



Development and Psychometric Evaluation of an Item Bank for Computerized Adaptive Testing of the EORTC Insomnia Dimension in Cancer Patients (EORTC CAT-SL)

Linda Dirven, et al. *[full author details at the end of the article]*
on behalf of the EORTC Quality of Life Group

Received: 8 May 2019 / Accepted: 19 November 2019 / Published online: 6 December 2019
© The Author(s) 2019

Abstract

To further advance assessment of patient-reported outcomes, the European Organisation of Research and Treatment of Cancer (EORTC) Quality of Life Group has developed computerized adaptive test (CAT) versions of all EORTC Quality of Life Core Questionnaire (QLQ-C30) scales/items. The aim of this study was to develop and evaluate an item bank for CAT measurement of insomnia (CAT-SL). In line with the EORTC guidelines, the developmental process comprised four phases: (I) defining the concept insomnia and literature search, (II) selection and formulation of new items, (III) pre-testing and (IV) field-testing, including psychometric analyses of the final item bank. In phase I, the literature search identified 155 items that were compatible with our conceptualisation of insomnia, including both quantity and quality of sleep. In phase II, following a multistep-approach, this number was reduced to 15 candidate items. Pre-testing of these items in cancer patients (phase III) resulted in an item list of 14 items, which were field-tested among 1094 patients in phase IV. Psychometric evaluations showed that eight items could be retained in a unidimensional model. The final item bank yielded greater measurement precision than the original QLQ-C30 insomnia item. It was estimated that administering two or more items from the insomnia item bank with CAT results in a saving in sample size between approximately 15–25%. The 8-item EORTC CAT-SL item bank facilitates precise and efficient measurement of insomnia as part of the EORTC CAT system of health-related quality life assessment in both clinical research and practice.

Keywords Computerized adaptive testing · Cancer · Insomnia · Sleeplessness · Item bank · Health-related quality of life · EORTC QLQ-C30

Introduction

Prevalence rates of insomnia as high as 60% have been reported in patients with cancer (Savard et al. 2009), which is markedly higher than the <15% reported in the general

population (Itani et al. 2016; Pallesen et al. 2014). Insomnia, or sleeplessness, has been defined as a subjective feeling of poor or unsatisfactory sleep that may involve insufficient sleep, difficulty falling asleep, difficulty maintaining sleep, early final waking, and/or poor quality or non-restorative sleep (O'Donnell 2004; Sateia and Lang 2008; Savard and Morin 2001). Factors contributing to insomnia include predisposing factors such as age and gender, precipitating factors like medical treatment, and perpetuating factors such as poor sleeping behaviour (Harris et al. 2014; Induru and Walsh 2014). Not only patients undergoing anti-tumour treatment report high levels of insomnia (Savard et al. 2009; Costa et al. 2014; Robertson et al. 2016); cancer survivors and advanced cancer patients report higher levels of insomnia when compared to the general population (Arndt et al. 2017; Lowery-Allison et al. 2017; Mercadante et al. 2017).

Insomnia is typically measured with patient-reported outcome measures, for which several instruments are available (Chen et al. 2018). Moreover, insomnia is often included in measures of health-related quality life (HRQoL). One frequently used tool to assess HRQoL in cancer patients is the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) (Aaronson et al. 1993). This multidimensional instrument consists of 30 items comprising five functional scales, three symptom scales, and six single item symptoms, including an item on insomnia.

Static measures such as the QLQ-C30 require that all patients respond to all questions, regardless of their relevance for any individual patient. Static instruments, with a fixed number of items per domain of interest, may also yield suboptimal measurement precision. The use of computerized adaptive testing (CAT) can overcome these issues (Bjorner et al. 2007; Cella et al. 2007; van der Linden and Glas 2010). CAT is a method of selecting the most relevant item set for an individual patient. A computer algorithm selects a new item based on the patient's response to the previous items. As a result, fewer items are typically needed to obtain the desired level of measurement precision. Importantly, results across patients or studies are still comparable, even if patients do not answer the same subset of items. This is possible as CAT is based on item response theory (IRT) methods (Hambleton et al. 1991).

The EORTC Quality of Life Group (QLG) has an on-going, large scale international project to develop CAT versions of all EORTC QLQ-C30 scales/items (Dirven et al. 2017; Gamper et al. 2014, 2016; Giesinger et al. 2011; Petersen et al. 2011, 2013a, b, 2016a, 2018a; Thamsborg et al. 2015), and for several scales higher measurement precision and reduced response burden has been achieved (Dirven et al. 2017; Gamper et al. 2016; Petersen et al. 2011, 2013b, 2016a, b). Whether this holds true for other scales remains to be determined. The EORTC CAT development process consists of four phases: (I) conceptualisation and literature search, (II) selection and formulation of new items, (III) pre-testing and (IV) larger scale field-testing, including psychometric analyses of the final item bank. Here we report the results of phases I-IV of the development and evaluation of an item bank for insomnia (EORTC CAT-SL).

Methods

The approach for the development of items banks for the EORTC CAT follows the EORTC guidelines (Petersen et al. 2010, 2018; Johnson et al. 2011), in which phase I

and II are conducted collaboratively. The methodology for the four phases is briefly summarized below.

Phase I: Conceptualization and Literature Search

To ensure comparability between the CAT instrument and the original EORTC QLQ-C30 questionnaire, the additional items for the insomnia item bank should reflect the same underlying concept as the original item (i.e., “Have you had trouble sleeping?”). Also, items should have the same response format and timeframe (i.e., a 4-point polytomous response scale – “not at all,” “a little,” “quite a bit,” “very much” during the past week’), which allows assessment of the severity of insomnia.

WHO ICD-11 defines insomnia as a condition of impaired quantity and/or quality of sleep resulting in daytime impairment (World Health Organization 2018). According to Morin, the clinical significance of insomnia is determined by its intensity/severity, frequency, duration, and daytime sequela (Morin 1993). Here we regard intensity/severity, frequency and duration as reflecting the level of insomnia, while daytime sequela is regarded as a potential consequence of insomnia. The original item asks about insomnia in general and responses to the item may reflect any aspect of the level of insomnia. To reflect the definition of insomnia and the general nature of the original item, item selection and development focused on items indicating (1) any trouble with sleeping (i.e. sleep quantity, such as trouble falling asleep or maintaining sleep) or (2) the quality of sleep (e.g. restless sleep).

To identify self-report instruments and items assessing insomnia, we performed a literature search in the e-resources PROQOLID, PubMed and the EORTC QLG Item Library up to November 2007. Keywords used were ‘insomnia’, ‘sleepless’, ‘sleeplessness’, ‘wakefulness’ and ‘sleep’. The identified items served as a basis for the generation of new items.

Phase II: Selection and Formulation of New Items

The item list resulting from phase I was refined following a multistep item selection procedure. The first five steps were conducted by two members of the CAT project. The first step comprised the classification of items into aspects of (1) ‘sleep quantity’, (2) ‘quality of sleep’ or (3) ‘other’. The next step focused on the deletion of duplicate items, while in step 3 remaining items were used as inspiration to formulate new items fitting the QLQ-C30 item style. In step 4 it was determined if the items covered all aspects of insomnia. In case of insufficient coverage of the measurement continuum (i.e. different levels of severity), new items could be generated in step 5. Lastly, items were assessed by two groups of experts, QLG members and international experts on the development of patient-reported outcome measures for cancer patients or on sleep measurement, for their relevance to the construct insomnia, their appropriateness, and for whether they were clear and well-defined.

Phase III: Pre-Testing

The preliminary item list resulting from phase II was pre-tested in an international convenience sample of adult cancer patients (the specific countries were based on a

convenience sample) to ensure content validity and appropriateness of the items. The English item list was first translated into relevant languages by the Translation Unit of the EORTC according to well-established guidelines, allowing that patients could complete the questionnaire in their own language. (Kulis et al. 2016) Next, the item list was administered to the patients. In addition, semi-structured interviews were conducted with each patient to evaluate each item for wording (i.e., whether the items were difficult, confusing, annoying, upsetting or intrusive) and whether relevant issues were missing. If there were no comments or just a few comments on an item, these were deemed relevant for phase IV where the psychometric analyses would provide additional information on item relevance.

Phase IV: Field-Testing and Psychometric Analyses

Sample

We recruited a heterogeneous convenience sample of cancer patients across oncology departments in Europe (Denmark, France, The Netherlands, Sweden and the United Kingdom) and Taiwan. Eligible patients had to be over 18 years of age and have a histologically verified cancer. Moreover, patients had to be able to understand and read the questionnaire and informed consent form. To facilitate item calibration, we aimed to include a minimum of 1000 patients with different cancer diagnoses, disease stages, treatment modalities, and sociodemographic characteristics (Fayers 2007; Jiang et al. 2016; Muraki and Bock 1996).

Patients were requested to complete the questionnaire resulting from phase III, as well as the EORTC QLQ-C30 and a short debriefing questionnaire to assess whether patients found any of the items problematic. In addition, information on patient- and disease-related characteristics was collected. The procedures performed in this study are in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments, and the ethical committees of all participating centres approved the study. Patients gave their written informed consent before participation.

Descriptive and Basic Statistical Analyses

Descriptive statistics were used to characterize the patient sample, to calculate response rates and item means with their standard deviations (SD), and investigate responses to the debriefing questionnaires.

Evaluation of Dimensionality and Local Dependence

To evaluate if the requirements for unidimensionality and local dependence were met, both exploratory and confirmatory factor analyses were performed (i.e. for ordinal categorical data, with polychoric correlations, conducted in Mplus (Muthen and Muthen 2002)). Eigenvalues and scree plot (Cattell 1966) were used to evaluate the dimensionality of the items. The root mean square error of approximation (RMSEA) <0.10 , the Comparative Fit Index (CFI) >0.90 , and the Tucker-Lewis Index (TLI) >0.90 (Browne and Cudek 1992; Kline 2005) were used to define an acceptable model fit. Local independence (i.e., item

responses are independent when controlling for the overall level of insomnia) was evaluated by means of residual correlations, with correlations <0.20 serving as indicators for local independence (Bjorner et al. 2003).

Item Bank Calibration and Evaluation of Item Fit

The generalized partial credit model (GPCM) (Muraki 1997) was used as basis for the IRT modelling, with Parscale (Scientific Software International [SSI], Skokie, IL, USA) for calibration of the model (Muraki and Bock 1996), as this model is a generalization of other well-known IRT models, such as the Partial Credit Model which can be further constrained to the Rating Scale Model. The GPCM model was also used in the development of the item banks for the other domains of the QLQ-C30. Nevertheless, if one of the more restrictive models had a better fit with the data, that model would have been applied for this specific item bank. Item fit was examined using the item-fit test (Orlando and Thissen 2000), $S-\chi^2$, and bias and indices of fit were evaluated by calculating the difference between expected and observed item responses and the infit and outfit indices, respectively (Bond and Fox 2007). Infit and outfit indices reflect the difference between expected responses of the model and the actual observed responses to an item, and values between 0.7 and 1.3 were deemed acceptable (Wright and Linacre 1994).

Test for Differential Item Functioning

To explore whether items function differently between subgroups of patients, we performed differential item functioning (DIF) analysis. Using ordinal logistic regression methods (performed in SAS version 9.3 (Sas I. I. 2004)), we tested for both uniform and non-uniform DIF with regard to gender (male versus female), age (<40 , $40-49$, $50-59$, $60-69$, ≥ 70 years), country (Denmark, France, The Netherlands, Sweden, Taiwan, the United Kingdom), cancer site (breast, gastrointestinal, genito-urinary, head & neck, lung, other cancers) and stage (stage I-II versus stage III-IV), current treatment (chemotherapy, other current treatment, no current treatment), cohabitation (living with a partner versus living alone), work (working, retired or other), and educational level ($0-10$, $11-13$, $14-16$, >16 years of education). Controlling for the insomnia score estimated in the previously calibrated IRT model, each item was entered as outcome, and group (DIF) variables as independent variables. DIF was considered relevant if $p < 0.001$ and if the regression coefficient for the group variable was $\beta > 0.64$ (Bjorner et al. 1998; Petersen et al. 2003). Next, the possible effect of DIF findings on the estimation of insomnia was evaluated (Hart et al. 2009), by comparing scores obtained with a model accounting for DIF with the scores obtained with the model from the previous step. DIF was only considered problematic if the estimates of these two models differed more than the median standard error of the insomnia estimates (Petersen et al. 2011, 2013b, 2016a; Hart et al. 2009).

Evaluation of Measurement Properties

The information function of the final item bank was calculated, and high measurement precision was defined as an information score > 20 , corresponding to a reliability of $>95\%$ (Nunnally 1978).

Simulations of CAT administration based on the collected data were performed to further evaluate the measurement properties of the final CAT-SL. To evaluate possible savings in sample size, the relative validity (RV, i.e. the ratio of two test statistics for comparing two subgroups) of these CATs as compared to the QLQ-C30 insomnia scale for detecting expected group differences was estimated using t-test statistics (Fayers and Machin 2007). We expected that patients with good emotional functioning would have less sleep problems than patients with poor functioning, and that patients with no/limited pain would have less sleep problems than those with severe pain. Hence, these groups were used to estimate the RVs. From the RVs we estimated the relative sample size required using the CAT measures to obtain the same power as the QLQ-C30 insomnia scale (Petersen et al. 2013a). In addition to the evaluations based on the observed data, we also evaluated the RV of the CATs based on simulated data across different groups and group sizes (Petersen et al. 2013a).

Results

Phase I: Literature Search

A total of 69 items were retrieved from PROQOLID, one item from the EORTC QLQ Item Library, and 85 items were identified with the PubMed search, resulting in a total of 155 items.

Phase II: Selection and Formulation of New Items

In the first step (classification of the items) 97 items were categorised as measuring ‘sleep quantity’, 24 items as measuring ‘sleep quality’ and 34 items as measuring something other than sleep quantity/quality and hence irrelevant for the current purpose (e.g. *‘How concerned are you about your sleeping?’*). In the next step, duplicate items and those that could not be reformulated into the QLQ-C30 item style were deleted, which resulted in the removal of 104 of 121 items. Based on the 17 remaining items, 22 candidate items fitting the QLQ-C30 item style were formulated in step 3 (for 5/17 items two formulations were included, as it was unclear which one would be best, resulting in $17 + 5 = 22$ items). In step 4 it was determined if each item was most relevant for patients with mild, moderate or severe insomnia (e.g. *‘Have you been unable to sleep?’* seemed mostly relevant for patients with severe insomnia problems, while *‘Have you had trouble staying asleep?’* seemed mostly relevant for patients with moderate insomnia). Five items were rated as mostly relevant for patients with severe insomnia, 16 items as mostly relevant for patients with moderate insomnia, and one item as relevant for patients with mild insomnia. Although these results indicated that more items for patients with mild insomnia are warranted, no new items were generated in step 5, as the two reviewers were unable to come up with relevant items (and thought this could possibly be added by external experts). Lastly (step 6), the items were assessed by two groups of experts. Review by EORTC QLG members resulted in the deletion of eight items, reformulation of three items, and the addition of three positively formulated items. After these internal ratings, ten experts from Denmark, Canada, The Netherlands, Sweden and the United Kingdom evaluated the remaining 17 items.

Based on these evaluations, three items were deleted, two rephrased and one new item was added, resulting in 15 candidate items of which two were positively formulated.

Phase III: Pre-Testing

The 15 items resulting from phase II were tested in a sample of 49 cancer patients from Denmark, Italy, The Netherlands, Taiwan and the United Kingdom. The patients varied in tumour type and stage, and treatment (see Table 1 for details on patients' characteristics). In general, the patients had few comments about the items. Based on the patients' comments, two items were deleted because of redundancy, and one item was rephrased because the response format did not fit the question. No new items were generated by patients. Hence, phase III resulted in an item list of 14 items, including the original QLQ-C30 item (i.e. 13 items from phase III plus the original item, resulting in 14 items). See Table 2 for the complete item list.

Phase IV: Field-Testing and Psychometric Analyses

Sample

A large ($n = 1094$) heterogeneous sample of cancer patients was included in phase IV of this study, with patients from Denmark, France, The Netherlands, Sweden, Taiwan and United Kingdom (Table 1).

Descriptive and Basic Statistical Analyses

Response rates for the 14 items ranged between 97.4% and 98.8%, and item means from 1.8–2.6 on a 1–4 scale (ranging from 'not at all' to 'very much'). Each item was rated as difficult, annoying or confusing by a maximum of 9 patients (0.8%). None of the items was rated as upsetting or intrusive. Item 14 ('felt rested') correlated -0.58 with the QLQ-C30 item, and the remaining items correlated $>\pm 0.72$. Based on these results, no items were deleted in this step.

Evaluation of Dimensionality and Local Dependence

Exploratory factor analysis showed that the first factor explained 75% of the total variation (eigenvalue = 9.6). The second factor explained 6% of the variation but had an eigenvalue <1 . All other factors explained $<5\%$ of the variation. Evaluations indicated that the two positively phrased items (#13 and #14) may form a separate factor. The scree plot, however, indicated that one factor may suffice. To obtain a reasonable fit, four items (items #2, #8, #13 and #14) had to be deleted. The remaining 10 items (items 1, 3–7, 9–12) constituted our candidate model and were evaluated using confirmatory methods. The fit indices of this 10-item model were: RMSEA = 0.103, CFI = 0.988 and TLI = 0.998, and one factor explained 80% of the variation. Although the RMSEA was just above 0.1, this set of items was deemed appropriate as a starting point for the IRT analysis.

All of the 45 residual correlations were <0.05 , indicating that there was no local dependence among the retained items.

Table 1 Sociodemographic and clinical characteristics of the patients participating in phase III (pre-testing) and phase IV (field-testing)

Characteristic	Sample phase III (<i>n</i> = 49) N (%) / mean	Sample phase IV (<i>n</i> = 1094) N (%) / mean
Gender		
Male	21 (42.9%)	541 (49.5%)
Female	28 (57.1%)	552 (50.5%)
Missing	–	1 (0.1%)
Country		
Denmark	10 (20.4%)	310 (28.3%)
France	–	163 (14.9%)
Italy	15 (30.6%)	–
The Netherlands	3 (6.1%)	28 (2.6%)
Sweden	–	106 (9.7%)
United Kingdom	11 (22.4%)	383 (35%)
Taiwan	10 (20.4%)	104 (9.5%)
Missing	–	–
Age in years, mean	56	61
Cancer stage		
I-II	18 (36.7%)	580 (53.0%)
III-IV	25 (51.0%)	485 (44.3%)
Unknown	6 (12.2%)	29 (2.7%)
Diagnosis		
Breast	8 (16.3%)	224 (20.5%)
Gastrointestinal	10 (20.4%)	116 (10.6%)
Genito-urinary	5 (10.2%)	237 (21.7%)
Gynaecological	6 (12.2%)	151 (13.8%)
Head & Neck	5 (10.2%)	128 (11.7%)
Lung	3 (6.1%)	46 (4.2%)
Other	9 (18.4%)	190 (17.4%)
Missing	3 (6.1%)	2 (0.2%)
Current treatment		
Chemotherapy	20 (40.8%)	277 (25.3%)
Other treatment	3 (6.1%)	281 (25.7%)
No treatment	24 (49.0%)	533 (48.7%)
Missing	2 (4.1%)	3 (0.3%)
Cohabitation		
Live with partner	30 (61.2%)	803 (73.4%)
Live alone	16 (32.7%)	277 (25.3%)
Missing	3 (6.1%)	14 (1.3%)
Education*		
0-10 years	8 (16.3%)	352 (32.2%)
11-13 years	13 (26.5%)	249 (22.8%)
14-16 years	14 (28.6%)	248 (22.7%)

Table 1 (continued)

Characteristic	Sample phase III (n = 49) N (%) / mean	Sample phase IV (n = 1094) N (%) / mean
More than 16 years	13 (26.5%)	221 (20.2%)
Missing	1 (2.0%)	24 (2.2%)
Work		
Fulltime	16 (32.7%)	272 (24.9%)
Part-time	5 (10.2%)	86 (7.9%)
Retired	15 (30.6%)	570 (52.1%)
Other	11 (22.4%)	159 (14.5%)
Missing	2 (4.1%)	7 (0.6%)

*All patients were literate, i.e. able to understand and read the questionnaires and informed consent form

Item Bank Calibration and Evaluation of Item Fit

Inspecting the mean item scores in relation to the ‘rest score’ (i.e. sum score of all items except the evaluated item) did not reveal problems with monotonicity, meaning that a mean item score did not decrease when values of the ‘rest score’ were increasing. Therefore, a GPCM (as this had a better fit than for example the GRM model (data not shown)) was calibrated to the 10 remaining items. Items #6 and #12 had to be deleted to obtain acceptable fit indices: bias estimates were all very close to zero, indicating no

Table 2 Parameter estimates and fit statistics for the 8 items in the final IRT model

Item	Slope	Location	Item fit p value	Bias	Infit	Outfit
<i>Item 1:</i> Has your sleep been restless?	2.33	0.69	0.20	-0.01	0.88	0.79
<i>Item 2:</i> Have you had trouble falling asleep?	Excluded in phase IV					
<i>Item 3:</i> Have you woken up during the night?	2.18	-0.03	0.42	0.00	0.85	0.82
<i>Item 4:</i> Have you woken up too early?	1.58	0.69	0.01	-0.01	0.92	0.86
<i>Item 5:</i> Have you had trouble staying asleep?	2.67	0.68	0.46	-0.02	0.85	0.73
<i>Item 6:</i> Have the nights been characterized by lack of sleep?	Excluded in phase IV					
<i>Item 7:</i> Have you had trouble sleeping? (QLQ-C30 item)	3.78	0.75	0.18	-0.02	0.74	0.56
<i>Item 8:</i> Have you had trouble returning to sleep if you woke up during the night?	Excluded in phase IV					
<i>Item 9:</i> Have you had trouble getting a good night’s sleep?	3.70	0.48	0.05	0.00	0.64	0.49
<i>Item 10:</i> Have you woken up for long periods during the night?	2.53	0.83	0.18	-0.01	0.86	0.73
<i>Item 11:</i> Have you felt tired (not rested) when you woke up?	1.44	0.59	0.39	-0.01	0.92	0.91
<i>Item 12:</i> Have you slept poorly?	Excluded in phase IV					
<i>Item 13:</i> Have you slept well?	Excluded in phase IV					
<i>Item 14:</i> Have you felt rested when you woke up?	Excluded in phase IV					

systematic bias, and the RMSEAs were between 0.37 and 0.62, indicating some variation in the precision of predicting item responses but not problematic (i.e. all <1), and p values >0.01 (see Table 2 for details). The infits ranged from 0.64–0.92 and the outfit from 0.49–0.91. Since values between 0.7 and 1.3 were deemed acceptable, items #7 (original QLQ-C30 item) and #9 were problematic and probably reflect a tendency for redundancy or local dependence. However, removing additional items resulted in poor fit of one or more of the retained items. It was therefore decided to continue with the 8-item pool in further analyses. The final 8-item model fulfilled all criteria for acceptable fit: RMSEA = 0.080 and both CFI and TLI were >0.99 .

Test for Differential Item Functioning

Five items showed significant DIF, but only for age (one item) and country (four items) (Table 3). Detailed evaluations showed that the effect of DIF had negligible impact on insomnia estimation (data not shown). Therefore, all items were retained in the model.

Evaluation of Measurement Properties

Since re-addition of the discarded items to the model resulted in significantly poorer model fit, the measurement properties of the item bank including eight items was evaluated. Figure 1 shows the test information function for the eight items in the final model, as well as for the original QLQ-C30 sleeplessness item. Insomnia scores ranged from -1.84 ('not at all' on all items) to 2.59 ('very much' on all items). The total score has very high measurement precision (reliability >0.95) for a small interval only (0.6–

Table 3 Results of the differential item functioning (DIF) analysis. For each group variable (e.g. country) one Beta is presented summarizing the potential DIF, and the reference categories that were used in each case. Conditional on the estimated insomnia scale score, a Beta coefficient >0 indicates that the second group in the bracket has a higher risk of reporting problems on the item. Beta <0 indicates that the first group has the highest risk

Item	DIF	β	P value
<i>Item 1:</i> Has your sleep been restless?	Country	0.84 (Denmark and Sweden and France versus United Kingdom and the Netherlands and Taiwan)	<0.0001
<i>Item 3:</i> Have you woken up during the night?	No DIF		
<i>Item 4:</i> Have you woken up too early?	Country	-0.83 (France and The Netherlands versus United Kingdom)	<0.0001
<i>Item 5:</i> Have you had trouble staying asleep?	No DIF		
<i>Item 7:</i> Have you had trouble sleeping?	Country	1.12 (Taiwan versus United kingdom)	<0.001
<i>Item 9:</i> Have you had trouble getting a good night's sleep?	Country	0.86 (Denmark and France versus all other countries)	<0.0001
<i>Item 10:</i> Have you woken up for long periods during the night?	No DIF		
<i>Item 11:</i> Have you felt tired (not rested) when you woke up?	Age	-0.92 (age groups ≤ 50 versus age groups >50 years)	<0.0001

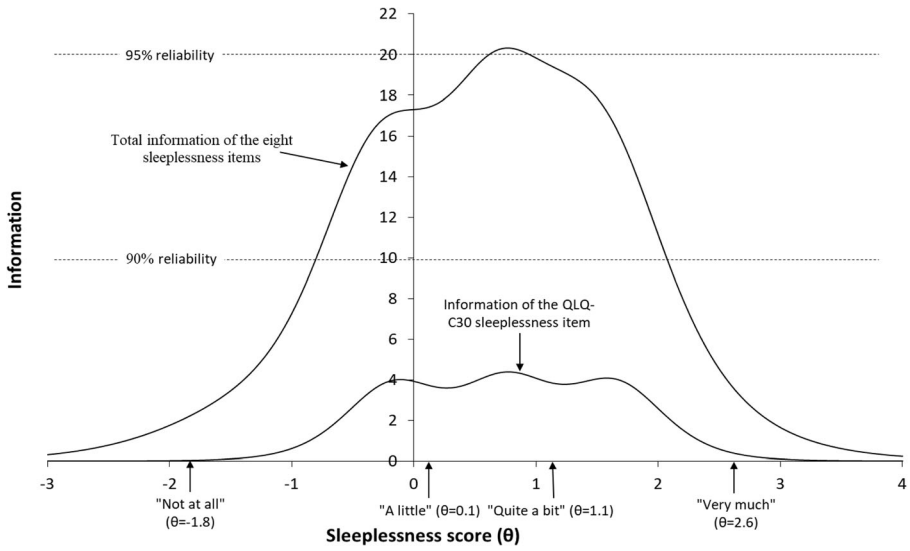


Fig. 1 Test information function for the 8 items in the final model and the original QLQ-C30 sleeplessness item. Scores for all response options (ranging from ‘not at all’ to ‘very much’) are presented. The θ of -1.84 reflects the score when ‘not at all’ is answered on all eight items, while the θ of 2.59 reflects the score when ‘very much’ is answered on all items. In addition, the level of measurement precision is presented on the y-axis, with 90% reliability as the cut-off for good measurement precision

0.9), but has a reliability >0.90 in a much broader interval between -0.8 and 2.1 (three standard deviation units). This means that the item bank is less precise for patients without insomnia or those with severe insomnia. The item bank does result in markedly higher measurement precision than the original item, with a reliability >0.5 across the whole continuum.

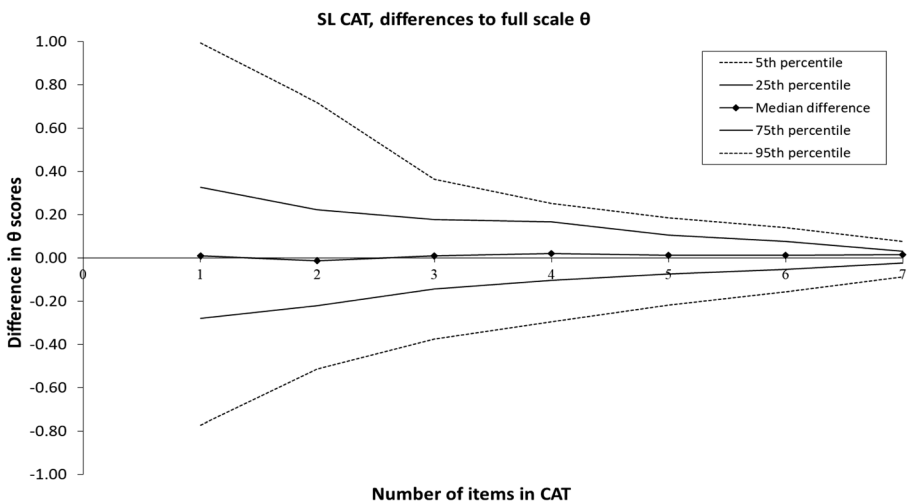


Fig. 2 Median and percentiles for differences between CAT versions with different lengths (θ estimate) and the entire item bank (full-length θ)

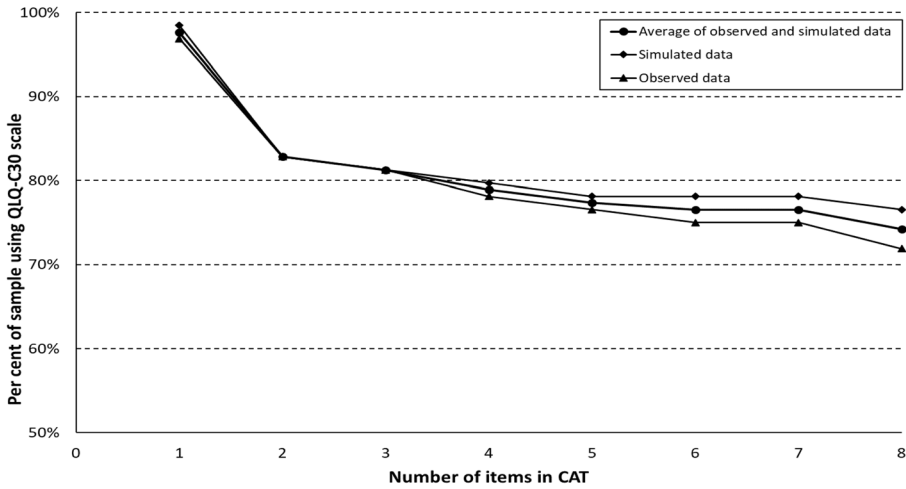


Fig. 3 The relative required sample size in percentage using CAT measurement compared to using the QLQ-C30 sleeplessness item for both observed data and simulated data. A smaller percentage indicates that smaller sample sizes are required without reducing the power

Figure 2 shows that for CATs of all lengths, the median insomnia score was very similar to the score based on the entire item bank (median differences <0.02). Possible savings in sample size were determined by known group comparisons based on both the observed and simulated data. Observed and simulated data resulted in quite similar findings (Fig. 3), with the observed data indicating slightly larger savings. On average, across the two methods, the savings were between approximately 15–25% when using two or more items.

Discussion

In this study we developed and evaluated an item bank for CAT measurement of insomnia. The resulting EORTC CAT-SL item bank consists of eight items, including the original item, and is applicable to a heterogeneous group of cancer patients across countries.

Overall, the measurement precision of the item bank was higher than the original QLQ-C30 SL item alone, across the whole continuum. Relatively high measurement precision was observed for patients with moderate levels of insomnia. This is in line with the finding that all items in the item bank were deemed particularly relevant for patients with moderate insomnia. Importantly, the measurement precision for patients without insomnia or with severe insomnia was suboptimal. In clinical practice this means that more detailed information would be particularly useful for these patients. To further enhance measurement precision of the item bank, new items primarily relevant for these latter patients could be constructed and added. In contrast, for patients without insomnia this may be irrelevant from a clinical perspective, because absence of insomnia is ‘normal’, and interventions would not be necessary. For patients with severe insomnia this could be relevant, as these patients would be eligible for further treatment.

The final item bank consists of eight items and is therefore relatively small. In comparison, the EORTC item banks for cognitive functioning and fatigue both consist of 34 items (Dirven et al. 2017; Petersen et al. 2013a, 2018). However, only two items are sufficient to improve the measurement precision and reduce sample sizes when compared to single item assessment. Moreover, the items in the item bank not only reflect the quantity of sleep, but also the quality of sleep. Therefore, the whole construct of insomnia as we conceptualised it is covered in this item bank. It should be noted though, that the literature search included articles up to 2007 only as this was the first step in the developmental process, and possible items assessing other concepts of insomnia may have been missed, although we think that this is unlikely. Whether the level of insomnia as measured with the CAT-SL is similar to the level of insomnia as measured with other validated insomnia tools (whether or not specifically developed for cancer patients) is not yet known. Future studies using these instruments in parallel may clarify this. Of note, the eight items in the final item bank are all negatively phrased. Although three positively phrased items were constructed during the developmental process, the two items tested in phase IV loaded on a different factor. Although mixing positively and negatively phrased items may lessen the acquiescent response bias, it may also cause problems with reliability and validity, questioning the use of both positively and negatively phrased items in one questionnaire. (Roszkowski and Soven 2010) Lastly, in this study we chose to comply with the standard EORTC response format to formulate our items, although this is not required for IRT. In future studies it may be worth exploring the impact of including items with different formats. Indeed, a response format including the options 'yes' or 'no' would also fit the EORTC items, as these are formulated in such a way that they can enquire about the severity of an issue, but also presence versus absence.

The fit indices of the model with the eight items were not optimal, partly because of the inclusion of the original item. However, since this item could not be deleted (a policy of the EORTC Quality of Life Group), further evaluation of the item bank was based on these eight items and showed that all could be retained in a unidimensional item bank, with generally acceptable item fit. Even though quite some items (5/8) showed DIF, this had negligible impact on the estimation of insomnia (i.e. if the scores obtained with the model accounting for DIF were similar to the model ignoring any possible DIF, the DIF for that item was considered not to have practical importance). This means that insomnia scores based on CAT-SL are not impacted by patient characteristics (i.e. gender, age, country of residence, cancer site and stage, current treatment, cohabitation, work and educational level), and can therefore be compared across studies. As all EORTC questionnaires are developed cross-culturally, including many languages, it may also be that DIF occurred due to linguistic reasons.

The estimated savings in sample size when using CAT varies between the domains of the EORTC QLQ-C30, ranging from 15% for social functioning to 55% for dyspnoea (Petersen et al. 2018). For the CAT-SL, savings in sample sizes between approximately 15–25% can be achieved when administering two or more items with CAT. Thus, fewer patients are needed to include in studies with insomnia as the primary outcome. The response burden would be slightly higher for patients, as at least two

items on insomnia need to be administered with CAT to all patients, compared to a single item in the original QLQ-C30 questionnaire. On the other hand, considering that HRQoL assessments comprises evaluation of multiple domains, the overall response burden may be reduced for patients, as more concise assessment of many domains is possible with the EORTC CAT measurement system.

The increased interest in CAT versions of patient-reported outcomes is based on their potential for yielding more precise and efficient assessments as compared to traditional instruments based on classical test theory. Enhanced measurement efficiency and precision can be useful in both clinical studies and clinical practice. In clinical studies, smaller sample sizes and reduced response burden can contribute to higher compliance rates. Poor compliance with HRQoL assessment is a major problem in cancer trials (Bernhard et al. 1998), especially during the follow-up period, hampering the interpretation of results. In clinical practice, CAT can be used to monitor patients' functioning during the disease trajectory, or to facilitate patient-physician communication (Detmar et al. 2002; Velikova et al. 2004, 2010). The burden of repeated assessment of HRQoL in daily practice may be attenuated by CAT measurement, by reducing the number of items that needs to be completed by patients and by presenting different sets of items at each assessment point, thus avoiding feelings of repetitiveness.

In conclusion, we have developed an eight item insomnia item bank for CAT measurement relevant to various cancer patients in different countries. This new item bank can be used in combination with the CAT versions of all of the other QLQ-C30 domains that have been developed or as 'short-forms' that can even be used in paper and pencil questionnaires (<http://qol.eortc.org/projectqol/eortc-cat/>). The resulting CAT instrument, the EORTC CAT Core, is now validated in an independent dataset. (Petersen et al. 2018) With the EORTC CAT Core instrument, precise and efficient assessment of HRQoL can be carried out, with clear utility in both clinical studies and clinical practice.

Acknowledgements This study was funded by grants from the EORTC Quality of Life Group, and the manuscript was developed on behalf of the Group. The authors would like to thank the participating patients for their time and effort.

Funding Information The study was funded by grants from the EORTC Quality of Life Group, and the paper was developed on behalf of the group.

Compliance with Ethical Standards The procedures performed in this study are in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of Interest All authors declare no conflict of interest with respect to this work.

Ethical Approval Danish Ethical Board, registration number (KF) 01318568; UK National Research Ethics Service, registration number 07/Q0204/65; Ethics Committee of Sant' Andrea Hospital Rome, registration number 827. The French, Dutch and Swedish ethical committees were notified of the study, but formal submission and registration of the study was not necessary at the time of conduct.

Informed Consent Informed consent was obtained from all individual participants included in the study.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85(5), 365–376.
- Arndt, V., Koch-Gallenkamp, L., Jansen, L., Bertram, H., Eberle, A., Holleczeck, B., et al. (2017). Quality of life in long-term and very long-term cancer survivors versus population controls in Germany. *Acta Oncologica*, 56(2), 190–197. <https://doi.org/10.1080/0284186x.2016.1266089>.
- Bernhard, J., Cella, D. F., Coates, A. S., Fallowfield, L., Ganz, P. A., Moinpour, C. M., et al. (1998). Missing quality of life data in cancer clinical trials: Serious problems and challenges. *Statistics in Medicine*, 17(5–7), 517–532.
- Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology*, 51(11), 1189–1202.
- Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the headache impact test (HIT). *Quality of Life Research*, 12(8), 913–933.
- Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research*, 16(Suppl 1), 95–108. <https://doi.org/10.1007/s11136-007-9168-6>.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.
- Browne, M. W., & Cudek, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21, 230e258.
- Cattell, R. B. (1966). Scree test for number of factors. *Multivariate Behavioral Research*, 1, 245.
- Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, 16(Suppl 1), 133–141. <https://doi.org/10.1007/s11136-007-9204-6>.
- Chen, D., Yin, Z., & Fang, B. (2018). Measurements and status of sleep quality in patients with cancers. *Support Care Cancer*, 26(2), 405–414. <https://doi.org/10.1007/s00520-017-3927-x>.
- Costa, A. R., Fontes, F., Pereira, S., Goncalves, M., Azevedo, A., & Lunet, N. (2014). Impact of breast cancer treatments on sleep disturbances – A systematic review. *Breast*, 23(6), 697–709. <https://doi.org/10.1016/j.breast.2014.09.003>.
- Detmar, S. B., Muller, M. J., Schormagel, J. H., Wever, L. D., & Aaronson, N. K. (2002). Health-related quality-of-life assessments and patient-physician communication: A randomized controlled trial. *JAMA*, 288(23), 3027–3034.
- Dirven, L., Taphoorn, M. J. B., Groenvold, M., Habets, E. J. J., Aaronson, N. K., Conroy, T., et al. (2016). Development of an item bank for computerized adaptive testing of self-reported cognitive difficulty in cancer patients. *Neuro-Oncology Practice*, 4(3), 189–196. <https://doi.org/10.1093/nop/npw026>.
- Dirven, L., Groenvold, M., Taphoorn, M. J. B., Conroy, T., Tomaszewski, K. A., Young, T., et al. (2017). Psychometric evaluation of an item bank for computerized adaptive testing of the EORTC QLQ-C30 cognitive functioning dimension in cancer patients. *Quality of Life Research*, 26(11), 2919–2929. <https://doi.org/10.1007/s11136-017-1648-8>.
- Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: The challenges for health outcomes assessment. *Quality of Life Research*, 16(Suppl 1), 187–194. <https://doi.org/10.1007/s11136-007-9197-1>.
- Fayers, P. M., & Machin, D. (2007). *Quality of life. The assessment, analysis and interpretation of patient-reported outcomes* (2nd ed.). Chichester: Wiley.
- Gamper, E. M., Groenvold, M., Petersen, M. A., Young, T., Costantini, A., Aaronson, N., et al. (2014). The EORTC emotional functioning computerized adaptive test: Phases I-III of a cross-cultural item bank development. *Psychooncology*, 23(4), 397–403. <https://doi.org/10.1002/pon.3427>.

- Gamper, E. M., Petersen, M. A., Aaronson, N., Costantini, A., Giesinger, J. M., Holzner, B., et al. (2016). Development of an item bank for the EORTC role functioning computer adaptive test (EORTC RF-CAT). *Health and Quality of Life Outcomes*, 14, 72. <https://doi.org/10.1186/s12955-016-0475-x>.
- Giesinger, J. M., Aa Petersen, M., Groenvold, M., Aaronson, N. K., Arraras, J. I., Conroy, T., et al. (2011). Cross-cultural development of an item list for computer-adaptive testing of fatigue in oncological patients. *Health and Quality of Life Outcomes*, 9, 19. <https://doi.org/10.1186/1477-7525-9-19>.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: SAGE.
- Harris, B., Ross, J., & Sanchez-Reilly, S. (2014). Sleeping in the arms of cancer: A review of sleeping disorders among patients with cancer. *Cancer Journal*, 20(5), 299–305. <https://doi.org/10.1097/ppo.000000000000067>.
- Hart, D. L., Deutscher, D., Crane, P. K., & Wang, Y. C. (2009). Differential item functioning was negligible in an adaptive test of functional status for patients with knee impairments who spoke English or Hebrew. *Quality of Life Research*, 18(8), 1067–1083. <https://doi.org/10.1007/s11136-009-9517-8>.
- Induru, R. R., & Walsh, D. (2014). Cancer-related insomnia. *The American Journal of Hospice & Palliative Care*, 31(7), 777–785. <https://doi.org/10.1177/1049909113508302>.
- Itani, O., Kaneita, Y., Munezawa, T., Mishima, K., Jike, M., Nakagome, S., et al. (2016). Nationwide epidemiological study of insomnia in Japan. *Sleep Medicine*, 25, 130–138. <https://doi.org/10.1016/j.sleep.2016.05.013>.
- Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, 7, 109. <https://doi.org/10.3389/fpsyg.2016.00109>.
- Johnson, C., Aaronson, N. K., Blazeby, J., Bottomley, A., Fayers, P., Koller, M., et al. (2011). *Guidelines for developing Quality of Life Questionnaires* (4th ed.). Brussels: EORTC Publications.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- Kulis, D., Bottomley, A., Velikova, G., Greimel, E. & Koller, M, on behalf of the EORTC Quality of Life Group (2016). *EORTC Quality of Life Group Translation Procedure* (4th edition). Brussels: EORTC Publications.
- Lowery-Allison, A. E., Passik, S. D., Cribbet, M. R., Reinsel, R. A., O’Sullivan, B., Norton, L., et al. (2017). Sleep problems in breast cancer survivors 1-10 years posttreatment. *Palliative & Supportive Care*, 1–10. <https://doi.org/10.1017/s1478951517000311>.
- Mercadante, S., Adile, C., Ferrera, P., Masedu, F., Valenti, M., & Aielli, F. (2017). Sleep disturbances in advanced cancer patients admitted to a supportive/palliative care unit. *Support Care Cancer*, 25(4), 1301–1306. <https://doi.org/10.1007/s00520-016-3524-4>.
- Morin, C. M. (1993). *Insomnia: Psychological assessment and management*. New York: Guilford Press.
- Muraki, E. (1997). *A generalized partial credit model (Handbook of modern item response theory)*. Berlin: Springer.
- Muraki, E., & Bock, R. D. (1996). *PARSCALE – IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago: Scientific Software International, Inc..
- Muthen, L. K., & Muthen, B. O. (2002). *Mplus user’s guide* (2nd ed.). Los Angeles: Muthen & Muthen.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- O’Donnell, J. F. (2004). Insomnia in cancer patients. *Clinical Cornerstone*, 6(Suppl 1D), S6–S14.
- Orlando, E., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64.
- Pallesen, S., Sivertsen, B., Nordhus, I. H., & Bjorvatn, B. (2014). A 10-year trend of insomnia prevalence in the adult Norwegian population. *Sleep Medicine*, 15(2), 173–179. <https://doi.org/10.1016/j.sleep.2013.10.009>.
- Petersen, M. A., Groenvold, M., Bjorner, J. B., Aaronson, N., Conroy, T., Cull, A., et al. (2003). Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire. *Quality of Life Research*, 12(4), 373–385.
- Petersen, M. A., Groenvold, M., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., et al. (2010). Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 dimensions - general approach and initial results for physical functioning. *European Journal of Cancer*, 46(8), 1352–1358. <https://doi.org/10.1016/j.ejca.2010.02.011>.
- Petersen, M. A., Groenvold, M., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., et al. (2011). Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. *Quality of Life Research*, 20(4), 479–490. <https://doi.org/10.1007/s11136-010-9770-x>.

- Petersen, M. A., Aaronson, N. K., Arraras, J. I., Chie, W. C., Conroy, T., Costantini, A., et al. (2013a). The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency. *Journal of Clinical Epidemiology*, *66*(3), 330–339. <https://doi.org/10.1016/j.jclinepi.2012.09.010>.
- Petersen, M. A., Giesinger, J. M., Holzner, B., Arraras, J. I., Conroy, T., Gamper, E. M., et al. (2013b). Psychometric evaluation of the EORTC computerized adaptive test (CAT) fatigue item pool. *Quality of Life Research*, *22*(9), 2443–2454. <https://doi.org/10.1007/s11136-013-0372-2>.
- Petersen, M. A., Gamper, E. M., Costantini, A., Giesinger, J. M., Holzner, B., Johnson, C., et al. (2016a). An emotional functioning item bank of 24 items for computerized adaptive testing (CAT) was established. *Journal of Clinical Epidemiology*, *70*, 90–100. <https://doi.org/10.1016/j.jclinepi.2015.09.002>.
- Petersen, M. A., Aaronson, N. K., Chie, W. C., Conroy, T., Costantini, A., Hammerlid, E., et al. (2016b). Development of an item bank for computerized adaptive test (CAT) measurement of pain. *Quality of Life Research*, *25*(1), 1–11. <https://doi.org/10.1007/s11136-015-1069-5>.
- Petersen, M. A., Aaronson, N. K., Arraras, J. I., Chie, W. C., Conroy, T., Costantini, A., et al. (2018). The EORTC CAT Core—the computer adaptive version of the EORTC QLQ-C30 questionnaire. *European Journal of Cancer*, *100*, 8–16. <https://doi.org/10.1016/j.ejca.2018.04.016>.
- Robertson, M. E., McSherry, F., Herndon, J. E., & Peters, K. B. (2016). Insomnia and its associations in patients with recurrent gliial neoplasms. *Springerplus*, *5*(1), 823. <https://doi.org/10.1186/s40064-016-2578-6>.
- Roszkowski, M. J., & Soven, M. (2010). Shifting gears: Consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education*, *35*(1), 113.
- SAS Institute Inc. (2004). *SAS/STAT 9.1 user's guide*. Cary NC: SAS Institute Inc
- Sateia, M. J., & Lang, B. J. (2008). Sleep and cancer: Recent developments. *Current Oncology Reports*, *10*(4), 309–318.
- Savard, J., & Morin, C. M. (2001). Insomnia in the context of cancer: A review of a neglected problem. *Journal of Clinical Oncology*, *19*(3), 895–908. <https://doi.org/10.1200/jco.2001.19.3.895>.
- Savard, J., Villa, J., Ivers, H., Simard, S., & Morin, C. M. (2009). Prevalence, natural course, and risk factors of insomnia comorbid with cancer over a 2-month period. *Journal of Clinical Oncology*, *27*(31), 5233–5239. <https://doi.org/10.1200/jco.2008.21.6333>.
- Thamsborg, L. H., Petersen, M. A., Aaronson, N. K., Chie, W. C., Costantini, A., Holzner, B., et al. (2015). Development of a lack of appetite item bank for computer-adaptive testing (CAT). *Support Care Cancer*, *23*(6), 1541–1548. <https://doi.org/10.1007/s00520-014-2498-3>.
- van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York: Springer.
- Velikova, G., Booth, L., Smith, A. B., Brown, P. M., Lynch, P., Brown, J. M., et al. (2004). Measuring quality of life in routine oncology practice improves communication and patient well-being: A randomized controlled trial. *Journal of Clinical Oncology*, *22*(4), 714–724. <https://doi.org/10.1200/jco.2004.06.078>.
- Velikova, G., Keding, A., Harley, C., Cocks, K., Booth, L., Smith, A. B., et al. (2010). Patients report improvements in continuity of care when quality of life assessments are used routinely in oncology practice: Secondary outcomes of a randomised controlled trial. *European Journal of Cancer*, *46*(13), 2381–2388. <https://doi.org/10.1016/j.ejca.2010.04.030>.
- World Health Organization (2018). International classification of diseases for mortality and morbidity statistics (11th Revision). Retrieved from <https://icd.who.int/browse11/l-m/en>.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370–371.

Affiliations

Linda Dirven^{1,2} · Morten Aa. Petersen³ · Neil K. Aaronson⁴ · Wei-Chu Chie⁵ · Thierry Conroy⁶ · Anna Costantini⁷ · Eva Hammerlid⁸ · Galina Velikova⁹ · Irma M. Verdonck-de Leeuw^{10,11} · Teresa Young¹² · Mogens Groenvold^{3,13}

✉ Linda Dirven
l.dirven@lumc.nl

¹ Department of Neurology, Leiden University Medical Center, PO BOX 9600, 2300 RC Leiden, The Netherlands

² Department of Neurology, Haaglanden Medical Center, The Hague, The Netherlands

³ The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, University of Copenhagen, Copenhagen, Denmark

⁴ Division of Psychosocial Research & Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

⁵ Department of Family Medicine, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

⁶ Medical Oncology Department, Institut de cancérologie de Lorraine, Vandoeuvre-lès-Nancy, France

⁷ Psychooncology Unit, Sant' Andrea Hospital, Faculty of Medicine and Psychology, Sapienza University, Rome, Italy

⁸ Department of Otolaryngology Head and Neck Surgery, Sahlgrenska University Hospital, Göteborg University, Göteborg, Sweden

⁹ Leeds Institute of Cancer and Pathology, Faculty of Medicine and Health, University of Leeds, Leeds, UK

¹⁰ Department of Otolaryngology – Head & Neck Surgery, Amsterdam University Medical Centers (location VU University Medical Center), Cancer Center Amsterdam, Amsterdam, The Netherlands

¹¹ Department of Clinical, Neuro and Development Psychology, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

¹² East & North Hertfordshire NHS Trust incorporating Mount Vernon Cancer Centre, Northwood, Middlesex, UK

¹³ Department of Public Health, University of Copenhagen, Copenhagen, Denmark