



# Measuring fatigue in cancer patients: a common metric for six fatigue instruments

Michael Friedrich<sup>1</sup> · Andreas Hinz<sup>1</sup> · Susanne Kuhnt<sup>1</sup> · Thomas Schulte<sup>2</sup> · Matthias Rose<sup>3,4</sup> · Felix Fischer<sup>3</sup>

Accepted: 21 February 2019 / Published online: 27 February 2019  
© Springer Nature Switzerland AG 2019

## Abstract

**Purpose** Fatigue is one of the most disabling symptoms in cancer patients. Many instruments exist to measure fatigue. This variety impedes the comparison of data across studies or to the general population. We aimed to estimate a common metric based on six different fatigue instruments (EORTC QLQ-C30 subscale fatigue, EORTC QLQ-FA12, MFI subscale General Fatigue, BFI, Fatigue Scale, and Fatigue Diagnostic Interview Guide) to convert the patients' scores from one of the instruments to another. Additionally, we linked the common metric to the general population.

**Methods** For  $n = 1225$  cancer patients, the common metric was estimated using the Item Response Theory framework. The linking between the common metric of the patients and the general population was estimated using linear regression.

**Results** The common metric was based on a model with acceptable fit ( $CFI = 0.94$ ,  $SRMR = 0.06$ ). Based on the standard error of measurement the reliability coefficients of the questionnaires ranged from 0.80 to 0.95. The common metric of the six questionnaires, also linked to the general population, is reported graphically and in supplementary crosswalk tables.

**Conclusions** Our study enables researchers and clinicians to directly compare results across studies using different fatigue questionnaires and to assess the degree of fatigue with respect to the general population.

**Keywords** Fatigue · Common metric · Cancer · General population

## Abbreviations

BFI9	Sum of the nine items of the BFI (item range = 0–10, scale range = 0–10)
CFI	Comparative fit index
Chi <sup>2</sup>	Chi-squared statistic
CI	Confidence interval
df	Degrees of freedom

DIG11	Sum of the eleven binary symptom items of the DIG for Fatigue (item range: 0/1, scale range = 0–11)
FA12	Sum of the twelve items of the EORTC QLQ-FA12 (item range = 1–4, scale range = 0–100)
FA3	Sum of the three-item fatigue scale of the EORTC QLQ-C30 (item range = 1–4, scale range = 0–100)
FS11	Sum of the eleven items of the FS (item range = 0–3, scale range = 0–33)

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11136-019-02147-3>) contains supplementary material, which is available to authorized users.

✉ Michael Friedrich  
michael.friedrich@medizin.uni-leipzig.de

Andreas Hinz  
andreas.hinz@medizin.uni-leipzig.de

Susanne Kuhnt  
susanne.kuhnt@medizin.uni-leipzig.de

Thomas Schulte  
tschulte@badoexen.de

Matthias Rose  
matthias.rose@charite.de

Felix Fischer  
felix.fischer@charite.de

<sup>1</sup> Department of Medical Psychology and Medical Sociology, University of Leipzig, Leipzig, Germany

<sup>2</sup> Rehabilitation Clinic Bad Oexen, Bad Oeynhausen, Germany

<sup>3</sup> Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

<sup>4</sup> Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

GP	General population
M	Mean
MFI4	Sum of the four-item General Fatigue scale of the MFI-20 (item range = 1–5, scale range = 4–20); RMSEA: Root mean square error of approximation
SD	Standard deviation
SRMR	Standardized root mean square residual
TLI	Tucker–Lewis Index
T-scores(GP)	Estimation of T-scores for the German general population (mean = 50, standard deviation = 10)

## Background

Fatigue is a subjective feeling of overwhelming and exhausting tiredness. In patients suffering from chronic diseases, e.g., cardiovascular diseases, neurological disorders, arthritis, chronic obstructive pulmonary diseases or cancer, fatigue can persist for months, debilitates the patient's functioning and interferes with self-care and quality of life [1–4]. In patients with cancer, it is the most frequent complaint [5]. Prevalence rates during the treatment range from 25 to 99%, and fatigue has been reported to occur even years after treatment [6].

Fatigue can be described as “a multidimensional phenomenon, with physical, emotional, and cognitive manifestations” [7], but there is no universally accepted definition [8] mostly because the pathophysiological mechanisms underlying the development of fatigue are unclear [9]. Therefore, it is not clear, if fatigue has to be differentiated in separate dimensions (e.g., physical, emotional and/or cognitive) with separate causes like in a *multiple symptom concept*, or if fatigue has to be understood as the one common cause for different kinds of tiredness like in a *multidimensional concept*. In the first case, it is assumed that different forms of tiredness are caused by different sources; consequently, the separate dimensions are expected to behave differently and have to be differentiated to enable different treatments. In the second case (multidimensional concept), it is assumed that different kinds of tiredness are caused by only one common source, thus the dimensions should not behave differently and do not have to be differentiated. Despite the multidimensional concept is being based on a consensus of experts [10, 11], there are studies indicating that fatigue should be considered as a multiple symptom concept [11, 12].

Against this background, it is not surprising that a recent systematic review comprising 40 instruments to measure fatigue stated that the definition of dimension and domain seems to be a matter of author's opinion [13]. Hence the comparability of data across studies is hampered in two ways: the question of the underlying concept and the large

quantity of instruments. In this study, we will assume the multidimensional concept, and address the question of converting scores into one another using a selection of six different instruments.

Recently, construct-based, instrument-dependent common metrics for specific outcomes have been developed in order to facilitate comparability of data collected with different instruments, for example, for anxiety [14], depression [15, 16], physical function [17–19], and also for fatigue [20–22].

Methods to develop common metrics are described by the PROsetta Stone Project funded by the National Institute of Health (NIH)/ National Cancer Institute (NCI) [23]. Details can be found under <http://www.prosettastone.org/AnalysisReport>. The general approach is to model item responses to different instruments as probabilistic functions of a common, single latent trait-based unidimensional IRT models. Linking procedures using multidimensional IRT models have been developed, but are seldom applied [24]. A more detailed overview of different IRT linking procedures can be found in [25]. First validation studies in independent samples show that differences between latent trait estimates derived from different instruments are small on sample level [26, 27].

Hypothesizing that there is one common cause for different dimensions of tiredness (multidimensional concept), Lai et al. showed that fatigue can be considered sufficiently unidimensional [28], based on a 72-item fatigue bank. Using the same statistical technique of bifactor modeling, Cella et al. [29] found support for unidimensionality of the 13 items of the Functional Assessment of Chronic Illness Therapy—Fatigue (FACIT-F). These studies provide evidence that fatigue instruments can be calibrated on a unidimensional common metric to facilitate the comparison of results from studies using diverse instruments even if they assess different dimensions of fatigue.

The main aim of this study was to estimate a common metric to convert patient's scores of six different fatigue instruments to one another. In many cases, it is also useful to compare the results with the general population (GP), which is a meaningful reference, to judge fatigue severity both for individuals and for groups of patients. Hence, we additionally anchor the common metric to the GP.

## Methods

### Sample and setting

The study participants were recruited in an oncological rehabilitation clinic in Germany. The rehabilitation programs, generally lasting 3 weeks, help restore patient's physical and psychosocial functioning. Patients receive physical fitness exercises, physiotherapy, relaxation techniques,

and counseling concerning occupational and nutritional behavior. Inclusion criteria for this study were age 18 years and above, sufficient ability of the German language, and absence of severe cognitive impairment. Informed consent was obtained from the study participants after they were given a full explanation of the purpose and nature of the data collection and storage. The study was approved by the Ethics Committee of the Leipzig University. A total of 1547 consecutive patients were asked to participate, and 1225 (79.2%) of them agreed to take part in the study.

## Instruments

The following socio-demographic and medical characteristics were measured: *gender*, *age (years)*, *education*, *occupation*, *partnership*, *diagnosis (ICD-10)*, and *time since diagnosis (months)*. All data concerning socio-demographic and medical characteristics as well as the data of all instruments mentioned below were based on self-reported information. In our analysis, we included the items from six measures of fatigue. The criteria for the selection of these measures were the frequency of the use in oncological research and treatment, novelty (especially the FA12 and the DIG11), and the availability of normative data sets for the linking to the general population (FA3, FA12, MFI4, FS11). The abbreviations of the questionnaires mentioned in parentheses are described below.

The *Quality-of-Life Questionnaire Core 30*, developed by the European Organization for Research and Treatment of Cancer (EORTC QLQ-C30 [30]), contains a three-item symptom scale for fatigue that we call FA3. The response options range from 1 to 4 with higher values for higher fatigue levels. The sum of the items ranges from 4 to 12; it is then transformed to a range from 0 to 100.

The *EORTC QLQ-FA12* is an additional module of the EORTC QLQ-C30 that measures specific forms of fatigue (physical, emotional, and cognitive fatigue) and its interference with daily activities and social life [31]. Kecke et al. [32] and Hinz et al. [33] suggested a summary score using all twelve items. The sum ranges from 12 to 48, and is transformed to a range from 0 to 100. To indicate this scale, we use the abbreviation FA12.

From the *Multidimensional Fatigue Inventory (MFI-20)*, we used its General Fatigue scale with four items (henceforth called MFI4). Response options range from 1 to 5 [34], and the sum ranges from 4 to 20.

The *Brief Fatigue Inventory (BFI9)* contains nine items which assess the actual severity of fatigue and fatigue-related impairment during the past 24 h [35]. Response options range from 0 to 10 with higher values for higher severity or interference. The sum ranges from 0 to 90; it is transformed to a range from 0 to 10.

The Fatigue Scale (FS11) measures physical and mental fatigue with eleven items [36]. The questionnaire is also known as the Fatigue Questionnaire [37] or the Chalder Fatigue Scale. Initially, it consisted of 14 items [38], subsequently revised to a version with 11 items [39]. Response options range from 0 to 3 with higher values corresponding to higher fatigue levels, the sum ranges from 0 to 33.

The *Diagnostic Interview Guide (DIG11)* contains eleven binary items (0 = no, 1 = yes) corresponding to the eleven symptoms that are named in the proposed ICD-10 criteria for fatigue [7]. To our knowledge, the questionnaire is not yet validated as an instrument to measure severity of fatigue as a sum score. We calculated the sum of the items, ranging from 0 to 11, indicating higher symptom burden with higher values.

Cronbach's alpha coefficients of these measures in our sample were:  $\alpha(\text{FA3}) = 0.89$ ,  $\alpha(\text{FA12}) = 0.94$ ,  $\alpha(\text{MFI4}) = 0.81$ ,  $\alpha(\text{BFI9}) = 0.93$ ,  $\alpha(\text{FS11}) = 0.92$ , and  $\alpha(\text{DIG11}) = 0.86$ .

## General population

Additionally to the sample of study participants, we used fatigue assessments from four different GP surveys to anchor the common metric. Each of the surveys included at least one of the fatigue questionnaires used in this study. The *first survey* included the quality-of-life questionnaire EORTC QLQ-C30 with the 3-item fatigue subscale FA3. The sample consisted of two subsamples with 1993 [40] and 2448 [41] participants, respectively. The *second survey* included the MFI-20 ( $n = 1993$ ) [42]. In our analysis, we only used the General Fatigue scale of the MFI-20. The *third survey* included the FA12 ( $n = 2424$ ) [33]. The *fourth survey* included the Fatigue Scale FS11 ( $n = 2464$ ) [36].

## Statistical analyses

Missing values can lead to biased parameter estimates, inflated standard errors, loss of information, and weakened generalizability of the results [43]. Therefore, we estimated missing values on item level using the Expectation Maximization algorithm [44] using IBM SPSS Statistics 23.

## Common metric of the six fatigue scales

In order to link the six questionnaires to one another, we used a single-group design, that is, all respondents came from the same population and answered all questionnaires at the same time of data collection [25]. To place the parameter estimates on a common scale, we estimated parameters for all items at once (concurrent calibration) [45]. Test scores were equated using expected a posteriori (EAP) estimates of the latent trait for each observable sum score. For each sum

score of the respective questionnaire, latent trait estimates were averaged over all possible response patterns to place the sum score on the latent trait continuum [46]. Using one common latent trait for calibration requires estimates from a model with acceptable fit, and that additionally meets the assumption of local independence, that is, item errors should vary independently.

We used confirmatory factor analysis with diagonally weighted least squares (DWLS) estimation and mean- and variance-adjusted test statistic to determine the underlying structure of the item pool. Model fit was assessed using a combinational rule of the comparative fit index (CFI) and the standardized root mean square residual (SRMR) [47]. Models were rejected if both CFI and SRMR indicated poor fit ( $CFI < 0.95$  and  $SRMR > 0.06$ ). Since these cutoffs are rules of thumb and were suggested for maximum likelihood (ML) estimates instead of DWLS, we additionally present the Tucker–Lewis Index (TLI), along with the root mean square error of approximation (RMSEA) and its 90% confidence interval (90% CI).

Initially, we fitted a *unidimensional model*, with all 50 items loading onto one single (general) latent factor. This model expresses the assumption of the multidimensional concept. The instruments differ in their underlying dimensional concepts of fatigue. Since, in the light of recent studies [11, 12], it seems possible that there is not only one *common* cause of fatigue. We additionally fitted a *bifactor model* that represents a limited version of the multidimensional concept insofar as it allows the common cause to be accompanied by additional differing causes. This is implemented through the modeling of three additional specific factors not correlated with the general factor to separate variance that cannot be attributed to the general factor: two factors to account for residual variance introduced by similar item content (content regarding *emotional fatigue* in FA12 and DIG11, and content regarding *cognitive fatigue* in FA12, FS11, and DIG11); and one factor for residual variance introduced by the *specific time frame* of the items in BFI9 (last 24 h) and their especially *wide response range* (from 0 to 10). The use of bifactor models is not uncommon for fatigue instruments that measure the content of different fatigue dimensions [29, 48].

Although multidimensional linking methods were proposed, especially for bifactor models [49, 50], multidimensional IRT models are seldom applied in practice [25]. We also did not apply a multidimensional model for calibration. Instead, we focused on the estimated parameters of the general factor. If the multidimensional solution was found to fit substantially better, our analytic plan was to fix the parameters of the specific factors at zero and keep the other parameters (those of the general factor) fixed at their estimated value for calibration. We used a unidimensional IRT model for item calibration; even though the estimates

of the general factor came from a multidimensional model. Therefore, we additionally examined the explained common variance (ECV) for the general factor to judge the degree of unidimensionality in the data [51], and we computed coefficient omega hierarchical (COH) to estimate the degree to which the total score reflects a common dimension, because COH is the appropriate model-based measure of reliability for data with a bifactor structure [52]. Values of  $COH > 0.80$  indicate that the items are sufficiently unidimensional [53]. To test the assumption of local independence, we examined the residual correlations between each pair of items. A value of 0.2 above the average correlation indicates local dependence [54]. These analyses were performed with R Statistics using the R package “lavaan” [55].

To establish the common metric, we chose the better fitting model (unidimensional or bifactor model) and estimated it within the Item Response Theory (IRT) framework [56] to predict the underlying latent trait using the specific item responses. Binary items followed a two-parameter model (2PL) whereas items with more than two response categories followed a graded response model (GRM) [57]. We used the item parameters of the general factor (common fatigue dimension) to compute theta values for each possible sum score from each questionnaire [58]. These analyses were done with R Statistics [59] using the R package “mirt” [60].

To judge the accuracy of the links between the six measurements, we examined Bland–Altman plots [61] for each pair of questionnaire. Based on the theta values estimated separately for each questionnaire, the figures show the difference between theta values of two measurements (Y axis) as a function of the mean value of theta values (X axis). While the mean is the best estimate of the true value [62], the difference indicates the bias between both methods for each estimated true value. Each figure contains a straight line that indicates the mean of the differences, and two dotted lines indicating the 90% limits of agreement (90%-LoA) [63]. The curved line shows a locally fitted nonparametric regression line [64], also called “loess” smoothing. With the help of this line, it is easy to judge visually, if the distortion between both methods crosses the line of the confidence interval, without focusing on individual data points.

### Linking the common metric of the patients to the general population

In the resulting IRT model, latent trait estimates (theta) of the calibration sample have a mean of  $M = 0$  and a standard deviation of  $SD = 1$  due to constraints necessary to fully identify the model. We anchored the common metric to the GP to facilitate interpretation. Therefore, we estimated T-scores ( $M = 50$ ,  $SD = 10$ ) for each sum score observed in the GP, assuming a normal distribution of the latent trait. We also calculated the theta values for each sum score of the

respective questionnaire in the patient sample. For each GP sample, the T-scores (GP) were regressed on the respective theta values (patients) to obtain a transformation formula for theta values into GP-based T-scores. For these analyses, we used Microsoft EXCEL 2010.

## Results

The questionnaire was completed by  $n = 1225$  patients. We estimated missing values for the 50 items from the instruments mentioned above. The proportion of missing values per item ranged from 0.2 to 1.2%. All patients answered more than half of the items. Of  $n = 1225$  patients,  $n = 1118$  (91.3%) answered all items completely. Altogether the proportion of missing values was low, i.e., 265 missing of 61,250 values (0.43%) were imputed, so that almost every technique for dealing with missing values would lead to similar results [43, 65].

### Socio-demographic and medical characteristics of the patient sample

Table 1 presents the socio-demographic and medical characteristics of the patient sample. On average, the patients were 55.8 years old (ranging from 18 to 88 years), 18.7% of them were between 18 and 40 years, and 45.9% were 60 years or older. 52.2% of the patients were women, and nearly a third (33.1%) had more than 10 years of education. Three-quarters of the patients (75.6%) answered the questionnaire within the first year of diagnosis. The three most common diagnoses were breast cancer (25.0%), prostate cancer (19.2%), and cancer of the gastrointestinal tract (18.1%).

### Common metric of the six fatigue scales

Table 2 presents the fit indices for the unidimensional model and the bifactor model. The unidimensional model that contains only the general factor for all 50 items did not show acceptable fit (CFI = 0.85, SRMR = 0.10). The bifactor model with the three additional specific factors showed acceptable fit (CFI = 0.94, SRMR = 0.06).

The average residual correlation was  $-0.01$ , ranging from  $-0.16$  to  $0.18$  except for two pairs of items: item 6 and item 7 of the FS11 ( $r = 0.30$ ) and item 5 and item 6 of the DIG11 ( $r = 0.36$ ) due to response-dependency. Signs of local dependence indicate inflated reliability and problems with construct validity [54]. To estimate the bias caused by local dependence, we additionally tested a bifactor model where these two pairs of items were parceled that is they were summed up, resulting in a model with 48 items instead of 50. We refrained from deleting either item because this

**Table 1** Socio-demographic and medical characteristics of the patient sample ( $n = 1225$ )

Socio-demographic characteristics	<i>n</i> (%)
Sample	Patients
	<i>n</i> = 1225
Gender	
Male	585 (47.8)
Female	640 (52.2)
Age	
18 to < 26 years	37 (3.0)
26 to < 40 years	192 (15.7)
40 to < 50 years	164 (13.4)
50 to < 60 years	270 (22.0)
60 to < 70 years	288 (23.5)
70 years and older	274 (22.4)
M (SD)	55.84 (15.99)
Education <sup>a</sup>	
Elementary school (8–9 years)	432 (35.4)
Junior high school (10 years)	348 (28.5)
High school/ university (> 10 years)	404 (33.1)
Other and no formal qualification	38 (3.1)
Employment status	
Employed	594 (48.5)
Unemployed	70 (5.7)
Retired	466 (38.0)
Other	95 (7.8)
Partnership <sup>a</sup> (yes)	911 (74.6)
Medical characteristics	
Time since diagnosis <sup>a</sup>	
Up to 6 months	566 (46.5)
>6 to 12 months	354 (29.1)
>12 months	297 (24.4)
Median (IQR)	7.00 (9.00)
Diagnosis	
Breast (C50)	306 (25.0)
Prostate (C61)	235 (19.2)
Gastrointestinal tract (C15–C25)	222 (18.1)
Hematological cancers (C81–C86, C91–C92)	135 (11.0)
Kidney/urinary tract (C64–C67)	112 (9.1)
Female genital organs (C51–C54, C56–C58)	68 (5.6)
Melanoma (C43–C44)	49 (4.0)
Male genital organs (C60, C62–C63)	39 (3.2)
Thyroid/endocrine glands (C73, C75)	18 (1.5)
Other	41 (3.3)

*M* mean, *SD* standard deviation, *IQR* interquartile range

<sup>a</sup>Missing values considered

might change content, convergent, and divergent validity of the instruments.

For the model with 50 items, the general factor explained more than 80% (ECV = 0.81) of the common variance that



**Table 2** Model fit ( $n = 1225$ )

Model	Chi <sup>2</sup>	df	Chi <sup>2</sup> /df	$p(\text{Chi}^2)$	SRMR	CFI	TLI	RMSEA (90%CI)
Unidimensional model	21569.5	1175	18.36	<0.001	0.098	0.852	0.845	0.119 (0.118–0.120)
Bifactor model	10036.3	1152	8.71	<0.001	0.059	0.935	0.931	0.079 (0.078–0.081)

$\text{Chi}^2$  scaled chi-squared statistic,  $df$  scaled degrees of freedom,  $p(\text{Chi}^2)$  scaled type-I-error-probability for  $\text{Chi}^2$ ,  $CFI$  scaled comparative fit index,  $TLI$  scaled Tucker–Lewis Index,  $SRMR$  Standardized root mean square residual,  $RMSEA$  scaled root mean square error of approximation,  $CI$  scaled RMSEA confidence interval

was explained by all four factors together. The reliability of the general factor was high ( $\text{COH}_{\text{general}} = 0.96$ ). The bias in reliability caused by local dependence was negligible:  $\text{COH}_{\text{general}}$  of the model with 48 items decreased by a value of less than 0.003 (ECV decreased by less than 0.004). Likewise, the model fit did not change substantially: CFI (scaled) decreased by a value of less than 0.005 and SRMR increased by less than 0.002. Because there was no substantial bias in these psychometric key features, we used the bifactor model with 50 items for calibration. An inspection of the item fit statistics of this model showed no significant misfit after correcting the significance level of 0.05 for multiple testing ( $\alpha_{\text{corrected}} = 0.001$ ). The lowest  $p$  value showed item 11 of the questionnaire FS11 with  $\text{Chi}^2(df) = 201.8(152)$ ,  $p$  value = 0.004. The IRT item parameters for this model are presented in Appendix 1.

For judging the accuracy of the links between the six measurements, Table 3 shows the mean of the differences in theta values between each pair along with the confidence intervals. The mean difference (bias) between two questionnaires ranges from 0.003 (DIG11 and FA3) to 0.037 (BFI9 and FA12). This means that the largest distortion between two linked questionnaires is less than 4% of standard deviation in theta value. This result is reflected in the Bland–Altman plots (see Appendix 3). No smoothed regression curve exceeds the limits of agreement for any pair. The less biased conversions are those where the straight line is near to zero and the curved line is nearly horizontal.

Figure 1 shows the standard error of measurement (measurement precision) depending on the theta value for each instrument within the common metric. In the theta range from minus one to one, all questionnaires achieved reliabilities above 0.8. The BFI9 was the most reliable instrument, achieving reliabilities above 0.90 in the broadest range of theta, i.e., from  $-2$  to  $3$ , essentially covering the full range of theta expected. Furthermore, one can see that the questionnaires with either few response options (DIG11) or few items (MFI4, FA3) are less reliable than the others.

Figure 2 visualizes the common metric of the six questionnaires, linked to the GP. The left axis shows the theta values (mean = 0, standard deviation = 1) of the patients. Each data point of the six questionnaires represents a sum score with a mean theta value (see also Appendix 2 with Supplementary Tables S1 to S6). For example, a score of 33 in the

**Table 3** Mean of differences in theta values between each pair of the questionnaire

Pair of questionnaires	MD (90%-LoA)
ALL-FA3	0.003 (−0.963–0.925)
ALL-FA12	−0.020 (−0.697–0.607)
ALL-MFI4	0.010 (−0.827–0.901)
ALL-BFI9	0.017 (−0.539–0.586)
ALL-FS11	−0.009 (−0.965–0.947)
ALL-DIG11	−0.001 (−0.973–0.866)
BFI9-FA3	−0.014 (−1.186–1.224)
BFI9-FA12	−0.037 (−1.134–1.026)
BFI9-MFI4	−0.007 (−1.201–1.089)
BFI9-FS11	−0.026 (−1.418–1.336)
BFI9-DIG11	−0.018 (−1.263–1.261)
DIG11-FA3	0.003 (−1.175–1.270)
DIG11-FA12	−0.020 (−1.041–0.907)
DIG11-MFI4	0.010 (−1.153–1.245)
DIG11-FS11	−0.008 (−1.100–1.301)
FA12-FA3	0.023 (−0.993–1.094)
FA12-MFI4	0.030 (−1.006–1.180)
FA12-FS11	0.012 (−1.183–1.242)
FA3-MFI4	0.007 (−1.041–1.075)
FA3-FS11	−0.011 (−1.336–1.386)
FS11-MFI4	0.019 (−1.227–1.302)

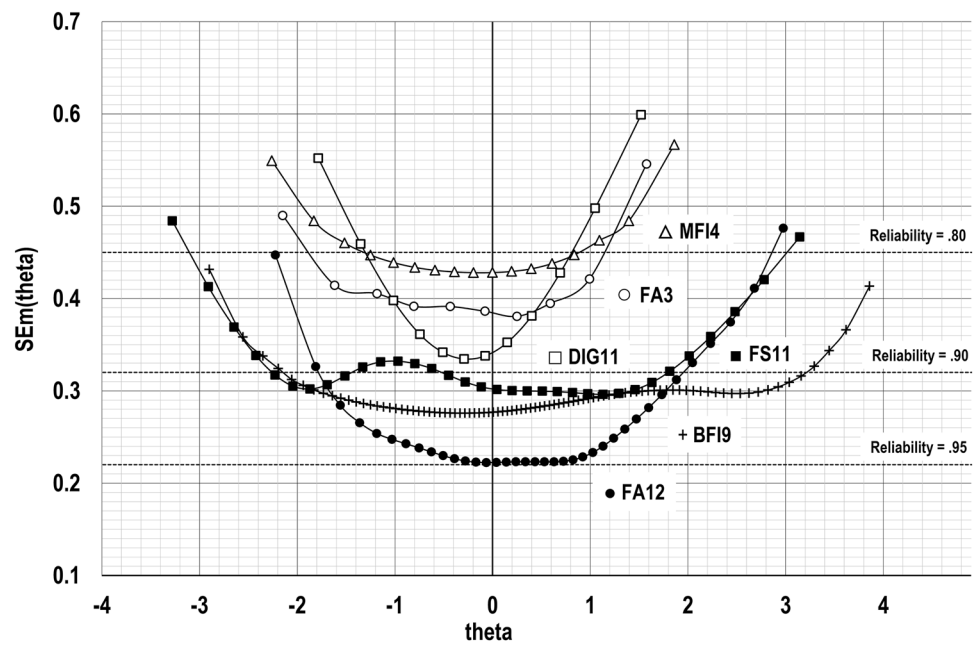
*MD* mean of the differences in theta values that were estimated separately for each questionnaire, *90%-LoA* limits of agreement, *ALL* theta values that were estimated from all 50 items, *FA3* three-item fatigue scale of the EORTC QLQ-C30, *FA12* twelve-item fatigue questionnaire EORTC QLQ-FA12, *MFI4* four-item General Fatigue scale of the MFI-20, *BFI9* nine-item fatigue questionnaire Brief Fatigue Inventory; *FS11* eleven-item fatigue questionnaire Fatigue Scale, *DIG11* eleven-item Diagnostic Interview Guide for Fatigue

three-item fatigue scale FA3 corresponds to a theta value of  $-0.81$ . An above-average burden of fatigue in patients ( $\text{theta} > 0$ ) corresponds to scores of  $\text{FA3} > 56$ ,  $\text{FA12} > 36$ ,  $\text{MFI4} > 13$ ,  $\text{BFI9} > 35$ ,  $\text{FS11} > 16$ , and  $\text{DIG11} > 6$ .

### Linking the common metric of the patients to the GP

The regression analyses for the T-scores of the GP (dependent variable) and the respective theta values of the patients (independent variable) led to the following transformation formula:

**Fig. 1** Measurement precision of the six fatigue scales. *SEm* standard error of measurement, *FA3* sum of the three-item fatigue scale of the EORTC QLQ-C30, *FA12* sum of the twelve items of the EORTC QLQ-FA12, *MFI4* sum of the four-item General Fatigue scale of the MFI-20, *BFI9* sum of the nine items of the BFI, *FS11* sum of the eleven items of the FS, *DIG11* sum of the eleven binary symptom items of the DIG for Fatigue



$$T\_score_{GP} = 61.34 + 8.95 \times \theta_{patients}$$

Based on this relation, we see that an average burden of fatigue in the patient sample ( $\theta(patients) = 0$ ) is more than one standard deviation above the average burden in the GP:  $T\ score(GP) = 61.3$ .

### Discussion

The central result of this study, the common metric of six different questionnaires measuring fatigue, is summarized in Fig. 2 and in Appendix 2, Supplementary Tables S1 to S6. We could show that in the theta range from plus/minus one standard deviation, i.e., for values of approximately 66% of the patients, all questionnaires achieved reliabilities above 0.8. For measuring in the medium range of theta, all questionnaires proved to be sufficiently reliable, whereas the BFI9 was the most reliable questionnaire in the whole theta range. Questionnaires with few items or few response options proved to be less reliable than others. Altogether, the reliability coefficients are sufficient to compare the questionnaires. While in the classical test theory (CTT), the standard error of measurement that is used to estimate the reliability, is a constant, in the IRT it is a variable depending on the test information function. Hence the expected score fluctuations due to error, which are critical to interpretations of individual scores, can vary [66]. This is one of the advantages of IRT over CTT. The theta values appear to be reasonable since they range between  $-3.28$  (FS11) and  $3.86$  (BFI9) that is, no

points deviate more than four standard deviations from the mean. Likewise, the mean ( $\theta = 0$ ) corresponds to scores that are located in the medium score range of the instruments ( $FA3_{0-100} = 58.6$ ,  $FA12_{0-100} = 37.9$ ,  $MFI4_{420} = 13.0$ ,  $BFI9_{0-10} = 4.0$ ,  $FS11_{0-33} = 16.7$ , and  $DIG11_{0-11} = 6.3$ ).

The data represented in Appendix 2, Supplementary Tables S1 to S6, indicate the position of the scores that actually occur when an individual completes a questionnaire. However, mean scores of groups of individuals generally differ from the points directly indicated in the figure or the supplementary tables. The FA3 mean score of 30.5 obtained in a study with 221 breast cancer patients [67] is between the scores of 22 and 33 which correspond to theta scores between  $-1.18$  and  $-0.81$  according to Fig. 2 or Supplementary Table S1. In such cases, we recommend a linear interpolation in the following way:

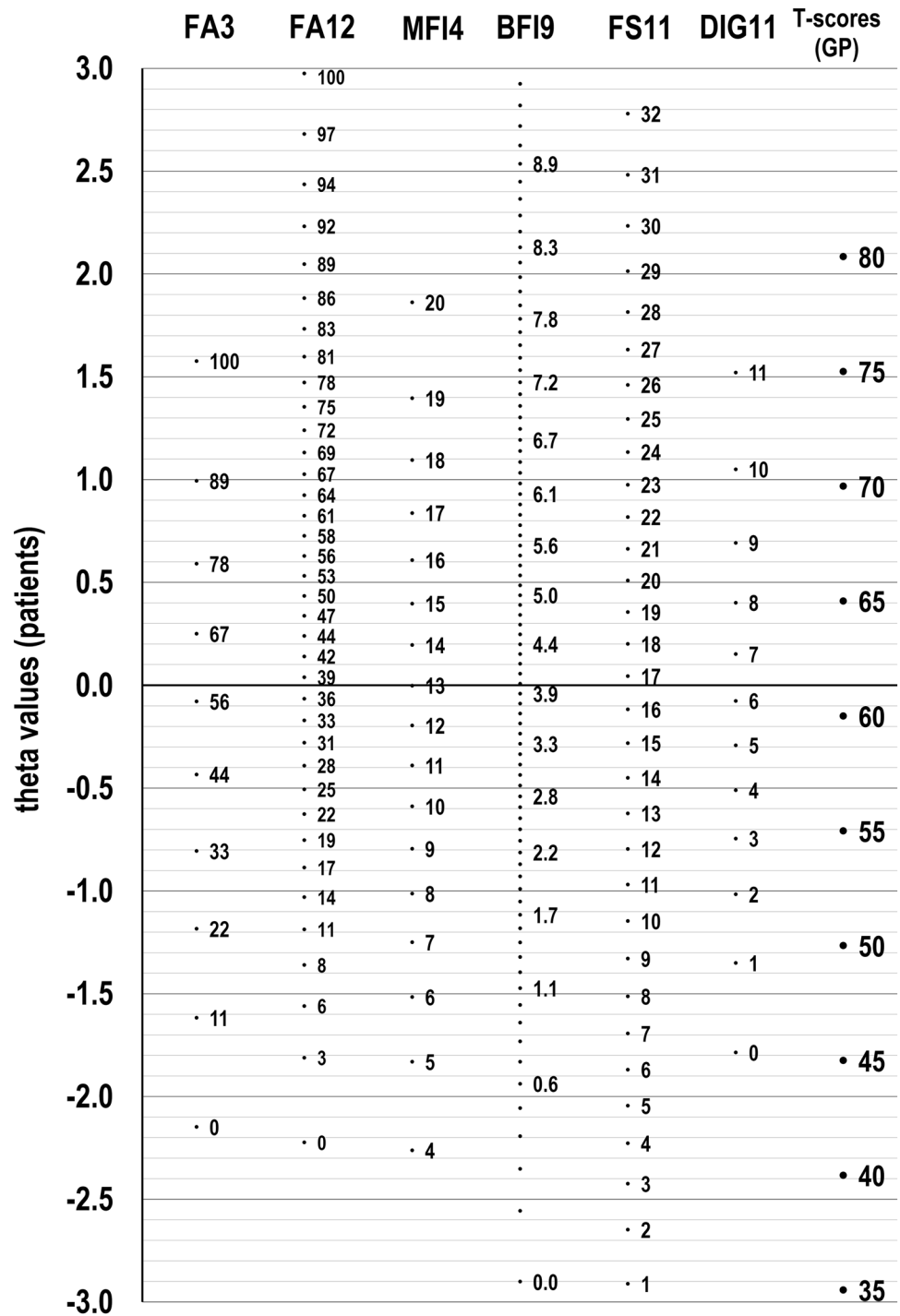
$$\begin{aligned} \theta(30.5) &= -1.18 + \frac{30.5 - 22}{33 - 22} \cdot (-0.81 - (-1.18)) \\ &= -1.18 + 0.77 \cdot 0.37 = -0.89. \end{aligned}$$

This interpolation can also be performed visually with a sufficient degree of accuracy using Fig. 2.

Another way to illustrate the value of this metric is the comparison of results from studies that used different questionnaires. For example, three studies with breast cancer patients [68–70] used three different fatigue questionnaires with the mean scores of BFI9 = 3.04 [68], FA3 = 36.5 [69], and MFI4 = 11.3 [70]. Using Fig. 2, the similarity of the findings can be assessed.

Furthermore, the common metric also allows extrapolating cutoffs for different questionnaires. Concerning a score in FA3 greater than 39 (proposed by Giesinger et al. [71]

**Fig. 2** Common metric of the six fatigue scales. *FA3* sum of the three-item fatigue scale of the EORTC QLQ-C30 (item range = 1–4, scale range = 0–100), *FA12* sum of the twelve items of the EORTC QLQ-FA12 (item range = 1–4, scale range = 0–100), *MFI4* sum of the four-item General Fatigue scale of the MFI-20 (item range = 1–5, scale range = 4–20), *BFI9* sum of the nine items of the BFI (item range = 0–10, scale range = 0–10), *FS11* sum of the eleven items of the FS (item range = 0–3, scale range = 0–33), *DIG11* sum of the eleven binary symptom items of the DIG for Fatigue (item range: 0/1, scale range = 0–11), *T-scores(GP)* estimation of T-scores for the German general population (mean = 50, standard deviation = 10)



as a threshold for clinical importance), we infer that this value corresponds to a value of theta(patients) ≈ -0.60:

$$\begin{aligned}
 \text{theta}(39) &= -0.81 + \frac{39 - 33}{44 - 33} \cdot (-0.43 - (-0.81)) \\
 &= -0.81 + 0.55 \cdot 0.37 = -0.60.
 \end{aligned}$$

Hence the FA3 threshold corresponds to scores of FA12 > 22, MFI4 > 9, BFI9 > 2.6, FS11 > 13, and DIG11 > 3.

Finally, we linked the theta values of the patient sample to T-scores of the GP. We presented a formula which can be used to estimate the burden of fatigue in the GP that corresponds to the burden in a comparable patient sample. This allows a straightforward comparison to a meaningful reference.



## Limitations

Assuming the multidimensional concept, we modeled a unidimensional and a bifactor model. Only with the bifactor model, which represents a weaker version of this concept, we did find the data to be reasonably unidimensional despite the differences in questionnaire design and background. However, the share of variance attributed to the specific factors was small compared to the general factor. On the basis of this general factor, we developed a common metric. We anchored the common metric to the GP on the basis of four large samples assuming a linear relationship between the burden of fatigue in the patient population and in the GP. This assumption was supported by highly explained variances of the four regression models ranging from 97.2 to 99.7%. Although estimating item parameters in the general population samples and subsequent linking would have been a viable way to scale the model, we chose the regression approach, because we expected that in the GP higher levels of fatigue would be underrepresented or even absent, making item parameter estimation unreliable. For example, the interquartile range of FA12 differed between patients and GP substantially: for patients from 22.2 to 58.3, versus for the GP from 0 to 27.8. To avoid presenting four different continuums of T-values for the GP, we averaged the regression constants and coefficients to obtain one transformation formula. The advantage of this approach is that, although it does not reflect the values of the general population in terms of a single questionnaire with the same accuracy, it is easier to apply because it does not require differentiation between different questionnaires in the GP.

A particular strength of this study is that we included six different instruments, but the plethora of fatigue items in the questionnaire booklet might also have tired the patients. To tone down this effect, we put questionnaires concerning other topics (quality of life, optimism, self-efficacy, coping, and others) between the fatigue questionnaires.

Although the selection of questionnaires might seem rather arbitrary, we selected fatigue instruments that are frequently used in samples of cancer patients. For the Fatigue Diagnostic Interview Guide (DIG), a detailed psychometric examination of a total scale is still pending; nevertheless, we included this scale in our analysis. From the Multidimensional Fatigue Inventory (MFI-20), we chose one subscale with four items in respect to the length of the questionnaire for the patients. This subscale, called General Fatigue, is assumed to assess general aspects of fatigue. For the EORTC QLQ-FA12, we used an overall score that includes all 12 items without regard to the respective sub-dimension. This is a consequence of assuming the multidimensional concept, and it was justified by the results of the bifactor model. It is beyond the

scope of this study to discuss which conceptualization of fatigue is appropriate; hence, we rather state possible critical decisions.

Another concern that is not only connected to the issue of dimensionality is that we did not remove any item from the calibration. The main reason for this is that a linking using all items does reflect the relationships between the questionnaires more realistically than a linking that uses only the matching items: a scoring in which individual items may not be used loses its practicability. Furthermore, inclusion of all items of each questionnaire allows transforming sum scores on the common metric even if raw item response data are not available. This is particularly evident if the data are to be compared with studies in which the scores are reported that are comprised of all items of the respective questionnaire. Given these arguments, we believe our approach provides an optimal trade-off between practicability and conceptual or theoretical grounds.

A further strength of our study is an analysis based on a large sample of cancer patients. Even in the range of severe fatigue, the sample size is sufficient for profound estimates. Existing common metrics for fatigue questionnaires [20–22] were based on samples from the general population only. They do not allow comparing the results from a sample of cancer patients against the general population. Furthermore, they did not consider more than three questionnaires. For example, Lai et al. [22] included three questionnaires: Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-Fatigue), Medical Outcomes Study Short Form (SF-36) Vitality Scale, and Quality of Life in Neurological Disorders (Neuro-QOL).

A limitation of the study is that we did not validate the common metric in an independent sample of patients. It is possible that item parameters differ across samples. Therefore, the generalizability of our results to other samples with different characteristics is unclear. This is true in particular for samples differing in age, gender, and time since diagnosis, because these characteristics are associated with fatigue [67, 72, 73]. Limited generalizability may also be the case when comparing samples of patients without cancer or with other diseases that show fatigue, but Cella et al., 2016 [1] presented support for the assumption that measures of the Patient-Reported Outcomes Measurement Information System (PROMIS) fatigue item bank may be validly compared across different chronic conditions. Furthermore, the comparison, or more precisely, the conversion of fatigue questionnaires on the basis of our common metric is reasonable only if the multidimensional concept underlies the interpretation.

## Conclusions

Our study enables researchers and clinicians to directly compare the results between multiple different studies using different fatigue instruments, and it enables the simultaneous comparison of the results with the metric derived from the general population.

**Acknowledgements** We thank all patients who participated in this study, and the German Cancer Aid for funding the study.

**Funding** This study was supported by the German Cancer Aid (Grant Number: 7011 2267).

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** The study received research ethics committee approval (File Number: 412/14-ff) from ethic board of the medical faculty of the University of Leipzig.

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

## References

1. Cella, D., Lai, J.-S., Jensen, S. E., Christodoulou, C., Junghaenel, D. U., Reeve, B. B., et al. (2016). PROMIS fatigue item bank had clinical validity across diverse chronic conditions. *Journal of Clinical Epidemiology*, *73*, 128–134. <https://doi.org/10.1016/j.jclinepi.2015.08.037>.
2. Finsterer, J., & Mahjoub, S. Z. (2014). Fatigue in healthy and diseased individuals. *The American journal of hospice & palliative care*, *31*(5), 562–575. <https://doi.org/10.1177/1049909113494748>.
3. Minton, O., Berger, A., Barsevick, A., Cramp, F., Goedendorp, M., Mitchell, S. A., et al. (2013). Cancer-related fatigue and its impact on functioning. *Cancer*, *119*(Suppl 11), 2124–2130. <https://doi.org/10.1002/cncr.28058>.
4. Richardson, A. (1995). Fatigue in cancer patients: A review of the literature. *European Journal of Cancer Care*, *4*(1), 20–32. <https://doi.org/10.1111/j.1365-2354.1995.tb00049.x>.
5. Kobashi-Schoot, J. A. M., Hanewald, G. J., van Dam, F. S., & Bruning, P. F. (1985). Assessment of malaise in cancer patients treated with radiotherapy. *Cancer Nursing*, *8*(6), 306–313. <https://doi.org/10.1097/00002820-198512000-00003>.
6. Bower, J. E. (2014). Cancer-related fatigue: Mechanisms, risk factors, and treatments. *Nature Reviews Clinical Oncology*, *11*(10), 597–609. <https://doi.org/10.1038/nrclinonc.2014.127>.
7. Cella, D., Peterman, A., Passik, S., Jacobsen, P., & Breitbart, W. (1998). Progress toward guidelines for the management of fatigue. *Oncology*, *12*(11A), 369–377.
8. Minton, O., & Stone, P. (2009). A systematic review of the scales used for the measurement of cancer-related fatigue (CRF). *Annals of Oncology*, *20*(1), 17–25. <https://doi.org/10.1093/annonc/mdn537>.
9. Wang, X. S., & Woodruff, J. F. (2015). Cancer-related and treatment-related fatigue. *Gynecologic Oncology*, *136*(3), 446–452. <https://doi.org/10.1016/j.ygyno.2014.10.013>.
10. Radbruch, L., Strasser, F., Elsner, F., Gonçalves, J. F., Løge, J., Kaasa, S., et al. (2008). Fatigue in palliative care patients: An EAPC approach. *Palliative Medicine*, *22*(1), 13–32. <https://doi.org/10.1177/0269216307085183>.
11. de Raaf, P. J., de Klerk, C., & van der Rijt, C. C. D. (2013). Elucidating the behavior of physical fatigue and mental fatigue in cancer patients: a review of the literature. *Psycho-Oncology*, *22*(9), 1919–1929. <https://doi.org/10.1002/pon.3225>.
12. Friedrich, M., Nowe, E., Hofmeister, D., Kuhn, S., Leuteritz, K., Sender, A., et al. (2018). Psychometric properties of the fatigue questionnaire EORTC QLQ-FA12 and proposal of a cut-off value for young adults with cancer. *Health and Quality of Life Outcomes*, *16*(1), 125. <https://doi.org/10.1186/s12955-018-0949-0>.
13. Seyidova-Khoshknabi, D., Davis, M. P., & Walsh, D. (2011). Review article: A systematic review of cancer-related fatigue measurement questionnaires. *The American Journal of Hospice & Palliative Care*, *28*(2), 119–129. <https://doi.org/10.1177/1049909110381590>.
14. Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders*, *28*(1), 88–96. <https://doi.org/10.1016/j.janxdis.2013.11.006>.
15. Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: Linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment*, *26*(2), 513–527. <https://doi.org/10.1037/a0035768>.
16. Kaat, A. J., Newcomb, M. E., Ryan, D. T., & Mustanski, B. (2017). Expanding a common metric for depression reporting: Linking two scales to PROMIS® depression. *Quality of Life Research*, *26*(5), 1119–1128. <https://doi.org/10.1007/s11136-016-1450-z>.
17. Schalet, B. D., Revicki, D. A., Cook, K. F., Krishnan, E., Fries, J. F., & Cella, D. (2015). Establishing a common metric for physical function: linking the HAQ-DI and SF-36 PF subscale to PROMIS® physical function. *Journal of General Internal Medicine*, *30*(10), 1517–1523. <https://doi.org/10.1007/s11606-015-3360-0>.
18. Kaat, A. J., Schalet, B. D., Rutsohn, J., Jensen, R. E., & Cella, D. (2018). Physical function metric over measure: An illustration with the Patient-Reported Outcomes Measurement Information System (PROMIS) and the Functional Assessment of Cancer Therapy (FACT). *Cancer*, *124*(1), 153–160. <https://doi.org/10.1002/cncr.30981>.
19. ten Klooster, P. M., Voshaar, O., Gandek, M. A. H., Rose, B., Bjorner, M., Taal, J. B., E., et al (2013). Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire disability index in rheumatoid arthritis. *Health and Quality of Life Outcomes*, *11*, 199. <https://doi.org/10.1186/1477-7525-11-199>.
20. Lai, J.-S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, *92*(10 Suppl), S20–S27. <https://doi.org/10.1016/j.apmr.2010.08.033>.
21. Noonan, V. K., Cook, K. F., Bamer, A. M., Choi, S. W., Kim, J., & Amtmann, D. (2012). Measuring fatigue in persons with multiple sclerosis: Creating a crosswalk between the Modified Fatigue Impact Scale and the PROMIS Fatigue Short Form. *Quality of Life Research*, *21*(7), 1123–1133. <https://doi.org/10.1007/s11136-011-0040-3>.

22. Lai, J.-S., Cella, D., Yanez, B., & Stone, A. (2014). Linking fatigue measures on a common reporting metric. *Journal of Pain and Symptom Management*, 48(4), 639–648. <https://doi.org/10.1016/j.jpainsymman.2013.12.236>.
23. Choi, S. W., Podrabsky, T., McKinney, N., Schalet, B. D., Cook, K. F., & Cella, D. (Eds.). (2012). *PROSetta Stone® analysis report: a rosetta stone for patient reported outcomes (Vol. 1)*. Chicago, IL.
24. Brossman, B. G., & Lee, W.-C. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement*, 37(6), 460–481. <https://doi.org/10.1177/0146621613484083>.
25. Lee, W.-C., & Lee, G. (2018). IRT linking and equating. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The wiley handbook of psychometric testing* (pp. 639–673). Chichester: Wiley.
26. Sunderland, M., Batterham, P., CEAR, A., & Carragher, N. (2018). Validity of the PROMIS depression and anxiety common metrics in an online sample of Australian adults. *Quality of Life Research*. <https://doi.org/10.1007/s11136-018-1905-5>.
27. Kim, J., Chung, H., Askew, R. L., Park, R., Jones, S. M. W., Cook, K. F., et al. (2017). Translating CESD-20 and PHQ-9 Scores to PROMIS Depression. *Assessment*, 24(3), 300–307. <https://doi.org/10.1177/1073191115607042>.
28. Lai, J.-S., Crane, P. K., & Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Quality of Life Research*, 15(7), 1179–1190. <https://doi.org/10.1007/s11136-006-0060-6>.
29. Cella, D., Lai, J.-S., & Stone, A. (2011). Self-reported fatigue: One dimension or more? Lessons from the Functional Assessment of Chronic Illness Therapy—Fatigue (FACIT-F) questionnaire. *Supportive Care in Cancer*, 19(9), 1441–1450. <https://doi.org/10.1007/s00520-010-0971-1>.
30. 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*. 1993(85), 365–376.
31. Weis, J., Tomaszewski, K. A., Hammerlid, E., Arraras, J. I., Conroy, T., Lanceley, A., et al. (2017). International psychometric validation of an EORTC quality of life module measuring cancer related fatigue (EORTC QLQ-FA12). *Journal of the National Cancer Institute*, 109(5), 1–8. <https://doi.org/10.1093/jnci/djw273>.
32. Kecke, S., Ernst, J., Einkenkel, J., Singer, S., & Hinz, A. (2017). Psychometric properties of the fatigue questionnaire EORTC QLQ-FA12 in a sample of female cancer patients. *Journal of Pain and Symptom Management*, 54(6), 922–928. <https://doi.org/10.1016/j.jpainsymman.2017.08.007>.
33. Hinz, A., Weis, J., Brähler, E., & Mehnert, A. (2018). Fatigue in the general population: German normative values of the EORTC QLQ-FA12. *Quality of Life Research*, 63–69. <https://doi.org/10.1007/s11136-018-1918-0>.
34. Smets, E. M. A., Garssen, B., Bonke, B., & de Haes, J. C. J. M. (1995). The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research*, 39(3), 315–325. [https://doi.org/10.1016/0022-3999\(94\)00125-0](https://doi.org/10.1016/0022-3999(94)00125-0).
35. Radbruch, L., Sabatowski, R., Elsner, F., Everts, J., Mendoza, T., & Cleeland, C. (2003). Validation of the German version of the Brief Fatigue Inventory. *Journal of pain and symptom management*, 25(5), 449–458. [https://doi.org/10.1016/S0885-3924\(03\)00073-3](https://doi.org/10.1016/S0885-3924(03)00073-3).
36. Martin, A., Staufenbiel, T., Gaab, J., Rief, W., & Brähler, E. (2010). Messung chronischer Erschöpfung—Teststatistische Prüfung der Fatigue Skala (FS). *Zeitschrift für Klinische Psychologie und Psychotherapie*, 39(1), 33–44. <https://doi.org/10.1026/1616-3443/a000010>.
37. Reinertsen, K. V., Cvancarova, M., Loge, J. H., Edvardsen, H., Wist, E., & Fosså, S. D. (2010). Predictors and course of chronic fatigue in long-term breast cancer survivors. *Journal of Cancer Survivorship: Research and Practice*, 4(4), 405–414. <https://doi.org/10.1007/s11764-010-0145-7>.
38. Chalder, T., Berelowitz, G., Pawlikowska, T., Watts, L., Wesely, S., Wright, D., et al. (1993). Development of a fatigue scale. *Journal of Psychosomatic Research*, 37(2), 147–153. [https://doi.org/10.1016/0022-3999\(93\)90081-P](https://doi.org/10.1016/0022-3999(93)90081-P).
39. Morriss, R., Wearden, A., & Mullis, R. (1998). Exploring the validity of the Chalder fatigue scale in chronic fatigue syndrome. *Journal of Psychosomatic Research*, 45(5), 411–417. [https://doi.org/10.1016/S0022-3999\(98\)00022-1](https://doi.org/10.1016/S0022-3999(98)00022-1).
40. Schwarz, R., & Hinz, A. (2001). Reference data for the quality of life questionnaire EORTC QLQ-C30 in the general German population. *European Journal of Cancer*, 37(11), 1345–1351. [https://doi.org/10.1016/S0959-8049\(00\)00447-0](https://doi.org/10.1016/S0959-8049(00)00447-0).
41. Hinz, A., Singer, S., & Brähler, E. (2014). European reference values for the quality of life questionnaire EORTC QLQ-C30: Results of a German investigation and a summarizing analysis of six European general population normative studies. *Acta Oncologica*, 53(7), 958–965. <https://doi.org/10.3109/0284186X.2013.879998>.
42. Schwarz, R., Krauss, O., & Hinz, A. (2003). Fatigue in the general population. *Oncology Research and Treatment*, 26(2), 140–144. <https://doi.org/10.1159/000069834>.
43. Kalkan, ÖK., Kara, Y., & Kelecioğlu, H. (2018). Evaluating performance of missing data imputation methods in IRT analyses. *International Journal of Assessment Tools in Education*, 403–416. <https://doi.org/10.21449/ijate.430720>.
44. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1), 1–38.
45. Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for Item Response Theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3–24. <https://doi.org/10.1177/0146621602026001001>.
46. Fischer, H. F., & Rose, M. (2018). Scoring depression on a common metric: a comparison of EAP estimation, plausible value imputation, and full bayesian IRT modeling. *Multivariate Behavioral Research*, 1–15. <https://doi.org/10.1080/00273171.2018.1491381>.
47. Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>.
48. Amtmann, D., Bamer, A. M., Noonan, V., Lang, N., Kim, J., & Cook, K. F. (2012). Comparison of the psychometric properties of two fatigue scales in multiple sclerosis. *Rehabilitation Psychology*, 57(2), 159–166. <https://doi.org/10.1037/a0027890>.
49. Lee, G., & Lee, W.-C. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education*, 29(3), 224–241. <https://doi.org/10.1080/08957347.2016.1171770>.
50. Kim, K. Y. (2017). *IRT linking methods for the bifactor model: a special case of the two-tier item factor analysis model*. Dissertation. University of Iowa, Iowa. <https://ir.uiowa.edu/etd/5792/>. Accessed 10 Dec 2018.
51. Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73(1), 5–26. <https://doi.org/10.1177/0013164412449831>.

52. Reise, S. P. (2012). Invited Paper: The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667–696. <https://doi.org/10.1080/00273171.2012.715555>.
53. Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, 98(3), 223–237. <https://doi.org/10.1080/00223891.2015.1089249>.
54. Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the rasch model using residual correlations. *Applied Psychological Measurement*, 41(3), 178–194. <https://doi.org/10.1177/0146621616677520>.
55. Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2). <https://doi.org/10.18637/jss.v048.i02>.
56. Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists (Multivariate applications series)*. Hoboken: Taylor and Francis.
57. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1), 1–97. <https://doi.org/10.1007/BF03372160>.
58. Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (2016). Item Response Theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, 19(1), 39–49. <https://doi.org/10.1177/014662169501900105>.
59. R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
60. Chalmers, R. P. (2012). mirt: A multidimensional item response theory Package for the R environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>.
61. Bland, J. M., & Altman, D. G. (1995). Comparing methods of measurement: Why plotting difference against standard method is misleading. *The Lancet*, 346(8982), 1085–1087. [https://doi.org/10.1016/S0140-6736\(95\)91748-9](https://doi.org/10.1016/S0140-6736(95)91748-9).
62. Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307–310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8).
63. Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, 25(2), 141–151. <https://doi.org/10.11613/BM.2015.015>.
64. Cleveland, W. S., Devlin, S. J., & Grosse, E. (1988). Regression by local fitting. *Journal of Econometrics*, 37(1), 87–114. [https://doi.org/10.1016/0304-4076\(88\)90077-2](https://doi.org/10.1016/0304-4076(88)90077-2).
65. Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th edn.). Boston: Pearson Education.
66. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists (Multivariate applications book series)*. Mahwah: L. Erlbaum Associates.
67. Reidunsdatter, R. J., Albrektsen, G., Hjermstad, M. J., Rannestad, T., Oldervoll, L. M., & Lundgren, S. (2013). One-year course of fatigue after post-operative radiotherapy in Norwegian breast cancer patients—comparison to general population. *Acta Oncologica*, 52(2), 239–248. <https://doi.org/10.3109/0284186X.2012.742563>.
68. So, W. K. W., Marsh, G., Ling, W. M., Leung, F. Y., Lo, J. C. K., Yeung, M., et al. (2009). The symptom cluster of fatigue, pain, anxiety, and depression and the effect on the quality of life of women receiving treatment for breast cancer: A multicenter study. *Oncology Nursing Forum*, 36(4), E205–E214. <https://doi.org/10.1188/09.ONF.E205-E214>.
69. Versmessen, H., Vinh-Hung, V., van Parijs, H., Miedema, G., Voordeckers, M., Adriaenssens, N., et al. (2012). Health-related quality of life in survivors of stage I-II breast cancer: Randomized trial of post-operative conventional radiotherapy and hypofractionated tomotherapy. *BMC Cancer*, 12, 495. <https://doi.org/10.1186/1471-2407-12-495>.
70. Rissanen, R., Arving, C., Ahlgren, J., & Nordin, K. (2014). Group versus individual stress management intervention in breast cancer patients for fatigue and emotional reactivity: A randomised intervention study. *Acta Oncologica*, 53(9), 1221–1229. <https://doi.org/10.3109/0284186X.2014.923935>.
71. Giesinger, J. M., Kuijpers, W., Young, T., Tomaszewski, K. A., Friend, E., Zabernigg, A., et al. (2016). Thresholds for clinical importance for four key domains of the EORTC QLQ-C30: Physical functioning, emotional functioning, fatigue and pain. *Health and Quality of Life Outcomes*, 14, 87. <https://doi.org/10.1186/s12955-016-0489-4>.
72. Butt, Z., Rao, A. V., Lai, J.-S., Abernethy, A. P., Rosenbloom, S. K., & Cella, D. (2010). Age-associated differences in fatigue among patients with cancer. *Journal of Pain and Symptom Management*, 40(2), 217–223. <https://doi.org/10.1016/j.jpainsymman.2009.12.016>.
73. Akechi, T., Kugaya, A., Okamura, H., Yamawaki, S., & Uchitomi, Y. (1999). Fatigue and its associated factors in ambulatory cancer patients. *Journal of Pain and Symptom Management*, 17(1), 42–48. [https://doi.org/10.1016/S0885-3924\(98\)00105-5](https://doi.org/10.1016/S0885-3924(98)00105-5).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.