Qual Life Res (2015) 24:2239–2255
DOI 10.1007/s11136-015-0963-1

Systematic content evaluation and review of measurement properties of questionnaires for measuring self-reported fatigue among older people

Thorlene Egerton[1] · Ingrid I. Riphagen[2] · Arnhild J. Nygård[1] · Pernille Thingstad[1] · Jorunn L. Helbostad[1,3]

Accepted: 11 March 2015 / Published online: 17 March 2015

Abstract
Purpose The assessment of fatigue in older people requires simple and user-friendly questionnaires that capture the phenomenon, yet are free from items indistinguishable from other disorders and experiences. This study aimed to evaluate the content, and systematically review and rate the measurement properties of self-report questionnaires for measuring fatigue, in order to identify the most suitable questionnaires for older people.
Methods This study firstly involved identification of questionnaires that purport to measure self-reported fatigue, and evaluation of the content using a rating scale developed for the purpose from contemporary understanding of the construct. Secondly, for the questionnaires that had acceptable content, we identified studies reporting measurement properties and rated the methodological quality of those studies according to the COSMIN system. Finally, we extracted and synthesised the results of the studies to give an overall rating for each questionnaire for each measurement property. The protocol was registered with PROSPERO (CRD42013005589).
Results Of the 77 identified questionnaires, twelve were selected for review after content evaluation. Methodological quality varied, and there was a lack of information on measurement error and responsiveness.
Conclusions The PROMIS-Fatigue item bank and short forms perform the best. The FACIT-Fatigue scale, Parkinsons Fatigue Scale, Perform Questionnaire, and Uni-dimensional Fatigue Impact Scale also perform well and can be recommended. Minor modifications to improve performance are suggested. Further evaluation of unresolved measurement properties, particularly with samples including older people, is needed for all the recommended questionnaires.

Keywords Review · Patient-reported outcome · Fatigue · Measurement properties · Older people

Introduction

Fatigue is one of the most common complaints among community-dwelling older people. It is often unexplained [1] and is frequently blamed for disability [2, 3]. Fatigue is a feature of many illnesses and thus commonly presents among those with co-morbidity. Many experts now believe the nature and experience of fatigue are not disease specific [4] and can be considered a single construct regardless of cause [5, 6]. Along these lines, the Patient-Reported Outcomes Measurement Information System (PROMIS) group carried out extensive work to develop a definition. They described fatigue as ranging from 'mild subjective feelings of tiredness to an overwhelming, debilitating, and sustained sense of exhaustion that is likely to decrease one's ability to carry out daily activities, including the ability to work effectively and to function at one's usual level in family or social roles' [7].

Electronic supplementary material The online version of this article (doi:10.1007/s11136-015-0963-1) contains supplementary material, which is available to authorized users.

✉ Thorlene Egerton
thor@sutmap.com

[1] Department of Neuroscience, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[2] Unit for Applied Clinical Research, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[3] Department of Clinical Services, St. Olavs Hospital, Trondheim, Norway

Research on causes, risk factors and treatments for fatigue requires measurement instruments that adequately capture the problem yet have minimal contamination caused by capturing concurrent problems. There is currently no clear gold standard for measuring fatigue experienced by older people, and no scale designed specifically for them. Measurement of fatigue in older people is particularly challenging because it is unlikely to occur in isolation, and can be complicated by association with other symptoms such as pain, sleepiness, depression, physical unfitness and/or other aspects of any disability. Many questionnaires have been developed to measure fatigue due to the wide range of conceptualisations of the problem and the concurrent development of questionnaires for many specific diseases. It is not clear which questionnaire is best for a general older population.

The aim of this study was to systematically review the measurement properties of published fatigue questionnaires and generate recommendations for the measurement of fatigue among older people. With the evolving of the concept of fatigue, many questionnaires can now no longer necessarily be considered valid. The degree to which a questionnaire measures important features of the construct (*comprehensiveness*), and includes items that measure the construct without inadvertently measuring other constructs (*relevance and specificity*), is often overlooked during both questionnaire development and in later psychometric studies and reviews. Yet without a strong case for this content validity, the questionnaire may be yielding misleading data [8], despite good performance on quantitative tests. In such a review, there is clearly a need to evaluate the extent to which each potential questionnaire measures the construct of fatigue as we currently understand it, as well as considering the evidence for other aspects of psychometric performance. In this review, we focus on fatigue as a uni-dimensional construct despite many questionnaires including a range of issues including the physical, cognitive, and emotional experiences of fatigue and its impact upon physical, mental, and social activity [9].

Methods

This is a systematic review and thus does not constitute a clinical study or contain patient data. The review protocol was registered with PROSPERO (CRD42013005589).

Stage 1: Identification of self-report fatigue questionnaires and content assessment

Search for questionnaires

PubMed and PsycINFO were searched until January 2015 to identify questionnaires that measure self-reported fatigue. The key concepts that were combined are summarised as (1) fatigue, (2) measurement instrument, and (3) measurement property. Our population of interest, older people, was not represented in our search. Details are provided in Online Resource 1(a). We also checked reference lists of previous reviews [10–14].

Questionnaire selection

Two reviewers independently screened all titles and abstracts for questionnaires. Selection criteria were:

1. The questionnaire was intended to measure self-reported fatigue (severity, nature and/or impact).
2. The questionnaire should be able to differentiate between small differences in levels of fatigue with a sufficient range of possible scores (evaluative rather than discriminative or diagnostic).
3. At least one study that evaluated one or more measurement properties was found.
4. The full questionnaire was available in English.
5. Fatigue subscales of larger questionnaires that had been validated to stand alone were eligible for selection.
6. Where there was more than one version of the questionnaire, for example a long and a short form, both were eligible for selection.

Reasons for exclusion were:

1. The purpose of the questionnaire was to measure sleepiness, fatigue experienced by children, adolescents, pregnant or post-partum women, fatigue resulting from work, occupation, or a specific activity such as driving, or fatigue due to participation in sport or exercise.
2. Single-item questionnaires.
3. Questionnaire designed to be completed by a clinician after observation or questioning (not self-report).
4. Not designed to generate a total score or consisting of open-ended questions.
5. Where a revision of a questionnaire was created to replace an original version, only the revision was eligible.

Development of content rating scale and rating of questionnaires

The content evaluation was based on the PROMIS definition of fatigue [7, 15]. A checklist was developed that operationalised this definition, was consistent with contemporary understanding of the construct, and guided by the COSMIN definition of content validity [16–18]. The checklist enabled us to semi-quantifiably determine the

extent to which the items in each questionnaire were suitable for the purpose of measuring fatigue among older people. The resulting scale (Online Resource 2) involved rating the following aspects:

1. *Comprehensiveness*. We rated the extent to which the questionnaire captured the key features of fatigue.
2. *Relevance and specificity* to the construct. We determined the proportion of items in the questionnaire that measured the right sort of fatigue without overlap with other problems such as mood, cognition, or physical impairments, symptoms that are side effects of medication, sleepiness, or 'normal' fatigue experiences (resulting directly from a bout of physical or mental exertion).
3. *Item-level validity*. We considered the suitability of content and language of each item for older people of either gender and in any setting.
4. *Scoring system*. We also considered suitability of the scoring system for older people because of their known difficulty with visual analogue scales and scales with many response options [19].

The scale generated an overall rating for the content, which could be 'inadequate', 'adequate', 'good', or 'excellent' (Online Resource 2). It was piloted by three authors, and several revisions were made to improve clarity and usability. As the content rating scale itself has not been assessed for reliability, each identified questionnaire was rated by three raters, all physiotherapists with experience in rehabilitation settings with older patients. Differences in scores were discussed until agreement was achieved. Questionnaires scoring adequate or better were reviewed in Stage 2 of the study.

Stage 2: Review of study quality

Additional search for and selection of studies of measurement properties

A second search was undertaken in MEDLINE, PsycINFO, Embase, and CINAHL (last search date 28 January2015) using the names of the questionnaires that achieved adequate or better in the content evaluation. The search was designed to identify studies evaluating measurement properties of the included questionnaires [Online Resource 1(b)]. The reference lists of included studies were also checked. Titles/abstracts were screened by one reviewer to detect studies that potentially met inclusion criteria. Full-texts of articles were obtained to identify studies published in English as original papers in peer-reviewed journals that reported at least one measurement property of an included questionnaire.

Methodological quality evaluation and quality rating

The COSMIN quality evaluation checklist [20, 21] and scoring system [22] were used for evaluation and rating of methodological quality of studies. The four-point scoring version was used (excellent, good, fair, or poor quality) for each measurement property except criterion validity (there is no gold standard) and cross-cultural validity. Studies that investigated measurement properties of questionnaire versions in languages other than the original language were reviewed provided cross-cultural adaptation processes seemed appropriate according to COSMIN recommendations. The measurement properties internal consistency, reproducibility, measurement error, construct validity (structural validity and hypothesis testing), content validity responsiveness, and interpretability were assessed and scored. Consistent with COSMIN procedures, an overall quality rating (the lowest rating from all items) was given for each measurement property evaluated. The exceptions we made to the COSMIN quality rating procedure are in Online Resource 3. Information on generalisability and results from other tests, including item response theory (IRT) analyses, were collated.

The terms and definitions used to describe measurement properties in many of the papers we assessed were often not consistent with COSMIN. In such cases, we applied the COSMIN taxonomy to determine which property was being reported [16]. A high level of inter-rater agreement for the COSMIN checklist has been demonstrated [23]. Our quality rating and data extraction form were created directly from the COSMIN checklist. The two reviewers in this stage were well versed with the COSMIN Checklist Manual [18]. During the early stages of reviewing, studies were rated by both reviewers to confirm similar interpretation of the criteria and facilitate consistency and accuracy. After 12 papers were rated, agreement was considered sufficiently high for studies to be rated by only one reviewer, with discussion whenever uncertainties arose.

Stage 3: Review of measurement properties of questionnaires

Extraction and rating of measurement property results

Results from the reviewed studies were extracted and collated. Rating for each measurement property could be positive (+), indeterminate (?), or negative (−). This rating method is commonly used to establish the relative merit of questionnaires [24]. Exceptions we made to the rating recommendations made by Terwee et al. [24] were:

1. *Internal consistency*. No upper limit for Cronbach's alpha scores. A negative rating was not given if Cronbach's alpha >0.95. In our opinion, a high Cronbach's alpha for the fatigue questionnaires indicated redundancy rather than a bad scale.
2. *Structural validity*. Rating was based on whether the factor analysis supported uni-dimensionality and thus the validity of the total score.
3. *Content validity*. Rather than rating each study, we rated the evidence accumulated across all studies generating one overall rating.
4. *Hypothesis testing* (*convergent and divergent validity*). Given the lack of explicit hypotheses in the majority of studies, we set the criterion for a positive (+) score as follows: (1) correlations with other fatigue scales ≥0.70, (2) correlations with other constructs predicted to have association with fatigue 0.40–0.70, and (3) correlations with other measures predicted to have minimal association with fatigue <0.40.
5. *Hypothesis testing* (*discriminative validity*). Criterion for a positive (+) score was (1) hypotheses were supported or (2) if no stated hypothesis, a significant difference in fatigue scores was found between groups compared.
6. *Interpretability*. Data did not exist to enable comparison of minimal clinically important change (MCID) with smallest detectable change (SDC). Thus, the rating for interpretability was based on whether or not MCID was ≤15 % of the score range.

Data collation also included characteristics of the questionnaires and details of the studies.

Best evidence synthesis

An overall rating was formulated for each property for each of the questionnaires. This 'best evidence synthesis' was performed using the criteria proposed by Terwee et al. [24], which ranges from − − − to +++ (details provided in Online Resource 4). The synthesis integrates the results from the methodological quality ratings (COSMIN scores), the study results, the number of studies evaluating each measurement property, and the consistency of the ratings. Results from studies with COSMIN score 'poor' were not considered in the best evidence synthesis.

Results

Stage 1: Identification and evaluation of questionnaires

The results of the searches and the flow diagram for the study are shown in Fig. 1. One hundred and eight questionnaires were initially identified and considered for inclusion. Thirty-one were excluded (Online Resource 5). A summary of the results of the content evaluation of the remaining 77 is provided in Online Resource 6. Only ten questionnaires achieved a rating of 'adequate' or better, of which only four were rated 'good' and none 'excellent'. Two questionnaires that have widespread use and availability, only marginally failed one key criterion and therefore were retained in the review despite being categorised as 'inadequate': the vitality subscale of the Medical Outcome Study SF-36 (SF-36 vitality), which did not have sufficient comprehensiveness, and the Functional Assessment of Chronic Illness Therapy Fatigue scale (FACIT-Fatigue), which had too many items potentially overlapping with other constructs.

Stage 2: Methodological quality of studies

The database searches retrieved 2244 records of potential interest. After screening of references, and full-text checking 264 papers, 141 papers remained. A further 26 papers were identified from references lists, resulting in 149 studies reviewed. Nineteen papers tested more than one questionnaire resulting in 169 COSMIN checklists completed for the 12 questionnaires. There was a wide range of methods used to assess measurement properties, some of which were difficult to fit to COSMIN criteria. Some measurement properties investigated in the studies were not included in the COSMIN taxonomy (e.g. Rasch analyses). These are reported descriptively for completeness. Table 1 shows the general properties of the 12 questionnaires reviewed.

The number of studies reporting measurement properties for each questionnaire ranged from only one for the Cancer-Related Fatigue Distress Scale (CRFDS) to 58 for the SF-36 vitality. The methodological quality was highest in studies that had the investigation of measurement properties as the primary aim. Study details, including sample characteristics, language version and scores, are provided in Online Resource 7. The number of studies investigating each property also varied widely. Hypothesis testing occurred in 112 studies with most of them testing multiple hypotheses (2 % were rated excellent quality, 23 % rated poor), internal consistency was investigated in 104 studies (79 % excellent, 1 % poor), interpretability (floor/ceiling effects and/or MCID) in 64, reproducibility in 42 (0 % excellent, 12 % poor), structural validity in 44 (70 % excellent, 5 % poor), responsiveness in 29 (0 % excellent, 50 % poor), and measurement error in only 18 (0 % excellent, 50 % poor). Eighty-one percent of the studies did not appropriately report the numbers of missing items and how the missing items were handled. IRT was used in 16 % of studies.

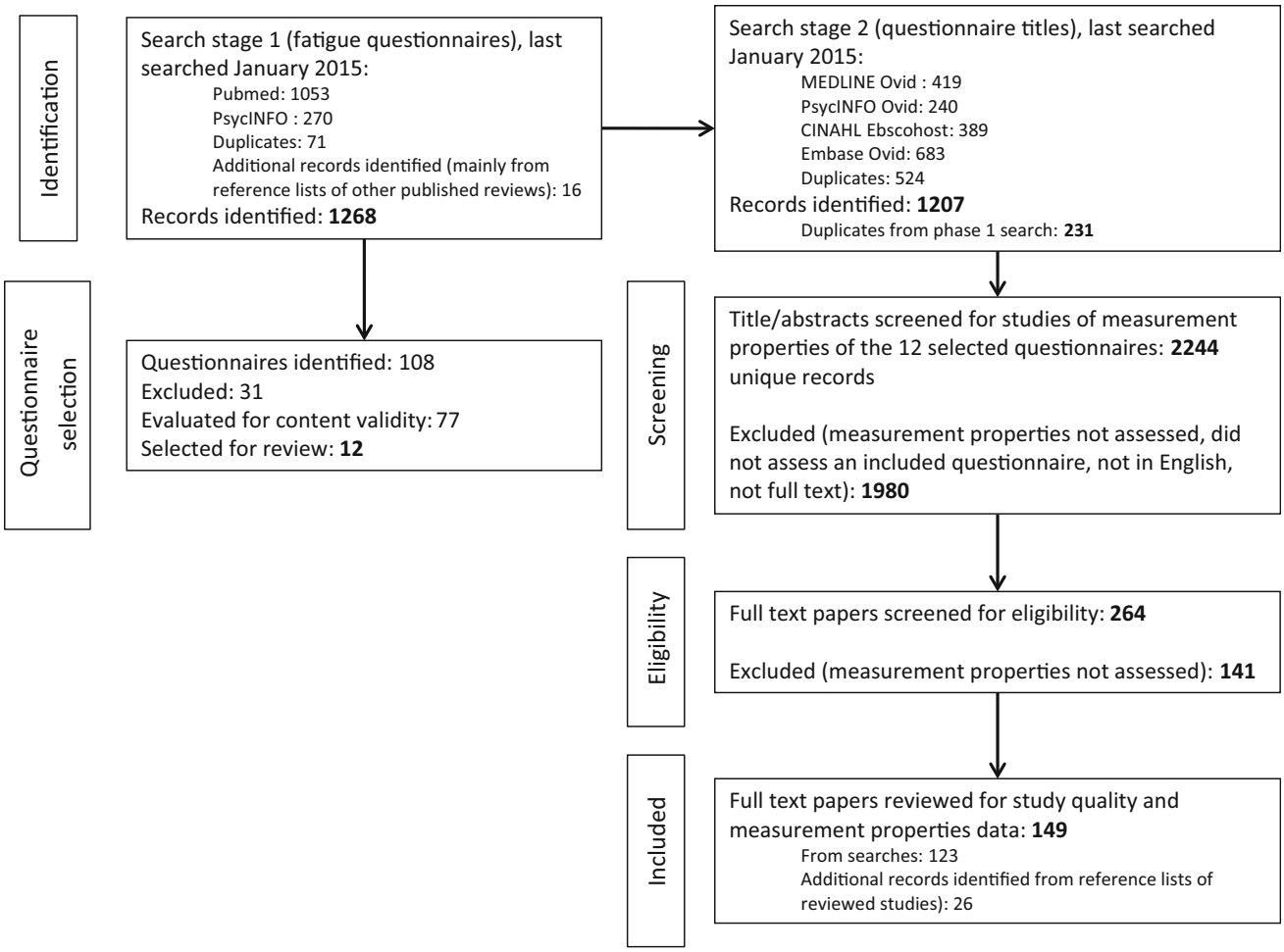

Fig. 1 Flow diagram of the systematic review (based on PRISMA 2009 [52])

Stage 3: Quality of the questionnaires

The results and ratings, for each measurement property, for each of the 12 selected questionnaires, are provided in Online Resource 8. The best evidence synthesis is shown in Table 2, along with additional information about the performance of the questionnaires. Main findings from the review are as follows:

Brief Fatigue Inventory—This questionnaire lacks information regarding reproducibility, measurement error, and responsiveness but performed satisfactorily on other properties. The main disadvantages are the 0–10 numeric rating scales as response options, which is difficult for older people, and the two distinct subscales.

Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire (BRAF-MDQ)—Many studies failed to provide support for this questionnaire, and most properties remain inconclusive. In particular, it may not adequately operate as a uni-dimensional scale, and therefore, the validity of the total score is in doubt. Some concern about the content of some items has been raised despite the thorough content development procedures. In addition, the large number of response options is difficult for older people.

Cancer-Related Fatigue Distress Scale (CRFDS)—This questionnaire currently lacks evidence to support its use. The large number of response options is also undesirable for older people.

Fatigue Associated with Depression (FAsD) questionnaire—Evidence suggests two distinct dimensions are measured by this scale, and therefore, the validity of the total score is in doubt. The scale development was thorough; however, there are items that are not relevant to the majority of older people. The scoring system allows for these items to be disregarded if not applicable, but this approach is not ideal.

Fatigue Impact Scale (FIS)—This questionnaire has considerable evidence to support its use, yet problems include doubt regarding uni-dimensionality, lack of evidence for measurement error and responsiveness, some items may not be applicable to older people, and the questionnaire is too long.

Table 1 General properties of the 12 questionnaires included in the review of measurement properties

Questionnaire (content evaluation rating)	Items (response options)	Range of scores	Comments
Brief Fatigue Inventory	9 (11)	0–10	The questionnaire consists of two components: fatigue severity dimension (three items) and fatigue interference dimension (six items) *Time to complete*: <5 min [53, 54]
Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire	20 (varies from 3–11)	0–70	Questionnaire development specifically targeted fatigue in people with RA, but the authors argue that abnormal fatigue experienced as a result of RA is the same as that experienced in other diseases [55]. Items can be separated into four dimensions which focus on physical (four items), living (seven items), cognition (five items), and emotion (four items) aspects of fatigue
Cancer-related fatigue distress scale	23 (11)	0–230	The questionnaire developed by Holley in 2000 [56] is actually a fatigue distress scale rather than a fatigue scale. However, distress caused by fatigue is an *impact* of fatigue, and therefore, the questionnaire should be consistent with measurement of the construct fatigue *Readability* assessed at third grade level [56]
Fatigue Associated with Depression Questionnaire	13 (5)	1.0–5.0	Two subscales, experience (six items) and impact (seven items). Scores are computed as the mean of answered items within the scale to accommodate for when questions are not answered due to irrelevance to the individual [57]. Version 2 (FAsD-V2) instructions are more generic as attribution of the symptoms to depression was removed [58]
Fatigue Impact Scale or Fisk Fatigue Severity Score	40 (5)	0–160	The questionnaire consists of 40 statements that measure fatigue in three areas: physical, cognitive, and psychosocial. It was made from other existing questionnaires *Readability* level assessed as <grade 8 [59] *Time to complete*: 5–8 min [60], 10–15 min [61], 5–10 min [62]
Functional Assessment of Chronic Illness Therapy (FACIT-Fatigue)	13 (5)	0–52[a]	The questionnaire was originally developed to assess fatigue associated with anaemia in cancer patients but has been successfully administered in a variety of other populations. A link between FACIT-Fatigue and the PROMIS-Fatigue item bank has been established such that scores on the FACIT-Fatigue can be converted to scores from the PROMIS-Fatigue item bank for direct comparison. Our content evaluation found it had too many items not specific to the construct, but it was retained for this review because of its widespread use *Time to complete*: Average time for completion is 15 min [63]
Modified Fatigue Impact Scale	21 (5)	0–84	This questionnaire was derived from the 40-item Fatigue Impact Scale (FIS) to assess the impact of fatigue on physical, cognitive, and psychosocial function [64]. As with the FIS, it was presumed the items could be aggregated into a total score, as well as into separate scores for the three dimensions [65]
Parkinson Fatigue Scale	16 (5)	1.0–5.0	The questionnaire was developed with a focus on the physical rather than the emotional or cognitive aspects of fatigue [28]
Perform Questionnaire	12 (5)	12–60[a]	The questionnaire includes beliefs and attitudes of patients about fatigue. Original language was Spanish. There have been no formal translations using cross-cultural validation procedures into English or any other language *Time to complete*: Mean time required <9 min [29]

Table 1 continued

Questionnaire (content evaluation rating)	Items (response options)	Range of scores	Comments
PROMIS-Fatigue Short Form 8a/full item bank	8/95 (5)	8–40/CAT scoring[b]	The PROMIS project aimed to develop item banks that achieve both precision and range in the measurement of patient-reported outcomes. Item response theory (IRT) was used during item bank development, and items can be presented in a computerized adaptive testing (CAT) format based on the IRT results. Short forms can be constructed from items either to cover the entire spectrum of fatigue severity or to target a certain range in the fatigue continuum
			The SF8a is the stand-alone short form that was evaluated in the first stage of this study. Other short forms exist, including SF7 and SF-MS, which were not specifically content evaluated. All studies on the PROMIS-Fatigue item bank were eligible to be included in the review of measurement properties. The paper by Kalkanis et al. [4] used 10 items from the PROMIS-Fatigue item bank (not clear which 10 items were selected or why). The Yost et al. [66] study reported on two fatigue short forms that were created from the PROMIS-Fatigue item bank, one with 17 and the other with 7 items. Broderick et al. [67] used both a seven-item short form plus the full CAT experience in their study
			Time to complete: Mean time 41 (18) seconds, *Number of items*: Mean number 4.2 [68]
SF-36 vitality	4 (6)	0–100[a]	The vitality subscale is part of the Medical Outcome Study health status questionnaire. It is very short and lacks comprehensiveness but was retained in this review as an exception because of its widespread availability and use.
Uni-dimensional Fatigue Impact Scale	22 (5)	0–88	An initial Rasch analysis of the original Fatigue Impact Scale (FIS) indicated that the subscales of the FIS could not be combined to create a uni-dimensional measure of fatigue impact (overall fit Chi-squared $P < 0.01$). It also revealed that there were a number of items that should be removed from the scale due to item misfit (five items) or DIF (four items). In addition, there were too few items measuring at the mild end of the scale. The FIS was modified to create the 22-item U-FIS which includes improvements such as a reduced number of items, shorter recall period and the ability to yield an empirically valid total score

DIF differential item functioning, *MS* multiple sclerosis, *PNH* paroxysmal nocturnal haemoglobinuria, *PROMIS* Patient-Reported Outcomes Measurement Information System, *SLE* systemic lupus erythematosus

[a] Questionnaires with lowest score representing the worst level of fatigue

[b] Scores reported on a T score metric (mean 50, SD 10) that is anchored to the distribution of scores in the US general population

FACIT-Fatigue—This scale has been used in a great many studies involving patients with a wide range of diseases. Our content evaluation identified several items that lacked specificity to fatigue. Other studies have similarly identified problems with up to four items, including the item regarding daytime sleepiness and being too tired to eat. These items may be measuring different underlying latent constructs or irrelevant information [15]. These findings support the concerns raised in our content evaluation. However, it has good coverage across the spectrum of fatigue and performed well on many of the measurement property tests. Some modifications have been suggested including removal of troublesome items [25], which may result in a very good questionnaire. About two-thirds of studies included samples with substantial proportions of older participants. One study focused on older people [26], providing evidence in support of internal consistency and hypothesis testing, and the Rasch model analysis.

Modified Fatigue Impact Scale (MFIS)—This scale performed the worst of those reviewed with negative

Table 2 Best evidence synthesis

Questionnaire	IC	RR	ME	SV	CV	HT	Res	Int	Other results	Modifications and recommendations for improvements
Brief Fatigue Inventory (BFI)	+++	?	0	+++	?	+++	?	+	Patients frequently had difficulty distinguishing the effect of their neurological impairment from the effect of fatigue. Also they found it difficult to use the 0–10 numeric rating scale and the question about 'work' was not relevant [34]	(1) Participants asked to comment on questionnaire items and their satisfaction with the questionnaire [54]. Two additional items were subsequently added (2) BFI modified by a) using a 0–7 scale for the response options instead of the 11-point scale, and b) increased the recall period from 24 h to 7 days [69].
Bristol Rheumatoid Arthritis Fatigue Multi-Dimensional Questionnaire (BRAF-MDQ)	+++*	?	?	?	+	+	?	?	Precision is best at higher levels of fatigue [70]	Two-round Delphi process carried out with a group of expert clinicians and RA patients [71]. As a result removal of the item 'Over the past 7 days have you felt embarrassed because of fatigue', and slight modifications to three other items, were recommended
Cancer-related fatigue distress scale (CRFDS)	+++	0	0	+++	+	0	0	0		
Fatigue Associated with Depression Questionnaire (FAsD)	+++*	0	0	− − −	+	?	+	+		The US Food & Drug administration recommended alterations to instructions to remove attribution of symptoms to a cause [58], resulting in FAsD-V2
Fatigue Impact Scale (FIS) or Fisk Fatigue Severity Score	+++*	+++	?	?	?	++	0	+	One study highlighted problems with items 28 ('I am less able to provide financial support for myself and my family') and 29 ('I engage in less sexual activity') for their older patients. They noted that removing these items did not substantially change the internal consistency and slightly improved the structure [72]	Scale reduced from 40 down to 25 items and called FIS-25 [73]. The reduction was based on theory, modification indices, and factor loadings, using data from a sample with mean age 62 years. Factor analysis of the new FIS-25 showed the nested factor model, comprising one general and three sub-dimensions, showed the best fit. Internal consistency, reproducibility and discriminative validity were all good. Many of the items that were omitted refer to situations less applicable to older people
Functional Assessment of Chronic Illness Therapy (FACIT-Fatigue)	+++	+++	?	+++	?	+++	?	?	FACIT had markedly better discrimination across the range of fatigue, particularly at average to high fatigue levels, than SF-36 vitality [74] Entire range fatigue was covered, with the exception of those with very little fatigue [8, 26, 74, 75] IRT analyses suggest that some changes to wording of items and the removal of the sleepiness item are required to improve the scale [26, 76] Three of these items ('I am able to do my usual activities', 'I need to sleep during the day' and 'I am too tired to eat') also demonstrated DIF and item misfit [8, 15]	Item reduction techniques showed the questionnaire performed better with 9 items [25]. One item removed because it had low correlation with other items and a high floor effect ('I am too tired to eat'). Two had high floor effect ('I need help to do my usual activities' and 'I need to sleep during the day'). The final item ('I feel tired') was considered redundant. However, the 9-item version was not uni-dimensional in their sample A four-item version was created from FACIT-An and called the Simplified

Table 2 continued

Questionnaire	IC	RR	ME	SV	CV	HT	Res	Int	Other results	Modifications and recommendations for improvements
									Crosswalk table has been developed linking FACIT-Fatigue scores to PROMIS-Fatigue item bank scores [36]	Evaluation of Fatigue (SEF) [77]. Items included were 'I am too tired to eat', 'I am forced to spend time in bed', 'I feel fatigued', and 'I feel weak all over'. Internal consistency reliability was >0.83, construct validity and responsiveness were supported. The estimated important difference was 1–2 points. However, the items selected have questionable content validity according to our criteria and other studies
Modified Fatigue Impact Scale (MFIS)	+++*	+++	− −	− − −	?	?	?	?	Analyses suggest some redundancy in the scale [78] Authors concluded MFIS cannot be used to generate a single overall score of fatigue [27]	A modified version exists [62]: (1) The response range was changed from 'past 4 weeks' to 'past week'. (2) Three items (4, 13 and 20) from the original scale were deleted because the content did not reflect the spinal cord injury experience. (3) Three items were added: 'I have had difficulty paying attention for short periods of time', 'I have trouble maintaining physical effort for short periods', and 'I have avoided/eliminated certain tasks, activities and lifestyles'. (4) Phrases 'away from home' and 'at home or at work' were removed from two items
Parkinson Fatigue Scale (PFS-16)	+++	+	?	+++	+	++	0	+	Item thresholds were relatively evenly spread along a range that covered the sample distribution, except those with lowest and highest fatigue [26] Evidence of DIF among those classified as non-fatigued (higher responses for 'need to sleep or rest' item) [26]	
Perform Questionnaire (PQ)	+++*	+	0	?	+	+	+	+	16 % did not answer item 8 ('felt bad about feeling tired at work') Questionnaire could be improved by removing/modifying item 12 [29]	
PROMIS-Fatigue short form 8a and full item bank	+++	++	?	+++	+	+++	0	+	CAT was compared to three generated short forms, CAT showed consistently better precision than short forms across the range of fatigue severity [79] Different methods of administration were compared. No statistically or clinically significant differences in score levels or measurement properties were found for interview, paper questionnaire, PDA or PC administration [80] Crosswalk tables have been developed using IRT methods linking PROMIS-Fatigue item bank with FACIT-Fatigue and SF-36 vitality [36]	The 92-item item bank had 20 items with misfit. With these items removed, the 72-item item bank performed well on IRT. Five items still had misfit slightly greater than the chosen cut-off but were retained [30] A modified 'daily' version of PROMIS-Fatigue SF7 was examined with IRT in general population (GP) and osteoarthritis samples. Found negligible DIF, good internal reliability from average to high levels of fatigue, and evidence for discriminative validity and sensitivity to change. But, 30 % floor effect in GP, inability to discriminate between low levels of fatigue with the daily measure, and differences in the first 2 days' scores after which scores stabilised [81]

Table 2 continued

Questionnaire	IC	RR	ME	SV	CV	HT	Res	Int	Other results	Modifications and recommendations for improvements
SF-36 Vitality	?	?	–	+++	+	+++	?	?	Scale showed an unexpectedly powerful capacity to discriminate between groups differing in mental distress [31] Several studies demonstrated scaling success [82–88], however, scaling failed for those >75 years [89] Several studies show that items correlated better with mental health subscale than with vitality scale [90–93], and correlation between vitality subscale was very high with mental health subscale and very low with physical health subscale [92–94] The questionnaire can differentiate people with relatively low fatigue well, but not people with moderate-to-severe fatigue [70, 74, 75] Study reported 5.5 % of participants considered items not relevant to them [83] Crosswalk table has been developed linking SF-36 vitality scores to PROMIS-Fatigue item bank scores [36]	
Uni-dimensional Fatigue Impact Scale (U-FIS)	+++	?	0	++	+	++	0	+		Rasch analysis indicated a four-point response option (with two responses collapsed) gives the optimal response scoring [37]

Possible ratings for content validity and interpretability +, −, 0 or ? only

CAT Computerized adaptive testing, *DIF* differential item functioning, *IRT* item response theory, *RA* rheumatoid arthritis, *IC* internal consistency, *RR* reproducibility, *ME* measurement error, *CV* content validity, *SV* structural validity, *HT* hypothesis testing, *Res* responsiveness, *Int* interpretability

* Where evidence for uni-dimensionality was lacking, the rating for methodological quality of the internal consistency studies may be overestimated and the rating for internal consistency of the total score may be inaccurate

findings for measurement error and structural validity, and mixed or inconclusive results for hypothesis testing, responsiveness, and interpretability. One study showed that the items cannot be aggregated into a single overall score [27].

Parkinsons Fatigue Scale—This questionnaire achieved positive scores for six out of the eight properties rated and has reasonable coverage across the range of fatigue. A unique aspect of the questionnaire development was the specific aim to minimise the overlap with other symptoms of Parkinson's disease [28], and it achieved one of the highest scores for relevance and specificity in our content evaluation. All studies included older people, adding further support for recommending this questionnaire. More evidence is required for measurement error and responsiveness.

Perform Questionnaire—This questionnaire also performed positively on six of eight properties, and again, all studies included older people. Evidence is lacking for measurement error and inconclusive for structural validity. One problematic item has been identified [29]. It is relatively short and appropriate for older people but requires formal cross-cultural adaptation from Spanish language.

PROMIS-Fatigue item bank and short forms—Several high-quality studies, most including older participants, provide good support for the item bank and the short forms taken from the item bank. The PROMIS item bank also has the advantage of extensive IRT analysis and the option of computerized adaptive testing (CAT) which offers the advantages of simultaneously minimising burden and missing data. The PROMIS items have been subjected to thorough development and content validation. However, misfit was found by one study, which led the authors to recommend removal of 20 items, with a further five still having slight misfit [30]. Misfit is likely due to items overlapping with other physical problems. For example, 'I have had energy to climb more than one flight of stairs'. More information is needed for measurement error and responsiveness. Overall, PROMIS-Fatigue appears to be the best available questionnaire due to the extensive evaluation and positivity of findings, plus the practical advantages of being an item bank.

SF-36 vitality—SF-36 vitality received positive ratings for structural validity, content validity, and hypothesis testing. However, there are doubts regarding internal consistency, reproducibility, measurement error, responsiveness, and interpretability. It lacks face validity for some items, and it is not good at separating those with higher levels of fatigue. This was the most extensively investigated of the questionnaires. While most studies focused on younger cohorts, there were nine studies with exclusively older participants and 16 others that included significant numbers of older people. Synthesising results from these 25 studies affected the findings for internal consistency (+++), measurement error (?), hypothesis testing (?), and interpretability (+)—an overall improvement. Several properties still showed mixed results. Evidence from several studies suggests that 'vitality' may not be equivalent to 'lack of fatigue', and other studies suggest the subscale may be measuring mental health or mental distress instead of vitality [31]. Our content evaluation also found it lacking. Thus, there is doubt about exactly what latent construct it is measuring despite its reasonable performance on most measurement properties. On the other hand, a recent study using a bi-factor factor analysis model showed that vitality can be considered a uni-dimensional construct with fatigue and energy representing the positive and negative sub-domains of the construct [32]. Plus, several authors are in favour of using SF-36 vitality as a measure of fatigue/exhaustion [33–35]. The full SF-36 health-related quality-of-life questionnaire is very widely used. As a result, information is frequently available even if measurement of fatigue was not intended. A crosswalk table to convert scores to PROMIS-Fatigue item bank t-scores has been produced [36]. Thus SF-36 vitality may be a useful proxy indicator of fatigue especially from existing population data banks.

Uni-dimensional Fatigue Impact Scale (U-FIS)—This questionnaire is another modification of the FIS and performs well on most measurement properties. One study showed that reducing the response options improved performance and this modification should be considered by future users [37]. Along with the FIS and the MFIS, the studies on U-FIS included predominantly younger people with almost no studies including significant numbers of older people. As a result, our recommendation to use this questionnaire for measuring fatigue in older people is not without some reservation.

Discussion

Measurement of fatigue is challenging. In this study, we identified existing fatigue questionnaires that met our inclusion criteria and evaluated their content against a priori criteria. Twelve questionnaires found to have adequate content were then reviewed for merit of measurement properties and appropriateness for older people.

We conclude that the FACIT-Fatigue, Parkinsons Fatigue Scale, Perform Questionnaire, PROMIS-Fatigue, and U-FIS can be recommended, although none were free of problems. All require further evaluation of at least two measurement properties. Minor modification to content is warranted for FACIT-Fatigue scale, Perform Questionnaire, and PROMIS-Fatigue item bank and to the response

options for U-FIS. We primarily support the PROMIS-Fatigue item bank because of its rigorous development and CAT-readiness. However, its performance may depend on exactly which items are used, and whether the selected items have overlap with other problems. Most of the studies supporting FACIT-Fatigue, Parkinsons Fatigue Scale, Perform Questionnaire and PROMIS-Fatigue included older participants, providing external validity to the findings for these questionnaires. The BRAF-MDQ, FAsD questionnaire, FIS, MFIS, and SF-36 vitality did not perform well enough on measurement property evaluation. The Brief Fatigue Inventory and CRFDS may be acceptable and are worthy of further investigation, but cannot be recommended at this stage.

Several reviews of self-report fatigue questionnaires have been published [10–14, 38–40]. Limitations of previous reviews include searches older than 5 years [10–12], reviewing only disease-specific measures [13, 14, 40], and limited search strategies [10, 38, 39]. Only one, which reviewed measures of fatigue for neurological conditions [13], critically appraised the methodological quality of the studies. Half of the reviews only considered studies that had been carried out on patients from their population of interest [13, 14, 39, 40]. Recommendations were made by three of the reviews, which include the Neurological Fatigue Indexes for neurological diseases [14], U-FIS or Fatigue Scale for Motor and Cognitive Functions for multiple sclerosis [13], FACIT-Fatigue or Fatigue Severity Scale for Parkinsons disease [13], Profile of Mood States Fatigue subscale for stroke [13], and FACIT-Fatigue for systemic lupus erythematosus [39]. All review authors commented that information on measurement properties for these scales was still lacking, and users should consider the details of the construct of fatigue purported to be measured by the scales before using it, as this was often outdated or inexplicit. The remaining reviews fell short of making recommendations for specific questionnaires due to inconclusive findings, but gave guidance on how to choose from the many available scales and summarised the findings from questionnaire evaluations [10–12, 38, 40]. The present review identified many more questionnaires than all the previous reviews, critically appraised the methodological quality of studies, and generated recommendations following best evidence synthesis.

Importantly, previous reviews did not consider the content of each questionnaire they evaluated. Lack of content validity is a strong argument for not using an instrument at all [17] (p. 155). Our inclusion of a content evaluation addresses this important omission and ensures the science of measurement of fatigue is keeping up with the knowledge of the problem. Items were assessed to ensure key qualitative aspects of fatigue were included and that overlap with non-fatigue-related constructs was avoided. This second element presented challenges because we found many items in many questionnaires that undoubtedly are consistent with fatigue, but could also be consistent with other problems. For example, 'Do you have difficulty concentrating?' Difficulty concentrating can be due to many disorders. If the wording had been 'Does your tiredness make it difficult for you to concentrate?' we would have considered it acceptable. Some popular questionnaires were excluded from this review because we found too many items had this problem of potential overlap, including the Fatigue Severity Scale [41] and the Chalder Fatigue Scale [42].

There are several limitations of the methods and available data in this study. Firstly, good questionnaires may have been excluded because of lack of evaluation of their measurement properties or studies written in non-English languages. In particular, we suggest Fatigue Pictogram [43, 44] and Symptom Fatigue Scale [45] warrant further consideration. Secondly, our search strategy may have missed studies. Qualitative studies that informed item development may be missing because they were published before questionnaires were given their name. Studies that did not focus primarily on evaluating measurement properties may be missing. However, these were likely to have lower scores for methodological quality anyway. Thirdly, while we attempted to create an objective content rating scale, there was still an element of subjectivity in determining whether criteria were met.

Having one reviewer for studies is a possible source of errors in methodological quality ratings. However, the COSMIN checklist has very good guidance documents, and the two reviewers in our study worked closely together to ensure accuracy and consistency. In our experience, most variation between scorers occurs at the top end of ratings (good vs. excellent) where errors have little impact on the final rating for the measurement property. Further, because of the very large number of studies included in our review, there was often a lot of data available during formulation of conclusions. As a result, occasional errors in quality rating would have had little impact on the final outcomes of the best evidence synthesis. Finally, the COSMIN manual acknowledges that some items need a subjective judgement and that how to deal with lack of reporting in the original article is unclear. Thus, there are grey areas where consistency is the goal of having two raters, rather than 'avoiding errors'. We are confident our review conclusions are not impaired by having only one rater for each study.

We used some modifications to the COSMIN methodological quality rating system and the criteria for positive measurement property scores. These were largely adaptations to suit the uniqueness of fatigue measurement and the shortcomings of the studies. Two COSMIN items that were

omitted from the final methodological quality ratings for most of the properties were about the number of missing items and handling of missing items. The majority of studies failed these criteria, and the methodological quality rating would have been 'fair' for otherwise good or excellent studies. These studies would then not have had prominence in formulating the final synthesised rating. As the questions are about the *reporting* of missing items, not whether missing items would have biased study results, omitting these two questions from the methodological quality rating is unlikely to have biased our conclusions towards favouring questionnaires with poorer quality properties.

The exceptions we made to our predefined selection criteria for questionnaires to review, by including FACIT-Fatigue and SF-36 vitality, are important to acknowledge. It is interesting that quantitative studies found problems with misfit for some FACIT-Fatigue items, and problems with convergent and divergent construct validity of SF-36 vitality. These findings support the validity of our content evaluation. Modifications to improve FACIT-Fatigue have been suggested, and we recommend SF-36 vitality be used as a fatigue measure with some caution because of concern regarding the latent construct it actually measures.

Our recommendations are indeed limited by both the quality and the quantity of the studies. The quality was generally disappointing especially for measurement error and responsiveness, where quantity was also lacking. The quantity of studies can also paradoxically have a negative effect on a questionnaire. Lack of evidence is not the same as evidence in support of a questionnaire. A good questionnaire may have more flaws identified than a lesser questionnaire, simply because of better quality and greater numbers of evaluation studies. This problem is difficult to accommodate within the COSMIN methodology, and some subjectivity in formulating conclusions and recommendations was necessary.

The generalisability of the measurement property findings to older people is a final potential limitation. The ability to apply measures developed for one clinical population to another is an important issue; however, it is not feasible to test every possible subpopulation in which a measure might be used. Consideration of the content is helpful for gaining confidence that a measure is suitable for the new purpose [46].

For the purpose of measuring of fatigue, questionnaires need to be able to generate a valid total score. For this reason, we made uni-dimensionality a requirement. This is appropriate since several studies give support to fatigue being a uni-dimensional construct [9, 47, 48] despite there being several possible sub-domains [49, 50]. We therefore negatively scored questionnaires that lacked support for uni-dimensionality; however, it does not mean that these questionnaires lack structural validity. Subscale scores from questionnaires with subscales may be of interest depending on the specific questions to be addressed with the data [50].

The very large number of questionnaires identified through our search may reflect a large interest in the problem of fatigue but leads to difficulty synthesising results from different studies and impedes progress towards understanding and managing the problem. The disease-specific approach and the multiple other uses of the word fatigue, which causes confusion for questionnaire respondents as well as clinicians and researchers [51], have likely lead to the plethora of questionnaires. In our opinion, the word fatigue continues to create problems in questionnaires because of differing individual interpretations of its meaning. For this reason, we believe that the item, for example, 'How often did you feel tired even when you had not done anything?' is preferable to 'How often did your fatigue interfere with your social activities?' Both items are from the PROMIS-Fatigue item bank.

Better understanding of fatigue, and progress in measurement science, necessitates that measurement tools be re-evaluated. Our findings suggest that questionnaires are currently available with sufficient support for their use, although further improvements and evaluation are still warranted. The best fatigue questionnaires appear to be FACIT-Fatigue, Parkinsons Fatigue Scale, Perform Questionnaire, PROMIS-Fatigue item bank and short forms, and U-FIS. PROMIS-Fatigue has several practical advantages as well as extensive evaluation and is our primary recommendation. The final choice of questionnaire will come down to the precise context and purpose of the clinician or researcher.

Acknowledgments TE's post-doctoral position was funded by the Norwegian Women's Health Association and the Norwegian Extra Foundation for Health and Rehabilitation through EXTRA funds.

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

Ethical standard This article only contains data from previously published studies. No patient data were collected since it is a systematic review. It does not contain any studies with human participants or animals performed by any of the authors.

References

1. Alexander, N. B., Taffet, G. E., Horne, F. M., Eldadah, B. A., Ferrucci, L., Nayfield, S., et al. (2010). Bedside-to-Bench conference: Research agenda for idiopathic fatigue and aging. *Journal of the American Geriatrics Society, 58*(5), 967–975. doi:10.1111/j.1532-5415.2010.02811.x.
2. Gill, T. M., Desai, M. M., Gahbauer, E. A., Holford, T. R., & Williams, C. S. (2001). Restricted activity among community-

living older persons: Incidence, precipitants, and health care utilization. *Annals of Internal Medicine, 135*(5), 313–321.
3. Leveille, S. G., Fried, L., & Guralnik, J. M. (2002). Disabling symptoms: What do older women report? *Journal of General Internal Medicine, 17*(10), 766–773.
4. Kalkanis, A., Yucel, R. M., & Judson, M. A. (2013). The internal consistency of PRO fatigue instruments in sarcoidosis: Superiority of the PFI over the FAS. *Sarcoidosis Vasculitis & Diffuse Lung Diseases, 30*(1), 60–64.
5. Jones, D. E., Gray, J. C., & Newton, J. (2009). Perceived fatigue is comparable between different disease groups. *QJM, 102*(9), 617–624. doi:10.1093/qjmed/hcp091.
6. Hagell, P., Rosblom, T., & Palhagen, S. (2012). A Swedish version of the 16-item Parkinson fatigue scale (PFS-16). *Acta Neurologica Scandinavica, 125*(4), 288–292. doi:10.1111/j.1600-0404.2011.01560.x.
7. Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A., et al. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Quality of Life Research, 19*(9), 1311–1321. doi:10.1007/s11136-010-9694-5.
8. Hobart, J., Cano, S., Baron, R., Thompson, A., Schwid, S., Zajicek, J., et al. (2013). Achieving valid patient-reported outcomes measurement: A lesson from fatigue in multiple sclerosis. *Multiple Sclerosis, 19*(13), 1773–1783.
9. 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. doi:10.1007/s11136-006-0060-6.
10. Dittner, A. J., Wessely, S. C., & Brown, R. G. (2004). The assessment of fatigue: A practical guide for clinicians and researchers. *Journal of Psychosomatic Research, 56*(2), 157–170. doi:10.1016/S0022-3999(03)00371-4.
11. Whitehead, L. (2009). The measurement of fatigue in chronic illness: A systematic review of unidimensional and multidimensional fatigue measures. *Journal of Pain and Symptom Management, 37*(1), 107–128.
12. Mota, D. D., & Pimenta, C. A. (2006). Self-report instruments for fatigue assessment: A systematic review. *Research and Theory of Nursing Practice, 20*(1), 49–78.
13. Elbers, R. G., Rietberg, M. B., van Wegen, E. E., Verhoef, J., Kramer, S. F., Terwee, C. B., et al. (2012). Self-report fatigue questionnaires in multiple sclerosis, Parkinson's disease and stroke: A systematic review of measurement properties. *Quality of Life Research, 21*(6), 925–944. doi:10.1007/s11136-011-0009-2.
14. Tyson, S. F., & Brown, P. (2014). How to measure fatigue in neurological conditions? A systematic review of psychometric properties and clinical utility of measures used so far. *Clinical Rehabilitation, 28*(8), 804–816. doi:10.1177/0269215514521043.
15. Lai, J. S., Cella, D., Chang, C. H., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research, 12*(5), 485–501.
16. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. doi:10.1016/j.jclinepi.2010.02.006.
17. de Vet, H. C. W., Terwee, C. B., & Mokkink, L. B. (2011). *Practical guides to biostatistics and epidemiology: Measurement in medicine: A practical guide*. Cambridge, GBR: Cambridge University Press.
18. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2012). COSMIN checklist manual. www.cosmin.nl: Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research.
19. Brink, T. L., Yesavage, J. A., Lum, O., Heersema, P. H., Adey, M., & Rose, T. L. (1982). Screening tests for geriatric depression. *Clinical Gerontologist, 1*(1), 37–43.
20. Mokkink, L. B., Terwee, C. B., Stratford, P. W., Alonso, J., Patrick, D. L., Riphagen, I., et al. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research, 18*(3), 313–333. doi:10.1007/s11136-009-9451-9.
21. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549. doi:10.1007/s11136-010-9606-8.
22. Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 21*(4), 651–657. doi:10.1007/s11136-011-9960-1.
23. Mokkink, L. B., Terwee, C. B., Gibbons, E., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010). Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. *BMC Medical Research Methodology, 10*, 82. doi:10.1186/1471-2288-10-82.
24. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42. doi:10.1016/j.jclinepi.2006.03.012.
25. Al-shair, K., Muellerova, H., Yorke, J., Rennard, S. I., Wouters, E. F., Hanania, N. A., et al. (2012). Examining fatigue in COPD: Development, validity and reliability of a modified version of FACIT-F scale. *Health & Quality of Life Outcomes, 10*, 100. doi:10.1186/1477-7525-10-100.
26. Nilsson, M. H., Bladh, S., & Hagell, P. (2013). Fatigue in Parkinson's disease: Measurement properties of a generic and a condition-specific rating scale. *Journal of Pain and Symptom Management, 46*(5), 737–746. doi:10.1016/j.jpainsymman.2012.11.004.
27. Mills, R. J., Young, C. A., Pallant, J. F., & Tennant, A. (2010). Rasch analysis of the Modified Fatigue Impact Scale (MFIS) in multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry, 81*(9), 1049–1051. doi:10.1136/jnnp.2008.151340.
28. Brown, R. G., Dittner, A., Findley, L., & Wessely, S. C. (2005). The Parkinson fatigue scale. *Parkinsonism Related Disorders, 11*(1), 49–55. doi:10.1016/j.parkreldis.2004.07.007.
29. Baro, E., Carulla, J., Cassinello, J., Colomer, R., Mata, J. G., Gascon, P., et al. (2011). Psychometric properties of the Perform Questionnaire: A brief scale for assessing patient perceptions of fatigue in cancer. *Supportive Care in Cancer, 19*(5), 657–666. doi:10.1007/s00520-010-0878-x.
30. Lai, J. S., Cella, D., Dineen, K., Bode, R., Von Roenn, J., Gershon, R. C., et al. (2005). An item bank was created to improve the measurement of cancer-related fatigue. *Journal of Clinical Epidemiology, 58*(2), 190–197. doi:10.1016/j.jclinepi.2003.07.016.
31. Persson, L. O., Karlsson, J., Bengtsson, C., Steen, B., & Sullivan, M. (1998). The Swedish SF-36 Health Survey II. Evaluation of clinical validity: Results from population studies of elderly and women in Gothenborg. *Journal of Clinical Epidemiology, 51*(11), 1095–1103.

32. Deng, N., Guyer, R., & Ware, J. E. (2015). Energy, fatigue, or both? A bifactor modeling approach to the conceptualization and measurement of vitality. *Quality of Life Research, 24*(1), 81–93. doi:10.1007/s11136-014-0839-9.
33. Lindeberg, S. I., Ostergren, P. O., & Lindbladh, E. (2006). Exhaustion is differentiable from depression and anxiety: Evidence provided by the SF-36 vitality scale. *Stress, 9*(2), 117–123. doi:10.1080/10253890600823485.
34. Mead, G., Lynch, J., Greig, C., Young, A., Lewis, S., & Sharpe, M. (2007). Evaluation of fatigue scales in stroke patients. *Stroke, 38*(7), 2090–2095. doi:10.1161/strokeaha.106.478941.
35. Brown, L. F., Kroenke, K., Theobald, D. E., & Wu, J. (2011). Comparison of SF-36 vitality scale and Fatigue Symptom Inventory in assessing cancer-related fatigue. *Supportive Care in Cancer, 19*(8), 1255–1259. doi:10.1007/s00520-011-1148-2.
36. 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. doi:10.1016/j.jpainsymman.2013.12.236.
37. Meads, D. M., Doward, L. C., McKenna, S. P., Fisk, J., Twiss, J., & Eckert, B. (2009). The development and validation of the Unidimensional Fatigue Impact Scale (U-FIS). *Multiple Sclerosis, 15*(10), 1228–1238. doi:10.1177/1352458509106714.
38. Shahid, A., Shen, J., & Shapiro, C. M. (2010). Measurements of sleepiness and fatigue. *Journal of Psychosomatic Research, 69*(1), 81–89. doi:10.1016/j.jpsychores.2010.04.001.
39. Holloway, L., Humphrey, L., Heron, L., Pilling, C., Kitchen, H., Hojbjerre, L., et al. (2014). Patient-reported outcome measures for systemic lupus erythematosus clinical trials: A review of content validity, face validity and psychometric performance. *Health and Quality of Life Outcomes,*. doi:10.1186/s12955-014-0116-1.
40. Barsevick, A. M., Cleeland, C. S., Manning, D. C., O'Mara, A. M., Reeve, B. B., Scott, J. A., et al. (2010). ASCPRO recommendations for the assessment of fatigue as an outcome in clinical trials. *Journal of Pain and Symptom Management, 39*(6), 1086–1099. doi:10.1016/j.jpainsymman.2010.02.006.
41. Krupp, L. B., LaRocca, N. G., Muir-Nash, J., & Steinberg, A. D. (1989). The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Archives of Neurology, 46*(10), 1121–1123.
42. Chalder, T., Berelowitz, G., Pawlikowska, T., Watts, L., Wessely, S., Wright, D., et al. (1993). Development of a fatigue scale. *Journal of Psychosomatic Research, 37*(2), 147–153.
43. Fitch, M. I., Bunston, T., Bakker, D., Mings, D., & Sevean, P. (2011). The fatigue pictogram: Psychometric evaluation of a new clinical tool. *Canadian Oncology Nursing Journal, 21*(4), 205–217.
44. Fitch, M. I., Bunston, T., & Mings, D. (2012). The Fatigue Pictogram: Assessing the psychometrics of a new screening tool. *Canadian Oncology Nursing Journal, 22*(1), 42–52.
45. Kjerulff, K. H., & Langenberg, P. W. (1995). A comparison of alternative ways of measuring fatigue among patients having hysterectomy. *Medical Care, 33*(4 Suppl), AS156–AS163.
46. Magasi, S., Ryan, G., Revicki, D., Lenderking, W., Hays, R. D., Brod, M., et al. (2012). Content validity of patient-reported outcome measures: Perspectives from a PROMIS meeting. *Quality of Life Research, 21*(5), 739–746. doi:10.1007/s11136-011-9990-8.
47. 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. doi:10.1007/s00520-010-0971-1.
48. Lai, J. S., Beaumont, J. L., Ogale, S., Brunetta, P., & Cella, D. (2011). Validation of the functional assessment of chronic illness therapy-fatigue scale in patients with moderately to severely active systemic lupus erythematosus, participating in a clinical trial. *Journal of Rheumatology, 38*(4), 672–679. doi:10.3899/jrheum.100799.
49. Cella, M., & Chalder, T. (2010). Measuring fatigue in clinical and community settings. *Journal of Psychosomatic Research, 69*(1), 17–22. doi:10.1016/j.jpsychores.2009.10.007.
50. McDonald, R. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Earlbaum.
51. Egerton, T. (2013). Self-reported aging-related fatigue: A concept description and its relevance to physical therapist practice. *Physical Therapy, 93*(10), 1403–1413. doi:10.2522/ptj.20130011.
52. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology, 62*(10), 1006–1012. doi:10.1016/j.jclinepi.2009.06.005.
53. Lin, C. C., Chang, A. P., Chen, M. L., Cleeland, C. S., Mendoza, T. R., & Wang, X. S. (2006). Validation of the Taiwanese version of the Brief Fatigue Inventory. *Journal of Pain and Symptom Management, 32*(1), 52–59. doi:10.1016/j.jpainsymman.2005.12.019.
54. 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.
55. Nicklin, J., Cramp, F., Kirwan, J., Urban, M., & Hewlett, S. (2010). Collaboration with patients in the design of patient-reported outcome measures: Capturing the experience of fatigue in rheumatoid arthritis. *Arthritis Care & Research, 62*(11), 1552–1558. doi:10.1002/acr.20264.
56. Holley, S. K. (2000). Evaluating patient distress from cancer-related fatigue: An instrument development study. *Oncology Nursing Forum, 27*(9), 1425–1431.
57. Matza, L. S., Phillips, G. A., Revicki, D. A., Murray, L., & Malley, K. G. (2011). Development and validation of a patient-report measure of fatigue associated with depression. *Journal of Affective Disorders, 134*(1–3), 294–303. doi:10.1016/j.jad.2011.06.028.
58. Matza, L. S., Murray, L. T., Phillips, G. A., Konechnik, T. J., Dennehy, E. B., Bush, E. N., et al. (2015). Qualitative Research on Fatigue Associated with Depression: Content Validity of the Fatigue Associated with Depression Questionnaire (FAsD-V2). *Patient,*. doi:10.1007/s40271-014-0107-7.
59. Fisk, J. D., Ritvo, P. G., Ross, L., Haase, D. A., Marrie, T. J., & Schlech, W. F. (1994). Measuring the functional impact of fatigue: initial validation of the fatigue impact scale. *Clinical Infectious Diseases, 18*(Suppl 1), S79–S83.
60. Prince, M. I., James, O. F., Holland, N. P., & Jones, D. E. (2000). Validation of a fatigue impact score in primary biliary cirrhosis: Towards a standard for clinical and trial use. *Journal of Hepatology, 32*(3), 368–373.
61. Mathiowetz, V. (2003). Test-retest reliability and convergent validity of the Fatigue Impact Scale for persons with multiple sclerosis. *American Journal of Occupational Therapy, 57*(4), 389–395.
62. Imam, B., Anton, H. A., & Miller, W. C. (2012). Measurement properties of a telephone version of the Modified Fatigue Impact Scale among individuals with a traumatic spinal cord injury. *Spinal Cord, 50*(12), 920–924. doi:10.1038/sc.2012.79.
63. Cella, D. F., Tulsky, D. S., Gray, G., Sarafian, B., Linn, E., Bonomi, A., et al. (1993). The Functional Assessment of Cancer Therapy scale: Development and validation of the general measure. *Journal of Clinical Oncology, 11*(3), 570–579.
64. Schiehser, D. M., Ayers, C. R., Liu, L., Lessig, S., Song, D. S., & Filoteo, J. V. (2012). Validation of the Modified Fatigue Impact Scale in Parkinson's disease. *Parkinsonism Related Disorders,*. doi:10.1016/j.parkreldis.2012.11.013.

65. Rietberg, M. B., Van Wegen, E. E., & Kwakkel, G. (2010). Measuring fatigue in patients with multiple sclerosis: Reproducibility, responsiveness and concurrent validity of three Dutch self-report questionnaires. *Disability and Rehabilitation, 32*(22), 1870–1876. doi:10.3109/09638281003734458.
66. Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology, 64*(5), 507–516. doi:10.1016/j.jclinepi.2010.11.018.
67. Broderick, J. E., Schneider, S., Junghaenel, D. U., Schwartz, J. E., & Stone, A. A. (2013). Validity and reliability of patient-reported outcomes measurement information system instruments in osteoarthritis. *Arthritis Care & Research, 65*(10), 1625–1633. doi:10.1002/acr.22025.
68. Senders, A., Hanes, D., Bourdette, D., Whitham, R., & Shinto, L. (2014). Reducing survey burden: Feasibility and validity of PROMIS measures in multiple sclerosis. *Multiple Sclerosis, 20*(8), 1102–1111. doi:10.1177/1352458513517279.
69. Aynehchi, B. B., Obourn, C., Sundaram, K., Bentsianov, B. L., & Rosenfeld, R. M. (2013). Validation of the Modified Brief Fatigue Inventory in head and neck cancer patients. *Otolaryngology Head & Neck Surgery, 148*(1), 69–74. doi:10.1177/0194599812460985.
70. Oude Voshaar, M. A., Ten Klooster, P. M., Bode, C., Vonkeman, H. E., Glas, C. A., Jansen, T., et al. (2015). Assessment of fatigue in rheumatoid arthritis: A psychometric comparison of single-item, multiitem, and multidimensional measures. *Journal of Rheumatology,*. doi:10.3899/jrheum.140389.
71. Nikolaus, S., Bode, C., Taal, E., & van der Laar, M. A. (2012). Expert evaluations of fatigue questionnaires used in rheumatoid arthritis: A Delphi study among patients, nurses and rheumatologists in the Netherlands. *Clinical and Experimental Rheumatology, 30*(1), 79–84.
72. Wu, C. W., Liu, Z. D., Zhang, Y. B., Li, J. M., & Wang, D. X. (2008). Validity and reliability of Chinese version of Fatigue Impact Scale in cerebral infarction patients. *Neural Regeneration Research, 3*(2), 177–181.
73. Theander, K., Cliffordson, C., Torstensson, O., Jakobsson, P., & Unosson, M. (2007). Fatigue Impact Scale: Its validation in patients with chronic obstructive pulmonary disease. *Psychol Health Medicine, 12*(4), 470–484. doi:10.1080/13548500601086771.
74. Harel, D., Thombs, B. D., Hudson, M., Baron, M., & Steele, R. (2012). Measuring fatigue in SSc: A comparison of the Short Form-36 Vitality subscale and Functional Assessment of Chronic Illness Therapy-Fatigue scale. *Rheumatology, 51*(12), 2177–2185. doi:10.1093/rheumatology/kes206.
75. Cella, D., Yount, S., Sorensen, M., Chartash, E., Sengupta, N., & Grober, J. (2005). Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. *Journal of Rheumatology, 32*(5), 811–819.
76. Hagell, P., Hoglund, A., Reimer, J., Eriksson, B., Knutsson, I., Widner, H., et al. (2006). Measuring fatigue in Parkinson's disease: A psychometric study of two brief generic fatigue questionnaires. *Journal of Pain and Symptom Management, 32*(5), 420–432. doi:10.1016/j.jpainsymman.2006.05.021.
77. Salsman, J. M., Beaumont, J. L., Wortman, K., Yan, Y., Friend, J., & Cella, D. (2014). Brief versions of the FACIT-fatigue and FAACT subscales for patients with non-small cell lung cancer cachexia. *Supportive Care in Cancer,*. doi:10.1007/s00520-014-2484-9.
78. 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. *Rehabiliation Psychology, 57*(2), 159–166. doi:10.1037/a0027890.
79. 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), S20–S27. doi:10.1016/j.apmr.2010.08.033.
80. Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E, Jr. (2014). Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. *Journal of Clinical Epidemiology, 67*(1), 108–113. doi:10.1016/j.jclinepi.2013.07.016.
81. Christodoulou, C., Schneider, S., Junghaenel, D. U., Broderick, J. E., & Stone, A. A. (2014). Measuring daily fatigue using a brief scale adapted from the Patient-Reported Outcomes Measurement Information System (PROMIS). *Quality of Life Research, 23*(4), 1245–1253. doi:10.1007/s11136-013-0553-z.
82. Kosinski, M., Keller, S. D., Ware, J. E, Jr, Hatoum, H. T., & Kong, S. X. (1999). The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: relative validity of scales in relation to clinical measures of arthritis severity. *Medical Care, 37*(5), MS23–MS39.
83. Lam, C. L., Gandek, B., Ren, X. S., & Chan, M. S. (1998). Tests of scaling assumptions and construct validity of the Chinese (HK) version of the SF-36 Health Survey. *Journal of Clinical Epidemiology, 51*(11), 1139–1147.
84. Lim, L. L. Y., Seubsman, S.-A., & Sleigh, A. (2008). Thai SF-36 health survey: tests of data quality, scaling assumptions, reliability and validity in healthy men and women. *Health & Quality of Life Outcomes, 6*, 52. doi:10.1186/1477-7525-6-52.
85. Loge, J. H., Kaasa, S., Hjermstad, M. J., & Kvien, T. K. (1998). Translation and performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. I. Data quality, scaling assumptions, reliability, and construct validity. *Journal of Clinical Epidemiology, 51*(11), 1069–1076.
86. McHorney, C. A., Ware, J. E, Jr, Lu, J. F., & Sherbourne, C. D. (1994). The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care, 32*(1), 40–66.
87. Montazeri, A., Goshtasebi, A., Vahdaninia, M., & Gandek, B. (2005). The Short Form Health Survey (SF-36): Translation and validation study of the Iranian version. *Quality of Life Research, 14*(3), 875–882.
88. Sabbah, I., Drouby, N., Sabbah, S., Retel-Rude, N., & Mercier, M. (2003). Quality of life in rural and urban populations in Lebanon using SF-36 health survey. *Health & Quality of Life Outcomes, 1*, 30. doi:10.1186/1477-7525-1-30.
89. Sullivan, M., Karlsson, J., & Ware, J. E, Jr. (1995). The Swedish SF-36 Health Survey—I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. *Social Science and Medicine, 41*(10), 1349–1358.
90. Ren, X. S., Amick, B, 3rd, Zhou, L., & Gandek, B. (1998). Translation and psychometric evaluation of a Chinese version of the SF-36 Health Survey in the United States. *Journal of Clinical Epidemiology, 51*(11), 1129–1138.
91. Chang, D. F., Chun, C. A., Takeuchi, D. T., & Shen, H. (2000). SF-36 health survey: Tests of data quality, scaling assumptions, and reliability in a community sample of Chinese Americans. *Medical Care, 38*(5), 542–548.
92. Li, L., Wang, H. M., & Shen, Y. (2003). Chinese SF-36 Health Survey: Translation, cultural adaptation, validation, and normalisation. *Journal of Epidemiology and Community Health, 57*(4), 259–263.

93. Tseng, H.-M., Lu, J.-F. R., & Gandek, B. (2003). Cultural issues in using the SF-36 Health Survey in Asia: Results from Taiwan. *Health & Quality of Life Outcomes, 1*, 72.
94. Gandek, B., Ware, J. E, Jr, Aaronson, N. K., Alonso, J., Apolone, G., Bjorner, J., et al. (1998). Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: Results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology, 51*(11), 1149–1158.