called an item bank or item pool. In a CAT item bank the items have been calibrated (fitted) to an item response theory (IRT) model [4, 5]. This means that scores based on any subset of the items are comparable. This unique property facilitates the adaptation to the individual without compromising comparability across individuals. The adaptability, i.e. the selection of the most informative item at each step, generally makes CAT instruments more precise than traditional, "static" questionnaires asking the same number of items and more efficient in the sense that fewer items are needed to obtain a specific precision. CAT instruments are also highly flexible as they can be adapted to the requirements of each study or setting. Because of these advantages of CAT several groups have developed and/or explored the use of CAT to assess patient-reported outcomes (PROs) [6-13].

The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life questionnaire (QLQ-C30) is an internationally widely used instrument for the assessment of health related quality of life (HRQOL) in cancer patients [14, 15]. 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 [16]. To improve the assessment of PROs in oncology, the EORTC Quality of Life Group is currently developing CAT versions of the EORTC QLQ-C30 scales [17-23]. The new CAT instrument operates within the same conceptual framework as the QLQ-C30. Hence, the aim is to develop a unidimensional item bank for each QLQ-C30 scale, which, in addition to the original QLQ-C30 items, consists of items covering the same aspects of the dimension as the QLQ-C30. That is, each new item bank will include all QLQ-C30 items from the relevant scale. To further enhance the compatibility with the QLQ-C30 and to ensure a homogeneous and simple format, the new items should have the same item style as the QLQ-C30 items, i.e. they should employ the same response options and recall period. In this way the CAT instrument will measure some well-validated and (to many) well-known HRQOL dimensions and it can be related to the substantial literature of studies using the QLQ-C30.

One of the key domains of the QLQ-C30 is emotional functioning (EF). The QLQ-C30 EF scale consists of four items measuring depression, anxiety (two items), and general distress that are assumed to represent a unidimensional construct. Responses to the four items are summed to form a unidimensional EF score. The new item bank should include the four QLQ-C30 EF items. Hence, the aim is a unidimensional item bank comprising the QLQ-C30 EF items and as many additional items on depression, anxiety, and general distress as possible.

As for any EORTC instrument, development of the CAT instrument takes place in an international, cross-cultural setting. The EORTC CAT development procedure consists of four phases: I) literature search, II) operationalisation, III) pre-testing, and IV) field testing. Phases I to III of the EF CAT development have been completed and described elsewhere [17]. In phase I we identified 1,729 EF items from existing questionnaires. The large majority of these items (1,480) were excluded mainly due to redundancy or lack of relevance for the EORTC measurement of EF. The remaining items formed the basis for formulating new EF items fitting the "QLQ-C30 item style". After a second round of evaluations of redundancy and relevance 63 items were retained. Evaluations by international samples of experts (phase II) and cancer patients (phase III) further reduced this to 38 candidate items. The present paper report on the phase IV field testing and psychometric evaluations of the 38 EF items.

2. Methods

The methods and analyses used in phase IV for the final development of the EF item bank are described below. They generally follow the approach previously reported for other dimensions [20-22]. Please refer to these publications for further details.

2.1. Sample

The EORTC CAT is intended for international use for cancer patients in general. Therefore, we accrued an international sample of cancer patients with different diagnoses, stages of disease, etc. Patients were recruited from oncology departments in Austria, Denmark, Italy, and the UK in the period February to December 2011. Patients were invited either by mail or when coming to the department. Eligibility requirements included a verified cancer diagnosis, age at least 18 years, and being physically and mentally competent to complete the questionnaire. Written informed consent was obtained following local requirements.

The study was approved by the local ethics committees of the participating countries.

2.2. Questionnaire

Similar to the QLQ-C30 EF items, all developed items have the recall period "during the past week" and employ a 4-point response scale: "not at all", "a little", "quite a bit" and "very much". The QLQ-C30 EF items are about distress and are formulated in the form "did you have this problem". The majority of the new items were formulated similarly covering the emotional aspects depression, anxiety and general distress/negative affect [17]. However, to try to also capture positive emotional states (i.e. emotional wellbeing/positive affect) and thereby possibly extend the measurement range, five items were formulated positively, e.g. "Have you felt cheerful?". The QLQ-C30 EF scale is scored so that higher values reflect better emotional functioning (less distress). To be in line with this scoring, we reversed the distress items for the analysis, so that higher scores reflected better functioning for all items. In addition to the 38 EF items we collected information to clarify whether the patients found any of the questions problematic and information on the patients' sociodemographic status.

2.3. Analysis plan

The psychometric evaluations and the selection of items were organized into seven steps:

1. Descriptive and basic statistical analyses: This included response rates, item means and standard deviations, and correlations with the original 4-item QLQ-C30 EF sum scale.

2. Assessing patient feedback on the items: The patients' qualitative comments and their ratings of the items were used to assess whether some of the items seemed problematic (e.g. difficult to understand or confusing).

3. Evaluation of dimensionality and local dependence: As the QLQ-C30 EF scale is unidimensional, the new CAT based EF measure is also intended to be unidimensional. We investigated the dimensionality of the items using factor analysis methods for ordinal categorical data [24]. This included evaluations of dimensionality using scree plot [25], parallel analysis [26, 27] and the Hull method [28]. These analyses were performed using the specialized software FACTOR v. 9.3.1 [29]. Further, we evaluated the fit of a unidimensional model. The following criteria were used as indication of reasonable model fit: the root mean square error of approximation (RMSEA)<0.10, the Tucker-Lewis Index (TLI)>0.90 and the Comparative Fit Index (CFI)>0.90 [30, 31]. The analyses were performed using Mplus [24].

Standard IRT models assume that the items are locally independent, i.e. item responses are

independent when controlling for the overall level of EF. This was examined using residual correlations from the factor model. Residual correlations <0.20 were regarded as indication of local independence [6].

4. Calibration of IRT model and evaluation of item fit: IRT models are a class of statistical models used to model latent variables, i.e. variables that are not directly observed but rather inferred from other (observed) variables [5, 18, 19]. We used the generalized partial credit model (GPCM) [32], a two-parameter IRT model, to form the basis for the EF CAT. In the GPCM each item has a discrimination/slope parameter describing an item's ability to discriminate between people with different scores, and a set of threshold parameters, which defines where on the EF continuum adjacent response options are equally likely to be endorsed. An item is generally most informative in the vicinity of the thresholds. The average of an item's thresholds is called the item location.

Standard IRT models assume monotonicity, i.e. the better EF the more likely it should be to give a response to the item reflecting good EF. This was examined by inspecting the average item scores across the rest scores (the sum score of all items except the item in question). If an item complies with monotonicity, the average item score should not decrease for increasing values of the rest score [33].

We used Parscale for estimating the IRT model [34]. Item fit was examined using the item-fit test S-X^2 proposed by Orlando and Thissen [35] and implemented for polytomous items in the SAS macro IRTFIT [36]. The performance of S-X^2 has been found to be superior to other fit indices [35, 37, 38]. Furthermore, we calculated the average difference between expected and observed item responses (bias) and the infit and outfit indices [39]. Infit and outfit are both based on mean square residuals, but assess slightly different aspects of fit. The infit gives more weight to responses from respondents with an EF score close to the item's location, whereas the outfit is more sensitive to unexpected responses from respondents far from the item's location. Here the infit is particularly important, since respondents in CAT measurement are primarily asked items with a location close to their EF score. For both indices values between 0.7 and 1.3 are often regarded as acceptable [40]. Large values (>1.3) indicate misfit to the model while small values (<0.7) indicate "overfit", i.e. better fit than expected statistically, e.g. because of redundancy. Misfit is our main concern, hence, large values are mainly regarded as problematic.

5. Test for differential item functioning (DIF): DIF analysis evaluates whether the items are perceived and "behave" similarly in different groups of patients [41]. If this is not the case the item

is said to exhibit DIF, which may make comparisons across groups problematic as the same item response may reflect different levels of EF in the groups. We tested for DIF using ordinal logistic regression methods [22, 42, 43] with regard to gender, age, country, cancer site, cancer stage, current treatment, education, work, and cohabitation. Each item was entered as the outcome and the group (DIF) variables were tested as independent variables controlling for the EF score estimated using the IRT model calibrated in the previous step.

Because of a large sample and multiple testing a difference was regarded as significant if $p<0.001$ and potentially relevant if also the coefficient for the group variable (numerically) exceeded 0.64 as this has been suggested as an indication of moderate to large DIF [43, 44]. For each item, each group variable was first tested individually for both non-uniform and uniform DIF. To eliminate false positive DIF findings caused by confounding of the group variables, the group variables significant in the individual tests were entered together and tested in a multiple logistic regression model. Only the results of these "multivariate" DIF analyses are reported.

We evaluated the practical impact of the significant DIF findings for estimation of EF using a method proposed by Hart et al [45] which we have further developed [10, 11]. This method compares the EF scores obtained with the model from step 4 (which does not account for DIF) with the scores obtained with an IRT model accounting for DIF. If the EF scores obtained with the two models differed substantially this was regarded as indicating practically problematic DIF, also termed "salient scale-level differential functioning" [20, 22, 45]. More precisely, if the EF scores obtained with the two models differed more than the median standard error for the EF estimates this was regarded as "salient scale-level differential functioning".

*6. Evaluation of discarded items: W*e added the discarded items one at a time to the list of items obtained after step 5 and evaluated whether the item still showed misfit. If this was not the case, i.e. if it had erroneously been discarded, it was included again.

For the final selection of items we calculated the information function. The information function is a measure of the measurement precision of an item at different levels of EF. The information of the individual items can be combined to obtain the total information of the item bank.

7. Evaluation of measurement properties: The measurement properties of the resulting EF CAT were evaluated using simulations of CAT administration based on the collected data. We simulated CATs asking 1, 2,… up to all but 1 items, respectively, estimated the EF score based on these

CATs, and compared these scores with the score based on all items. We evaluated the relative validity (RV) of these CATs as compared to the QLQ-C30 EF scale in detecting expected group differences [46]. The RV is the ratio of two test statistics for comparing two (known) groups. We used the t-test statistic for each of the CATs as the numerator and the t-test for the QLQ-C30 EF scale as the denominator. An RV>1 indicates that the CAT measure has greater discriminating power than the QLQ-C30 scale. We hypothesized that patients with stage I or II vs. stage III or IV, patients not in treatment vs. patients in treatment, patients working vs. patients not working, men and older patients would have better emotional functioning. The known groups variables being significant for at least one of the outcomes (the QLQ-C30 EF scale or one of the CAT based scores) were used for calculating RVs. We also evaluated the RV of the CATs based on simulated data. We simulated responses to the items based on EF 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 on these simulations please see Petersen et al. [21].

The descriptive analyses in step 1 were based on all available data for each item. The analyses in step 3-6 were based on complete cases, i.e. those responding to all items, while the observed data evaluations in step 7 were based on those responding to all items in the final model.

242 >> Insert Table 1. Sociodemographic and clinical characteristics of the study sample (N=1,023).<<

3. Results

We obtained responses from 1,023 cancer patients coming from Austria, Denmark, Italy, and the UK. Patient characteristics are presented in >> Insert Table 1.

1. Descriptive and basic statistical analyses. Response rates to the 38 EF items were generally high (98.2%-99.5%). The average item scores ranged 1.2-2.8 on a 0-3 scale, with 3="not at all", generally reflecting relatively good emotional functioning/little distress. Polychoric correlations between the 34 new items and the sum scale of the original four QLQ-C30 EF items ranged 0.52-0.81 for the (reversed) distress items while the five (unreversed) positive items correlated 0.32-0.55. The lowest were for "Have you felt that you have inner strengths and abilities?" (0.32) and "Have you achieved satisfaction from things that you did?" (0.33).

2. Assessing patient feedback on the items. Each item had been rated problematic (difficult to understand, annoying etc.) by no more than four patients (0.4%) and none of the items had been rated intrusive. Hence, generally the patients did not find the items problematic.

3. Evaluation of dimensionality and local dependence. Inspection of eigenvalues from an exploratory factor analysis indicated that 57% of the total variation was explained by the first factor. Three additional factors had eigenvalues above 1, and the second factor explained > 5% of the variation. The scree plot suggested one clearly important factor and three potentially important factors. The parallel analysis indicated two possibly three important factors while the Hull method indicated that one factor might be sufficient to explain the variation in the data (details omitted). In a 3-factor and a 4-factor solution the five positive items had their primary loadings on one factor while all the distress items had their primary loadings on other factors. Although the conclusion varied with the method, taken together, these analyses indicated that it might not be sensible to include all 38 items in a unidimensional model, and that the positive items may measure something distinct from the other items. This was confirmed by poor fit indices for a 1-factor model including all 38 items: RMSEA=0.144, CFI=0.690 and TLI=0.954. Excluding the five positive items improved fit, but it was still not acceptable: RMSEA=0.124, CFI=0.788 and TLI=0.971. Next, among the 33 distress items we deleted the poorest fitting items, one at a time, until a 1-factor model had acceptable fit. This resulted in a 24-item model with RMSEA=0.089, CFI=0.906 and TLI=0.987. One factor explained 65% of the variation for these 24 items. Scree plot, parallel analysis and the Hull method all indicated that one factor was sufficient to explain the variation in the 24 items.

All 276 residual correlations for the 24 items were < 0.20, and except one all were < 0.15. Hence, these correlations did not indicate any local dependence among the retained items.

281

282 >> Insert Table 2. Parameter estimates and fit statistics for the 24 items in the final IRT model. <<

283

4. Calibration of the IRT model and evaluation of item fit. There were no indications of violations of the assumption of monotonicity for the 24 items (details omitted). Therefore, we calibrated a GPCM to the 24 items and evaluated the item fit (results are summarized in >> Insert Table 2). The tests of item fit indicated good fit to the model for all items. Estimates of bias (average difference between expected and observed item responses) were all ≤0.01. The infit statistics were between 0.93 and 1.07 indicating good fit, and the outfit statistics ranged 0.59-0.97. Items 14, 20 and 26 had outfits just below 0.7 indicating that the information from these items might be slightly redundant to the information from other items. However, as all residual correlations for the three items were <0.15 the assumption of local independence did not seem significantly violated and as the items may contain unique information making them preferable in specific situations we retained them in the item bank. Hence, all in all, the 24 items seemed to have acceptable fit to the GPCM.

5. Test for DIF. >> Insert Table 3 presents the findings of significant DIF. There was no significant DIF with regard to treatment, cohabitation, or work. Of the 24 items, 12 showed potential problems with DIF. Most differences were found between countries; seven items showed country DIF while 1-2 items showed possible DIF with regard to age, gender, stage, cancer site, or education.

We evaluated the impact of the possible DIF for the estimation of EF. These evaluations generally indicated that the DIF findings had almost no effect when using all 24 items to estimate the EF score (details omitted). Even in the extreme case where only the DIF item was used to estimate EF, the DIF findings did not seem to have any significant impact, i.e. there were no indications of salient scale-level differential functioning regardless of the number of items used for the estimation. Therefore, we concluded that the DIF findings likely did not have any practically relevant impact on EF estimation, and therefore, retained all items in the model.

>> Insert Table 3. Results of the DIF analysis. Regression coefficients and p-values for the significant findings of DIF. <<

6. Evaluation of discarded items. The evaluations of the 14 items discarded in the previous steps indicated that adding any of these to the model again would result in significantly poorer model fit/lack of unidimensionality. Therefore, no items were reinstated and the 24 items with parameter estimates shown in >> Insert Table 2 constitute the final EF item bank.

The 24 items consisted of 15 items mainly covering depression related aspects, five items about anxiety and four about general distress. >>Insert Fig. 1 shows the total information for the 24 items. For comparison the figure also shows the information of the four QLQ-C30 EF items and the information of the four most informative items at each point along the EF continuum. For illustration, the EF scores obtained if reporting “very much”, “quite a bit”, “a little”, or “not at all” problem, respectively, to all items are shown. Measurement with the entire item bank provides reliability ≥ 95% (information ≥ 20) from -2.6 to 0.1 (about 3 standard deviation units) and reliability of at least 90% (information ≥ 10) from -3.0 to 0.6. Hence, the item bank provides particularly precise measurement for patients having some level of emotional problems, while for patients having at most “a little” EF problems the item bank is less precise. The four most informative items provide reliability ≥ 90% from -2.3 to -0.2 while the four QLQ-C30 items have reliability < 90% for the entire continuum.

>>Insert Fig. 1. Test information function for the 24 items in the final model, information of the four EORTC QLQ-C30 EF items, and of the four most informative items, respectively.<<

>>Insert Fig. 2. Correlations between EF scores based on CATs asking 1, 2,.., 23 items, respectively, and EF scores based on all 24 items.<<

7. Evaluation of measurement properties. For CATs of all lengths the mean and median EF score was very close to the score based on all 24 items (details omitted). Scores based on three or more items correlated >0.90 with the scores based on all items (see >>Insert Fig. 2). >>Insert Fig. 3 summarizes the results of the known groups comparisons. Contrary to expectation there was no significant difference between work statuses. Stage, treatment, gender and age showed significant differences as hypothesised and were therefore used to calculate mean RVs for the observed data. Both observed and simulated data indicated low power if using only one item. The observed data indicated that when 2 or more items are asked in the CAT, sample sizes may be reduced by about 20-50% without loss of power as compared to using the QLQ-C30 EF scale. The reduction generally increased with the length of the CAT with approximately 20% savings with 2 items, 40% with 6 items and 50% with 18 or more items. The simulated data on the other hand indicated that using the CAT may only reduce sample size requirements by up to 15%. Hence, there were significant differences in the expected reductions in sample size requirements based on the observed and simulated data.

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

4. Discussion

The overall aim of the EORTC CAT-project is to develop a new version of the EORTC QLQ-C30 with better and more precise measurement of HRQOL. This will hopefully improve the identification and treatment of patients' symptoms and problems. In this study we have expanded the existing QLQ-C30 EF measure to an item bank covering a broader range of the EF continuum. We obtained an EF item bank of 24 items showing good psychometric properties: factor analysis indicated acceptable unidimensionality and IRT calibration and evaluations showed good fit to a GPCM. We found some indications of DIF. However, evaluations indicated that even though there were statistically significant differences these did not have any practically relevant impact on the estimation of EF, i.e. EF scores based on the item bank can be compared across patients and studies regardless of patient characteristics.

The item bank provides high measurement precision for a wide range of EF. However, for patients having at most "a little" EF problems the item bank may lack precision. The measurement may be improved at a later stage by adding new items particularly relevant for these patients, although from a clinical point of view it may rarely be relevant to determine with high precision whether a patient has "little" or "very little" emotional problems, since neither is likely to need treatment. Compared to the original QLQ-C30 EF scale the new EF item bank has markedly higher precision across the entire continuum, indicating a significant improvement in measurement precision. Also if selecting the four most informative items at each point across the continuum, there is a significant gain in precision compared to the four QLQ-C30 items; the four "maximum information" items provide at least 50% more information than the QLQ-C30 items for about 3 standard deviation units.

The list of candidate items included five items on emotional wellbeing/positive affect. The evaluations indicated that they did not fit well into a unidimensional model with the distress items and were therefore excluded. These positive items were "experimental" as the QLQ-C30 emotional scale covers distress only. Further, the assessment of distress is the main interest in the context of identifying impairments and intervention needs. This split into an emotional distress and an emotional wellbeing component seems in line with findings, e.g., for the General Health Questionnaire [47]. Hence, although we refer to the item bank as measuring "emotional functioning" to be in line with the terminology of the QLQ-C30, it may be more appropriate to refer

to it as a measure of emotional distress.

In addition to the positive items, nine items were deleted because of poor fit to a unidimensional model. Of these, two items on feeling “anxious” and “restless” may have been somewhat ambiguous as they, besides anxiety, may reflect feeling eager/excited. An item on “felt life was meaningless” and one on “felt life isn't worthwhile” were also deleted. Of all items tested, they had the fewest reported problems. The highly skewed response distributions may have affected their fit. When new data with more patients having severe emotional problems become available it might be that these items will show more favorable fit. Hence, they may be candidates for inclusion in a “version 2” of the item bank. However, based on the current sample it does not seem appropriate to include the two items. Finally, four items on being “furious”, “angry”, “bad-tempered”, and “impatient” were deleted. Likely all four, and particularly the first three, relate to anger/aggressive feelings. This aspect of emotional functioning did not seem to fit well with the other aspects and is therefore, not covered by our item bank (anger is not part of the original QLQ-C30 EF scale either). This is an example of the limitation of our unidimensionality requirement: anger may be a relevant aspect to measure, but seemingly cannot be included in a unidimensional measure with the other aspects, and is therefore not included here. If we want to measure anger, a separate item bank may be constructed. Anger has also in other studies been modelled as a separate domain when measuring emotional distress [48].

The item bank is dominated by items on depression (constituting 15 of the 24 items). To get a more even distribution of the content areas more items on anxiety and general distress could be added. Note, however, that the main reason for the smaller numbers of anxiety and general distress items is that it was difficult to formulate many distinct, relevant, and fitting items for these areas. But even though the item bank is dominated by items on depression, the CAT can be programmed to always include items from all content areas, thereby ensuring coverage of all areas constituting the EORTC EF construct. As this will keep the content of the new measure as close as possible to the QLQ-C30 EF scale we propose to use such content balance.

The CAT generally provided precise measurement when comparing scores based on CATs of different lengths to scores based on the entire item bank. The known groups comparisons generally indicated that by asking two or more items, sample sizes may be reduced without loss of power as compared to the QLQ-C30 scale. However, the predicted reduction in sample size requirements using the new measure differed significantly between the observed data and simulated data analyses; the simulated data indicated a maximum of 15% reduction while the observed data indicated up to 50% reduction. This suggests that the actual gain may vary across studies depending

on patient samples etc. and underline that detailed evaluations of the power of the CAT measure in independent data are needed. A validation project has been initiated to evaluate the measurement properties of the EORTC CAT instrument. This validation study will include new countries, including countries from Eastern Europe and from outside Europe, allowing for evaluation of the generalisability of the current findings. Although this validation is not completed, the current, preliminary version the EORTC CAT may be used by other researchers. For more information on this preliminary use of the EORTC CAT please visit http://groups.eortc.be/qol/eortc-cat.

The uncertainty of the increased power using the CAT illustrates that although there are clear theoretical advantages of CAT it is not obvious how these translate into practical gain like increased power. CAT is more complex to develop, use and understand than traditional questionnaires. When is this additional complexity worthwhile, and when may a simpler tool suffice? More studies on the practical advantages of CAT for PRO measurement are relevant.

In conclusion, we have developed an item bank of 24 items for CAT measurement of emotional functioning. The CAT measure was developed in an international setting targeted at cancer patients in general in a multitude of languages, but it may also be applied to other patients (and the general population) as well. It 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, particularly for patients with some degree of emotional problems. Evaluations indicated that sample sizes may be reduced up to 50% without loss of power, compared to the QLQ-C30. However, these evaluations were subject to some uncertainty and the measurement properties should be validated with new data before drawing any final conclusions.

Acknowledgments

The study was funded by grants from the EORTC Quality of Life Group. The work of Eva-Maria Gamper and Johannes M. Giesinger was funded by a grant from the Austrian Science Fund (FWF L502 and FWF J3353).

The authors would like to thank the patients responding to our items and our collaborators for collecting these essential patient responses.

None of the authors have any conflicts of interest that might have biased the work.

References

1. Wainer H. Computerized Adaptive testing: A Primer (2nd). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.; 2000.

2. van der Linden WJ, Glas CAW. Computerized Adaptive testing: Theory and Practice. The Netherlands: Kluwer Academic Publishers; 2000.

3. van der Linden WJ, Glas CAW. Elements of Adaptive Testing. New York: Springer; 2010.

4. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. Newbury Park: Sage Publications, Inc; 1991.

5. van der Linden WJ, Hambleton RK. Handbook of Modern Item Response Theory. Berlin: Springer-Verlag; 1997.

6. Bjorner JB, Kosinski M, Ware JE, Jr. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). Quality of Life Research 2003; 12(8):913-33.

7. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res. 2007; 16 Suppl 1:133-41.

8. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Quality of Life Research 2005; 14(10):2277-91.

9. Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB et al. Development of the CAT-ANX: A Computerized Adaptive Test for Anxiety. Am.J.Psychiatry 2013;

10. Haley SM, Ni P, Fragala-Pinkham MA, Skrinar AM, Corzo D. A computer adaptive testing approach for assessing physical functioning in children and adolescents. Dev.Med Child Neurol. 2005; 47(2):113-20.

11. Hart DL, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function. Quality of Life Research 2008; 17(8):1081-91.

12. Jette AM, Haley SM, Ni P, Olarsch S, Moed R. Creating a computer adaptive test version of the late-life function and disability instrument. J Gerontol.A Biol.Sci.Med.Sci. 2008; 63(11):1246-56.

13. Riley WT, Pilkonis P, Cella D. Application of the National Institutes of Health Patient-reported Outcome Measurement Information System (PROMIS) to mental health research. J.Ment.Health Policy Econ. 2011; 14(4):201-8.

14. 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(5):365-76.

15. Fayers P,.Bottomley A. Quality of life research within the EORTC-the EORTC QLQ - C30. European Organisation for Research and Treatment of Cancer. European Journal of Cancer 2002; 38 Suppl 4:S125-S133.

16. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A. The EORTC QLQ-C30 Scoring Manual (third). Brussels: European Organisation for Research and Treatment of Cancer; 2001.

17. Gamper EM, Groenvold M, Petersen MAa, Young T, Costantini A, Aaronson N et al. The EORTC Emotional Functioning computer adaptive test (CAT): phase I-III of a cross-cultural item bank development. Psycho-Oncology 2014; 23:397-403.

18. 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)

19. Petersen MAa, Groenvold M, Aaronson NK, Chie W-C, 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. European Journal of Cancer 2010; 46:1352-8.

20. Petersen MAa, Groenvold M, Aaronson NK, Chie W-C, Conroy T, Costantini A et al. Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. Quality of Life Research 2011; 20(4):479-90.

21. Petersen MAa, Aaronson NK, Arraras JI, Chie W-C, Conroy T, Costantini A et al. The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency. Journal of Clinical Epidemiology 2012; 66(3):330-9.

22. 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. Quality of Life Research 2013; 22(9):2443-54.

23. Thamsborg LH, Petersen MAa, Aaronson NK, Chie W-C, Costantini A, Holzner B et al. Development of a lack of appetite item bank for computer-adaptive testing (CAT). Support Care Cancer 2014; 23(6):1541-8.

24. Muthen LK, Muthen BO. Mplus User's Guide (2nd edition). Los Angeles, CA: Muthen & Muthen; 2002.

25. Cattell RB. Scree Test for Number of Factors. Multivariate Behavioral Research 1966; 1(2):245-76.

26. Horn JL. A Rationale and Test for the Number of Factors in Factor-Analysis. Psychometrika 1965; 30(2):179-85.

27. Timmerman ME,.Lorenzo-Seva U. Dimensionality Assessment of Ordered Polytomous Items With Parallel Analysis. Psychol Methods 2011; 16(2):209-20.

28. Lorenzo-Seva U, Timmerman ME, Kiers HAL. The Hull Method for Selecting the Number of Common Factors. Multivariate Behavioral Research 2011; 46(2):340-64.

29. Lorenzo-Seva U,.Ferrando PJ. FACTOR v. 9.3.1 [computer program]. Available at http://psico.fcep.urv.es/utilitats/factor (Accessed May 2015). 2015;

30. Browne MW,.Cudeck R. Alternative Ways of Assessing Model Fit. Sociological Methods & Research 1992; 21(2):230-58.

31. Kline RB. Principles and practice of structural equation modeling (2nd). New York: The Guilford Press; 2005.

32. Muraki E. A Generalized Partial Credit Model. In: van der Linden WJ, Hambleton RK, Eds. Handbook of Modern Item Response Theory Berlin: Springer; 1997:153-68.

33. Junker BW,.Sijtsma K. Latent and manifest monotonicity in item response models. Applied Psychological Measurement 2000; 24(1):65-81.

34. 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.

35. Orlando M,.Thissen D. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. Applied Psychological Measurement 2000; 24(1):50-64.

36. Bjorner JB, Smith KJ, Stone C, Sun X. Software: IRTFIT: A macro for item fit and local dependence tests under IRT models (obtained at: http://outcomes.cancer.gov/areas/measurement/irt_model_fit.html). 2007;

37. Orlando M,.Thissen D. Further investigation of the performance of S-X-2: An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement 2003; 27(4):289-98.

38. Kang T,.Chen TT. Performance of the Generalized S-X(2) Item Fit Index for Polytomous IRT Models. Journal of Educational Measurement 2008; 45(4):391-406.

39. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences (2nd edition). New Jersey: Lawrence Erlbaum Associates, Inc.; 2007.

40. Wright BD,.Linacre JM. Reasonable mean-square fit values. Rasch Measurement Transactions 1994; 8(3):370.

41. Holland PW, Wainer H. Differential Item Functioning. Hilsdale, NJ: Lawrence Erlbaum Associates; 1993.

42. 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.

43. Petersen MAa, Groenvold M, Bjorner JB, Aaronson NK, Conroy T, Cull A et al. Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. Quality of Life Research 2003; 12(4):373-85.

44. Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. Journal of Clinical Epidemiology 1998; 51(11):1189-202.

45. Hart DL, Deutscher D, Crane PK, Wang YC. Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. Quality of Life Research 2009; 18(8):1067-83.

46. Fayers PM, Machin D. Quality of Life. The assessment, analysis and Interpretation of patient-reported outcomes (2nd ed.). Chichester: John Wiley & Sons Ltd; 2007.

47. Stochl J, Jones PB, Croudace TJ. Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. BMC Med Res Methodol. 2012; 12:74.

48. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)): depression, anxiety, and anger. Assessment. 2011; 18(3):263-83.

Table 1. Sociodemographic and clinical characteristics of the study sample (N=1,023).

		N/mean
Age (mean years)		62 (range 22-88)
Gender	Male	483 (47%)
	Female	540 (53%)
Country	Austria	204 (20%)
	Denmark	205 (20%)
	Italy	94 (9%)
	UK	520 (51%)
Education	0-10 years	376 (37%)
	11-13 years	258 (25%)
	14-16 years	218 (21%)
	>16 years	158 (15%)
Work	Working	322 (31%)
	Retired	564 (55%)
	Other	125 (12%)
Cohabitation	Living with a partner	759 (74%)
	Living alone	244 (24%)
Cancer stage	I-II	456 (45%)
	III-IV	420 (41%)
Cancer site	Breast	130 (13%)
	Gastrointestinal	199 (20%)
	Gynaecological	97 (10%)
	Head and neck	74 (7%)
	Lung	90 (9%)
	Urogenital	104 (10%)
	Other	235 (23%)
Current treatment	Chemotherapy	316 (31%)
	Other treatment	117 (11%)
	No current treatment	486 (48%)

Table 2. Parameter estimates and fit statistics for the 24 items in the final IRT model.

		Thresholds							
Item	**Slope**	**T1**	**T2**	**T3**	**Location**	**Item fit p-value**	**Bias**	**Infit**	**Outfit**
Item 3: Did you feel tense? (from QLQ-C30)	1.62	-1.79	-1.04	0.10	-0.91	0.837	0.01	0.98	0.89
Item 4: Have you felt helpless?	2.41	-1.67	-1.23	-0.67	-1.19	0.985	0.01	1.00	0.73
Item 5: Have you felt panic?	1.78	-1.80	-1.53	-1.19	-1.51	0.996	0.01	0.99	0.97
Item 6: Have you lost interest in things, such as recreational or social activities (independently of your actual ability to do them)?	1.14	-1.91	-1.21	-0.81	-1.31	0.640	0.01	0.99	0.97
Item 7: Have you felt vulnerable?	2.11	-1.79	-1.24	-0.26	-1.10	0.953	0.01	0.99	0.82
Item 8: Have you felt frustrated?	1.72	-1.65	-1.04	-0.14	-0.94	0.359	0.01	0.99	0.85
Item 9: Have you felt worthless?	2.29	-1.86	-1.48	-1.01	-1.45	0.969	0.01	1.02	0.82
Item 12: Have you felt discouraged?	2.58	-1.93	-1.42	-0.44	-1.27	0.820	0.01	0.97	0.75
Item 13: Have you had emotional outbursts?	1.24	-1.65	-1.60	-0.75	-1.33	0.225	0.01	1.01	0.89
Item 14: Have you felt that nothing could cheer you up?	3.32	-1.95	-1.53	-0.80	-1.43	0.995	0.01	0.98	0.63
Item 15: Have you felt afraid?	1.69	-1.63	-1.53	-0.53	-1.23	0.616	0.01	1.00	0.83
Item 16: Have you felt that pleasure has gone from your life?	1.97	-1.71	-1.47	-0.66	-1.28	0.985	0.01	0.97	0.84
Item 17: Have you had difficulty relaxing?	1.41	-1.78	-1.25	-0.25	-1.10	0.560	0.01	0.99	0.93

Item 18: Have you lost interest in your appearance?	1.33	-2.31	-1.92	-1.33	-1.85	0.543	0.01	1.02	0.93
Item 20: Have you felt miserable?	3.19	-1.75	-1.16	-0.29	-1.06	0.878	0.01	0.93	0.68
Item 22: Did you feel depressed? (from QLQ-C30)	2.53	-1.66	-1.25	-0.29	-1.07	0.946	0.01	0.95	0.82
Item 23: Did you feel irritable? (from QLQ-C30)	1.64	-2.28	-1.50	-0.02	-1.27	0.420	0.01	0.97	0.88
Item 24: Have you felt useless?	2.25	-1.85	-1.45	-0.89	-1.40	0.651	0.01	0.98	0.86
Item 25: Did you worry? (from QLQ-C30)	1.64	-1.66	-0.94	0.44	-0.72	0.718	0.01	0.95	0.92
Item 26: Have you felt desperate?	3.03	-2.10	-1.43	-1.03	-1.52	0.952	0.01	1.00	0.59
Item 29: Have you been afraid of losing control?	1.81	-2.06	-1.71	-1.06	-1.61	0.871	0.01	1.03	0.87
Item 30: Have you felt sad?	2.61	-1.59	-1.14	0.09	-0.88	0.927	0.01	0.93	0.83
Item 31: Have you felt like giving up?	2.17	-1.80	-1.58	-1.37	-1.58	0.992	0.01	1.07	0.71
Item 32: Have you felt that you have nothing to look forward to?*	2.40	-1.52	-0.97		-1.24	0.982	0.01	1.00	0.72

*: Response options “Quite a bit” and “Very much” were collapsed for item 32 because of reversed thresholds.

Table 3. Results of the DIF analysis. Regression coefficients and p-values for the significant findings of DIF.

Item	**DIF**	**β**	**p-value**	**DIF**	**β**	**p-value**
Item 3	Country:	-0.88 (Austria)	<0.0001			
Item 4	Stage:	0.76	<0.0001			
Item 5	No DIF					
Item 6	No DIF					
Item 7	Education:	0.96 (0-10 years)	0.0005			
Item 8	Country:	1.61 (Italy)	<0.0001			
Item 9	No DIF					
Item 12	No DIF					
Item 13	Age:	-1.49 (<40 years)	<0.0001			
Item 14	No DIF					
Item 15	Site:	0.80	0.0001			
Item 16	Country:	-1.61 (Austria)	<0.0001	Gender:	-0.69	<0.0001
Item 17	No DIF					
Item 18	No DIF					
Item 20	Country:	1.16 (Denmark)	<0.0001			
Item 22	Country:	-1.01 (Austria)	<0.0001			
Item 23	No DIF					
Item 24	No DIF					
Item 25	No DIF					

Item 26	Country:	-1.76 (Austria)	<0.0001
Item 29	No DIF		
Item 30	No DIF		
Item 31	Country:	-1.42 (Italy)	<0.0001
Item 32	Gender:	-0.73	0.0001

For country the largest regression coefficient (and the country) when comparing with UK is shown. A coefficient>0 indicates UK patients are more likely to report problems on the item (when controlling for the EF score).

For age the largest regression coefficient (and age group) when comparing with patients ≥70 years is shown. A coefficient>0 indicates that patients ≥70 years are more likely to report problems on the item.

For gender a coefficient>0 indicates that women are more likely to report problems on the item.

For cancer site a regression coefficient>0 indicates that breast cancer patients are more likely to report problems.

For cancer stage a regression coefficient>0 indicates that patients with advanced cancer (III-IV) are more likely to report problems.

For education the largest regression coefficient (and group, years of education) when comparing with patients having more than 16 years of education is shown. A coefficient>0 indicates patients having more than 16 years of education are more likely to report problems on the item.

Fig. 1. Test information function for the 24 items in the final model, information of the four EORTC QLQ-C30 EF items, and of the four most informative items, respectively.

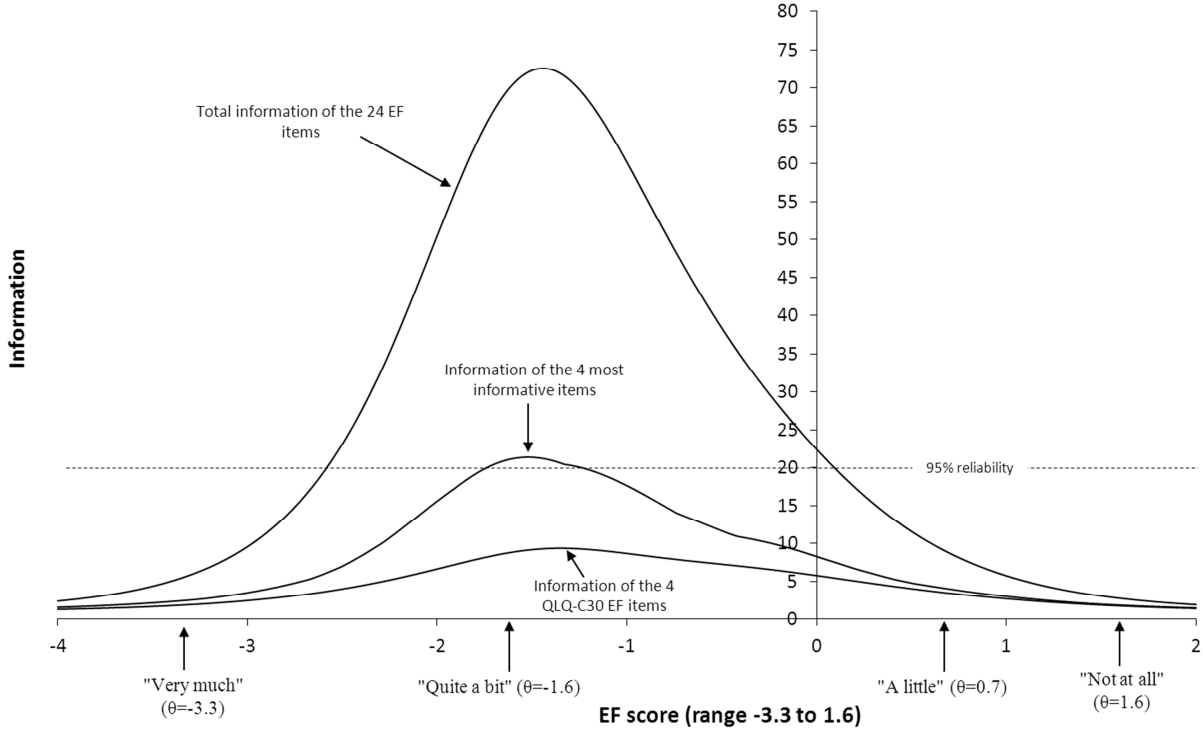


Fig. 2. Correlations between EF scores based on CATs asking 1, 2,.., 23 items, respectively, and EF scores based on all 24 items.

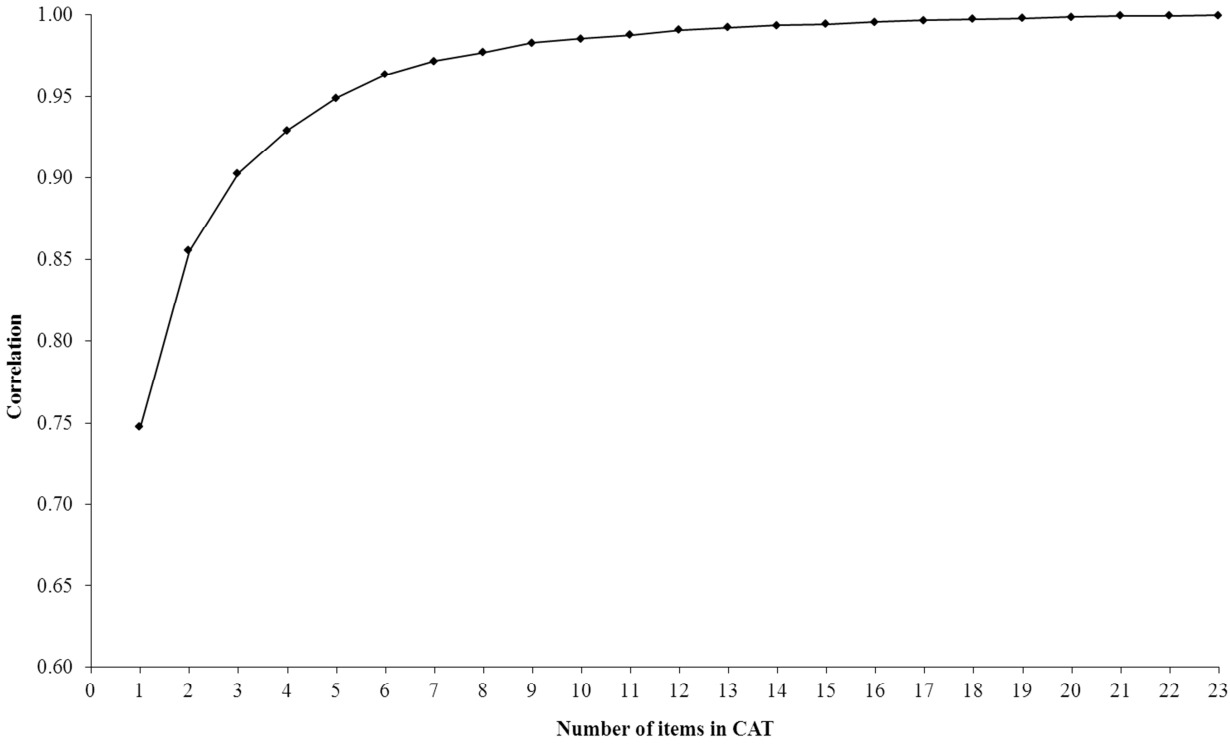
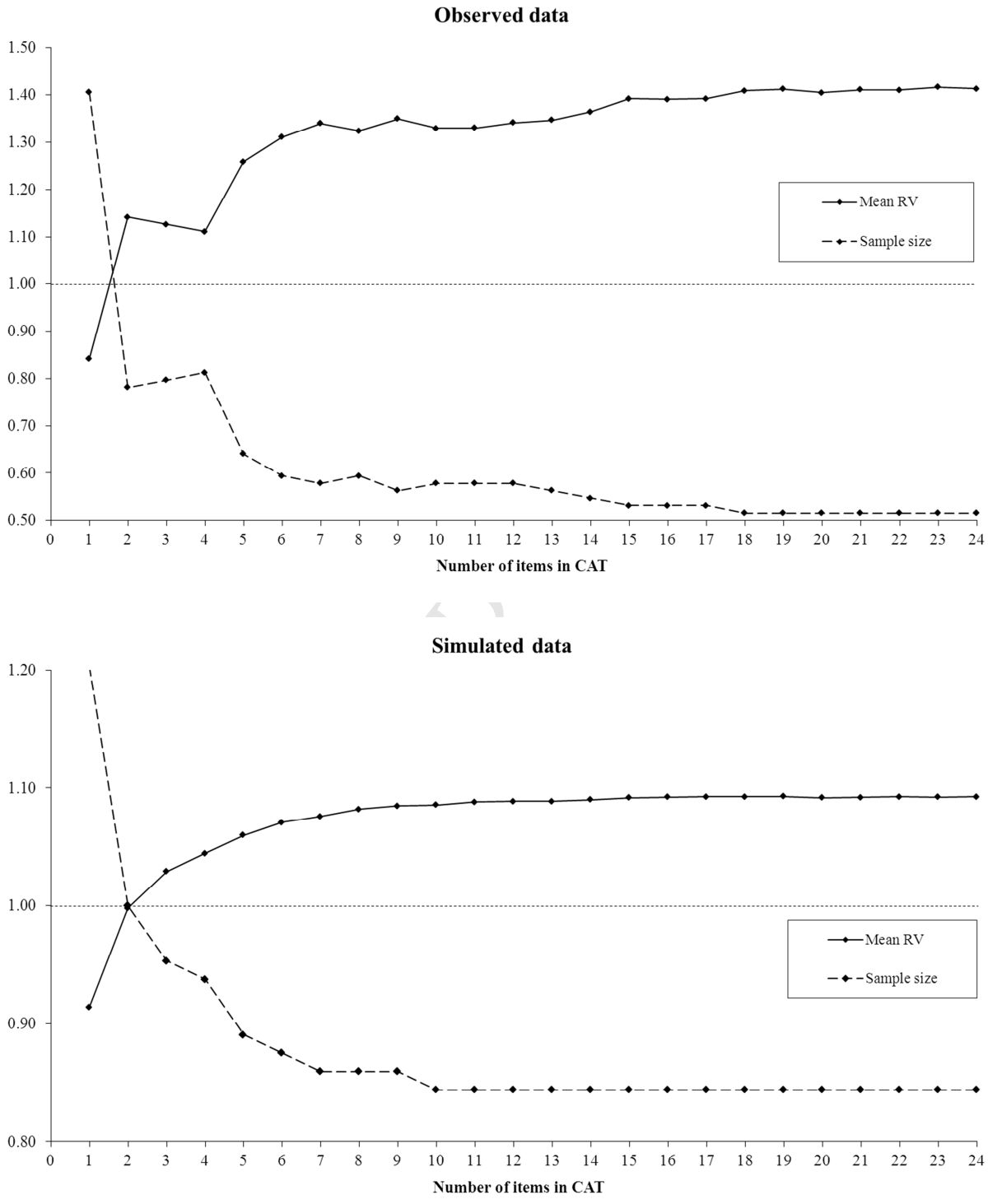


Fig. 3. The average relative validity (RV) and relative required sample size using CAT measurement compared to using the QLQ-C30 EF sum scale based on observed and simulated data, respectively.



Footnote to Fig. 3: For both RV and sample size the plots show the ratio of using the CAT compared to using the QLQ-C30 sum scale. For example, using a CAT with six items the observed data indicates that the validity of the CAT is 1.31 times that of the QLQ-C30 scale (RV=1.31) resulting in that only 0.59 (59%) of a sample size used with the QLQ-C30 scale is required when using a 6-item CAT to obtain the same power.