Qual Life Res (2012) 21:925–944
DOI 10.1007/s11136-011-0009-2

REVIEW

Self-report fatigue questionnaires in multiple sclerosis, Parkinson's disease and stroke: a systematic review of measurement properties

Roy G. Elbers · Marc B. Rietberg · Erwin E. H. van Wegen · John Verhoef · Sharon F. Kramer · Caroline B. Terwee · Gert Kwakkel

Accepted: 2 September 2011 / Published online: 20 October 2011

Abstract

Purpose To critically appraise, compare and summarize the measurement properties of self-report fatigue questionnaires validated in patients with multiple sclerosis (MS), Parkinson's disease (PD) or stroke.

Methods MEDLINE, EMBASE, PsycINFO, CINAHL and SPORTdiscus were searched. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist was used to assess the methodological quality of studies. A qualitative data synthesis was performed to rate the measurement properties for each questionnaire.

Results Thirty-eight studies out of 5,336 records met the inclusion criteria, evaluating 31 questionnaires. Moderate evidence was found for adequate internal consistency and structural validity of the Fatigue Scale for Motor and Cognitive functions (FSMC) and for adequate reliability and structural validity of the Unidimensional Fatigue Impact Scale (U-FIS) in MS.

Conclusions We recommend the FSMC and U-FIS in MS. The Functional Assessment of Chronic Illness Therapy Fatigue subscale (FACIT-F) and Fatigue Severity Scale (FSS) show promise in PD, and the Profile of Mood States Fatigue subscale (POMS-F) for stroke. Future studies should focus on measurement error, responsiveness and interpretability. Studies should also put emphasis on providing input for the theoretical construct of fatigue, allowing the development of questionnaires that reflect generic and disease-specific symptoms of fatigue.

Keywords Fatigue · Assessment · Psychometrics · Multiple sclerosis · Parkinson disease · Cerebrovascular disorders

Abbreviations

AUC	Area under the receiver operator characteristic curve
CC	Correlation coefficient
CIS-20R	Checklist individual strength
CTT	Classical test theory
COSMIN	Consensus-based standards for the selection of health measurement instruments
D-FIS	Fatigue impact scale for daily use
DIF	Differential item functioning

Electronic supplementary material The online version of this article (doi:10.1007/s11136-011-0009-2) contains supplementary material, which is available to authorized users.

R. G. Elbers (✉) · J. Verhoef
Department of Physiotherapy, University of Applied Sciences
Leiden, Zernikedreef 11, PO Box 382, 2300 AJ Leiden,
The Netherlands
e-mail: elbers.r@hsleiden.nl

R. G. Elbers · M. B. Rietberg · E. E. H. van Wegen ·
G. Kwakkel
Department of Rehabilitation Medicine, Research Institute
MOVE, VU University Medical Center, De Boelelaan 1117,
1081 HV Amsterdam, The Netherlands

S. F. Kramer
Department of Clinical Epidemiology and Biostatistics,
The Dutch Cochrane Centre, Academic Medical Center,
Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

C. B. Terwee
Department of Epidemiology and Biostatistics, EMGO Institute
for Health and Care Research, VU University Medical Center,
Van der Boechorststraat 7, 1081 BT Amsterdam,
The Netherlands

EDSS	Expanded disability status scale
EMIF-SEP	Adapted French version of fatigue impact scale
FACIT-F	Functional assessment of chronic illness therapy fatigue subscale
FAI	Fatigue assessment instrument
FAS	Fatigue assessment scale
FIS	Fatigue impact scale
FSMC	Fatigue scale for motor and cognitive functions
FSS	Fatigue severity scale
FSS-7	Fatigue severity scale 7 item version
FSS-5	Fatigue severity scale 5 item version
HR-PRO	Health-related patient-reported outcomes
ICC	Intraclass correlation coefficient
IQR	Interquartile range
IRT	Item response theory
LOA	Limits of agreement
MFI	Multidimensional fatigue inventory
MFIS	Modified fatigue impact scale
MFIS C-5/MFIS P-8	Modified fatigue impact scale cognitive and physical
MFSI-G	Multidimensional fatigue symptom inventory general subscale
MFSS	Multiple sclerosis-specific fatigue severity scale
MIC	Minimal important change
MS	Multiple sclerosis
NFI-MS	Neurological fatigue index for multiple sclerosis
NHP-E	Nottingham health profile energy subscale
PD	Parkinson's disease
PFS-16 (2)	Parkinson fatigue Scale 2-point scale version
PFS-16 (5)	Parkinson fatigue scale 5-point scale version
POMS-F	Profile of mood states fatigue subscale
PROMIS	Patient-reported outcomes measurement information system
PS-F	Performance scale fatigue subscale
RFS	Rhoten fatigue scale
S&E	Schwab and England score
SA-SIP-30	Stroke-adapted sickness impact profile 30 item version
SD	Standard deviation
SDC	Smallest detectable change
SF-36-V	Short-form-36 vitality subscale
SF-36-V (V2.0)	Short-form-36 vitality subscale version 2.0
SOFI	Swedish occupational fatigue inventory
U-FIS	Unidimensional fatigue impact scale
VAS-1, 2 or 3	Visual analogue scale-1, 2 or 3
WEIMUS	Würzburger Erschöpfungsinventars bei Multiple sclerosis

Introduction

Fatigue is common in chronic neurological disorders [1]. Prevalence rates in conditions often seen in neurological rehabilitation, such as multiple sclerosis (MS), Parkinson's disease (PD) and stroke, range from 58% [2] to 90% [3].

One of the challenges in assessing fatigue is the lack of a widely accepted definition [4] and with that, differentiating its many dimensions [2, 5]. Fatigue usually refers to the difficulty initiating or sustaining voluntary activity [6]. Its multidimensionality is believed to result from a complex interplay between the underlying disease process, peripheral control systems (i.e. muscle fatigability), central control systems (i.e. subjective sense of fatigue) and environmental factors [6]. This may reflect the large number of generic and disease-specific self-report questionnaires that are available to measure fatigue as either a multidimensional or a unidimensional assessment in patients considered for rehabilitation services. These questionnaires may measure different aspects or even different theoretical constructs of fatigue [7]. The clinician or researcher has to consider that each questionnaire is characterized by its own underlying concept, measurement properties and practical feasibility. A systematic review of the characteristics and measurement properties of self-report fatigue questionnaires can assist in selecting an appropriate questionnaire to evaluate fatigue in patients with MS, PD and stroke.

Several systematic reviews [7–13] have evaluated the measurement properties of fatigue questionnaires. Three of these reviews [7, 12, 13] focused on patients with chronic disease, including samples of patients with MS and PD. Unfortunately, no recommendations were made specifically for patients with MS or PD. One review [10] focused on patients with MS. The authors recommended the Fatigue Impact Scale (FIS) and the Modified Fatigue Impact Scale (MFIS) [10]. Another review [8] recommended the Multidimensional Fatigue Inventory (MFI) and the Fatigue Severity Scale (FSS) for patients with PD. No systematic review evaluated questionnaires validated in patients with stroke.

A limitation of the aforementioned reviews is that no uniform definitions and standards for the assessment of the methodological quality of the included studies were used. Therefore, the methodological quality of these studies was not taken into account when formulating conclusions, which makes it difficult to judge the strength of the evidence underlying the formulated recommendations. Recently, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist [14] was developed to systematically evaluate the methodological quality of studies on measurement properties. This makes it possible to appraise the methodological quality of the included studies and take this into account when formulating conclusions.

The aim of the present study was to critically appraise, compare and summarize the quality of the measurement properties of all published self-report fatigue questionnaires validated in patients with MS, PD or stroke, in order to assist clinicians and researchers in selecting a fatigue questionnaire.

Methods

Search

Five databases were searched up to November 2010 (MEDLINE (1966–2010), EMBASE (1974–2010), PsycINFO (1806–2010), CINAHL (1981–2010) and SPORTdiscus (1985–2010)). Text words and MESH terms for fatigue, MS, PD and stroke were combined with a sensitive filter (designed for PubMed) to identify studies on measurement properties of self-report questionnaires [15] (see supplementary file 1). References of the included studies were screened for additional articles.

Selection of studies

Two reviewers (RE/EvW) independently screened all titles and abstracts. The full text papers of relevant studies were obtained, and two reviewers (RE/MR) independently applied the a priori defined criteria for study selection. Studies were included if they met the following criteria: the study (1) focused on the development or evaluation of measurement properties of self-report questionnaires that assess subjective fatigue; (2) included patients with a clinical diagnosis of MS, PD or stroke and (3) included questionnaires that could be used for evaluative purposes. Studies were excluded if: the study (1) explicitly focused on the diagnostic test accuracy of the included questionnaire(s); (2) was published in a language other than Dutch, English, French or German. In case of disagreement, a third reviewer (EvW) was asked for advice to reach consensus.

Assessment of methodological quality

The methodological quality of a study was evaluated using the COSMIN checklist [14]. This checklist consists of 114 items, grouped in twelve boxes. Nine of these boxes contain standards for measurement properties (i.e. internal consistency, reliability, measurement error, content validity, structural validity, hypotheses testing, cross-cultural validity, criterion validity and responsiveness). One box contains standards for studies on interpretability, which is an important characteristic of a measurement scale [16]. In addition, two boxes contain requirements for studies in which Item Response Theory (IRT) methods are applied, and requirements for the generalizability of the results, respectively [14]. Each item was scored on a 4-point rating scale (i.e. ‘poor’, ‘fair’, ‘good’, or ‘excellent’) [17]. The methodological quality of a study was evaluated per measurement property and determined by the lowest rating of any of the items in a box. Pairs of reviewers (RE/EvW, RE/JV, RE/MR or RE/SK) independently scored the methodological quality of the included studies. Disagreement was resolved during consensus meetings.

Data extraction

A data extraction form was designed and tested before the pairs of reviewers independently extracted data on the: (1) characteristics of the study samples; (2) characteristics of the questionnaires (i.e. language version, theoretical construct of fatigue and dimensions, recall period, number of items, response options, range of scores, time to administer and ease of scoring); (3) evaluated measurement properties and (4) the interpretability and generalizability of the results.

Data synthesis

The theoretical construct of fatigue measured by a questionnaire was categorized by either ‘impact of fatigue on daily life’, ‘fatigue severity’ or ‘factors influencing fatigue’. Ease of scoring was categorized as ‘easy’ if items were simply summed, ‘moderate’ if a visual analogue scale (VAS) or simple formula was used, or ‘difficult’ if either a VAS in combination with a formula or a complex formula was used.

Measurement properties were summarized according to the COSMIN taxonomy [16]. For each study, the estimates of the investigated measurement properties were rated as ‘adequate’ (+), ‘not adequate’ (−) or ‘unclear’ (?), based on predefined criteria [18] as described below.

A qualitative data synthesis was performed to determine the overall quality of the measurement properties for each self-report questionnaire by taking into account the:

(1) ratings for each measurement property; (2) consistency of results between studies; (3) methodological quality of studies and (4) the number of studies that investigated the measurement property. The possible overall quality of a measurement property was either ‘adequate’ (+), ‘not adequate’ (−), ‘conflicting’ (±) or ‘unclear’ (?). As shown in Table 1, levels of evidence were defined to express whether the strength of the evidence for the overall quality was, for example, convincing (‘strong’ level of evidence) or unconvincing (‘unknown’ level of evidence) [19].

Criteria for the quality of measurement properties

Reliability

The domain reliability contains three measurement properties: internal consistency, reliability and measurement error [16].

Internal consistency is the degree of the interrelatedness among items, assuming the questionnaire to be unidimensional [16]. Cronbach’s α was considered an acceptable measure of internal consistency and scored adequate if it ranged between 0.70 and 0.95 [18]. If a questionnaire was multidimensional, internal consistency was considered per subscale.

Reliability was defined as the proportion of the total variance in the measurements which is because of ‘true’ differences between patients [16]. The intraclass correlation coefficient (ICC) and weighted kappa are acceptable measures for reliability and considered adequate if they were $\geq$0.70 [18]. If a Pearson or Spearman correlation coefficient (CC) was presented, which do not account for systematic differences between two tests [20], an estimate of $\geq$0.80 was considered adequate.

Measurement error, defined as the systematic and random error of a score that is not attributed to true changes in the construct to be measured [16], was scored adequate if the smallest detectable change (SDC) was smaller than the minimal important change (MIC), or if the MIC was outside the limits of agreement (LOA) [18].

Validity

Validity contains the measurement properties content validity, construct validity and criterion validity [16]. Content validity includes face validity and extends to the degree to which the content of a questionnaire is an adequate reflection of the construct to be measured [16]. It was rated adequate if the target population and experts considered all items in the questionnaire relevant and considered the questionnaire to be complete. Construct validity was defined as the degree to which scores of a questionnaire are consistent with hypothesis, based on the assumption that the instrument validly measures the construct to be measured [16]. Construct validity is divided into structural validity, hypothesis testing and cross-cultural validity. Structural validity, defined as the degree to which scores of a questionnaire are an adequate reflection of the dimensionality of the construct to be measured [16], was scored adequate if factor analysis showed that all factors together explained $\geq$50% of the total variance, or when IRT methods were applied to confirm unidimensionality. Hypothesis testing was scored adequate if the correlation with a questionnaire that assessed fatigue (convergent validity) was $\geq$0.50, or $\geq$75% of the results were in accordance with a priori defined hypotheses, and the correlations with other constructs (divergent validity) were lower than the correlations with fatigue. A score unclear was given if only the correlation with questionnaires measuring another construct than fatigue (divergent validity) was investigated. Cross-cultural validity was defined as the degree to which the performance of the items on a translated or culturally adapted health-related patient-reported outcomes (HR-PRO) instrument is an adequate reflection of the performance of the items of the original version of the HR-PRO instrument [16].

As no gold standard exits for fatigue questionnaires, criterion validity was not evaluated.

Responsiveness

Responsiveness was defined as the ability of a questionnaire to detect change over time in the construct to be measured [16]. Responsiveness refers to the validity of a change score

Table 1 Levels of evidence for the overall quality of a measurement property

Level	Rating	Criteria
Strong	‘Adequate’ or ‘Not adequate’ (+ or −)	Consistent findings in multiple studies of ‘good’ methodological quality OR in one study of ‘excellent’ methodological quality
Moderate	‘Adequate’ or ‘Not adequate’ (+ or −)	Consistent findings in multiple studies of ‘fair’ methodological quality OR in one study of ‘good’ methodological quality
Limited	‘Adequate’ or ‘Not adequate’ (+ or −)	One study of ‘fair’ methodological quality
Conflicting	‘Conflicting’ (±)	Conflicting findings
Unknown	‘Unknown’ (?)	Only studies of ‘poor’ methodological quality

[21] and scored adequate if the change score correlated $\geq$0.50 with the change score of an instrument assessing fatigue, or if $\geq$75% of the results were in accordance with a priori defined hypotheses, or if the area under the receiver operator characteristic curve (AUC) was $\geq$0.70 [18].

Interpretability

Interpretability was defined as the degree to which one can assign qualitative meaning to an instruments' quantitative scores or change in scores. Authors should provide information about clinically relevant differences in scores between subgroups (mean or median with distribution of scores), floor and ceiling effects and the MIC [21]. A floor or ceiling effect was present if >15% of patients achieved the lowest or highest possible score on a questionnaire [18].

Results

Search

The search yielded 5,336 records, of which 56 studies were retrieved in full text for further assessment. This resulted in the exclusion of another 18 studies [10, 22–38] (see Fig. 1). Thirty-eight studies were included in the review, investigating 31 different self-report fatigue questionnaires [3, 39–75]. The FSS was most frequently investigated ($n = 20$) and the only questionnaire validated in patients with MS, PD and stroke. Characteristics of the included studies are presented in Table 2.

Characteristics of questionnaires

Table 3 presents the characteristics of the included self-report questionnaires. Most questionnaires aimed to assess the impact of fatigue on activities in daily life (Fatigue Impact Scale for Daily use (D-FIS), Adapted French version of Fatigue Impact Scale (EMIF-SEP), Fatigue Assessment Scale (FAS), FIS, Fatigue Severity Scale 5 item version (FSS-5), MFI, MFIS, Modified Fatigue Impact Scale Cognitive and Physical (MFIS C-5/MFIS P-8), Parkinson Fatigue Scale 2-point scale version (PFS-16 (2)), Parkinson Fatigue Scale 5-point scale version (PFS-16 (5)), Performance Scale Fatigue subscale (PS-F), Unidimensional Fatigue Impact Scale (U-FIS), Visual Analogue Scale-1, 2 or 3 (VAS-1, VAS-2, VAS-3), Würzburger Erschöpfungsinventars bei Multiple sclerosis (WEIMUS)), whereas six questionnaires focused primarily on fatigue severity (Multidimensional Fatigue Symptom Inventory general subscale (MFSI-G), Profile Of Mood States Fatigue subscale (POMS-F), Rhoten Fatigue Scale (RFS), Short-form-36 Vitality subscale (SF-36-V), Short-form-36 Vitality subscale version 2.0 (SF-36-V (V2.0)), Swedish Occupational Fatigue Inventory (SOFI)).

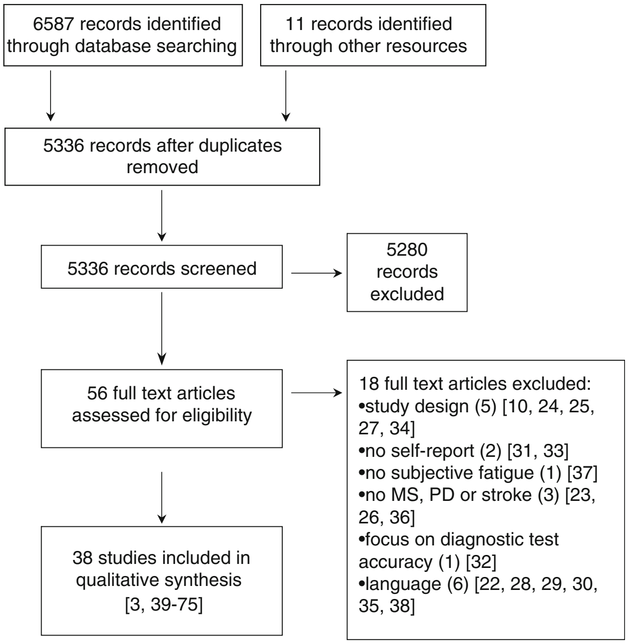
6587 records identified through database searching
11 records identified through other resources
5336 records after duplicates removed
5336 records screened
5280 records excluded
56 full text articles assessed for eligibility
18 full text articles excluded:
•study design (5) [10, 24, 25, 27, 34]
•no self-report (2) [31, 33]
•no subjective fatigue (1) [37]
•no MS, PD or stroke (3) [23, 26, 36]
•focus on diagnostic test accuracy (1) [32]
•language (6) [22, 28, 29, 30, 35, 38]
38 studies included in qualitative synthesis [3, 39-75]

Fig. 1 Flow diagram for study selection

Fifteen unidimensional (D-FIS, Functional Assessment of Chronic Illness Therapy Fatigue subscale (FACIT-F), FAS, FSS, Fatigue Severity Scale 7 item version (FSS-7), FSS-5, MFSI-G, Multiple sclerosis-specific Fatigue Severity Scale (MFSS), Nottingham Health Profile Energy subscale (NHP-E), PFS-16 (2), PFS-16 (5), POMS-F, SF-36-V, SF-36-V (2.0), U-FIS) and eleven multidimensional questionnaires (Checklist Individual Strength (CIS-20R), EMIF-SEP, Fatigue Assessment Instrument (FAI), FIS, Fatigue Scale for Motor and Cognitive functions (FSMC), MFI, MFIS, MFIS C-5/MFIS P-8, Neurological Fatigue Index MS (NFI-MS), SOFI, WEIMUS) were identified. The total number of items per questionnaire varied from 3 (NHP-E) to 40 (EMIF-SEP, FIS). Three visual analogue scales (VAS-1, VAS-2 and VAS-3) and two single-item Likert scales (PS-F, RFS) were included. Six disease-specific questionnaires were found: the MFSS, NFI-MS, PS-F and WEIMUS for patients with MS and the PFS-16 (2) and PFS-16 (5) for patients with PD.

Most questionnaires were found easy to administer. One questionnaire (EMIF-SEP) uses a complex formula to calculate an adjusted total score from 0 to 100, and for two questionnaires (FSS-5, NFI-MS), a nomogram was provided [65, 66] for ordinal-interval (Rasch) transformation. None of the included studies reported on the time needed to complete the questionnaires.

Measurement properties and methodological quality

Details about the investigated measurement properties and the methodological quality of the included studies are

Table 2 Characteristics of included studies

References	Patient characteristics					Questionnaire	
	Population	N	Age Years Mean (SD)	Disease duration Years Mean (SD)	Disease severity EDSS/S&E/SA-SIP-30 Median (IQR)	Investigated	Language version
Armutlu [39]	MS	72	38.16 (10.03)	9.5 (6.43)	EDSS 4.0 (1.0–9.5)[a]	FSS	Turkish
Armutlu [40]	MS	71	38.6 (9.9)	9.42 (6.39)	EDSS 3.94 (1.0–9.5)[a]	FIS	Turkish
Benito-León [41]	MS	68	37.0 (9.0)	6.0 (4.0–10.0)[b]	EDSS 2.5 (2.0–4.0)	D-FIS MFI	Spanish
Brown [42]	PD	39–495[c]	64.2 (9.6)–70.4 (9.5)[c]	10.0 (7.6)–7.9 (6.7)[c]	S&E 66.4 (23.0)–70.3 (15.5)[c]	PFS-16 (2) PFS-16 (5) RFS	English
Debouverie [3]	MS	237	42.5 (10.9)	9.8 (7.4)	EDSS 3.7 (1.7)[d]	EMIF-SEP FIS	French
Doward [43]	MS	9–167[c]	39.0 (12.9)–54.3 (5.9)[c]	8.4 (11.6)–22.7 (13.7)[c]	Not reported	NHP-E U-FIS	Canadian-English Canadian-French French German Italian Swedish US-English
Fisk [44]	MS	105	42.5 (11.6)	Not reported	Not reported	FIS	English
Flachenecker [45]	MS	151	39.0 (9.3)	9.9 (6.7)	EDSS 3.5 (0–8.5)[a]	FSS MFIS MFSS	German
Flachenecker [46]	MS	67–158[c]	39.2 (8.7)–39.2 (9.2)[c]	9.7 (6.8)–9.9 (6.7)	EDSS 3.5 (0–6.5)[a]–3.5 (0–8.5)[a,c]	FSS MFIS MFSS WEIMUS	German
Flachenecker [47]	MS	25–580[c]	44.1 (11.6)–47.2 (11.0)[c]	11.0 (8.1)–15 (9.5)[c]	EDSS 4.5 (1–8)[a]–5.5 (0–9)[a,c]	FSS MFIS MFSS WEIMUS	German
Flensner [48]	MS	161	47.9 (10.1)[e] 48.0 (11.1)[f]	Not reported	Not reported	FIS	Swedish
Grace [49]	PD	50	71.66 (1.39)	Not reported	Not reported	FSS PFS-16 (5)	English
Hagell [50]	PD	118	63.9 (9.6)	8.4 (5.7)	S&E 90 (80–90)[g]	FACIT-F FSS NHP-E	Swedish
Johansson [51]	MS	219	47.0 (12.0)	14 (10)	EDSS 1.0–3.5: 130[h] 4.0–5.5: 37[h] 6.0–9.5: 52[h]	FSS SOFI	Swedish
Kim [52]	MS	49	47 (25–67)[i]	15.7 (1.3–48.0)[i]	EDSS 3.2 (0–7)[i]	FSS MFIS	English

Table 2 continued

References	Patient characteristics					Questionnaire	
	Population	N	Age Years Mean (SD)	Disease duration Years Mean (SD)	Disease severity EDSS/S&E/ SA-SIP-30 Median (IQR)	Investigated	Language version
Kos [53]	MS	51	51.9 (10.5)	16.6 (8.9)	EDSS 6.5 (3–8.5)[a]	FSS MFIS	Dutch
Kos [54]	MS	30–51[c]	44.6 (11.7)–52.9 (10.5)[c]	11.3 (6.8)–16.6 (8.9)[c]	EDSS 6 (3.5–7.5)–6.5 (3–8.5)[c]	FSS MFIS	Dutch Italian Slovenian Spanish
Kos [55]	MS	62	52 (10.5)	Not reported	EDSS 6.5 (3–8.5)	FSS MFIS VAS-1 VAS-2 VAS-3	Dutch
Krupp [56]	MS	25	44.8 (10)	Not reported	Not reported	FSS	English
Kummer [57]	PD	87	56.9 (10.3)	8.7 (4.9)	S&E 76.7 (14.5)–86.1 (8.7)[c]	PFS-16 (2) PFS-16 (5)	Brazilian-Portuguese
Lerdal [58]	MS	227–368[c]	46.6 (12.4)–49.1 (11.7)[c]	11.4 (8.3)–14.0 (10.4)[c]	Not reported	FSS FSS-7 FSS-5	Norwegian Swedish
Losonczi [59]	MS	111	43.82 (11.62)	11.12 (8.29)	EDSS 1.94 (1.37)[d]	FIS	Hungarian
Marrie [60]	MS	9324	52.3 (10.8)	Not reported	Not reported	FSS MFIS PS-F	English
Martínez–Martín [61]	PD	96	66.7 (9.6)[j]	8 (4–13)[b,j]	S&E 80 (70–90)[j]	D-FIS MFI	Spanish
Mathiowetz [62]	MS	54	50 (31–74)[i]	9.5 (1–34)[i]	Not reported	FIS FSS SF-36-V	English
Mead [63]	Stroke	55	73 (66–81)[b]	23 (10–53)[b,k] 137 (93–217)[b,l]	Not reported	FAS MFSI-G POMS-F SF-36-V (V2.0)	English
Meads [64]	MS	15–135[c]	24–77[m]	0.4–59[m]	Not reported	NHP-E U-FIS	English
Mills [65]	MS	416	45.8 (10.5)	17.0 (9.5)	EDSS 0.0–4.0: 143[h] 4.5–6.5: 126[h] 7.0–7.5: 81[h] 8.0–9.5: 58[h] Unknown: 8[h]	FSS FSS-5	English

Table 2 continued

References	Patient characteristics					Questionnaire	
	Population	N	Age Years Mean (SD)	Disease duration Years Mean (SD)	Disease severity EDSS/S&E/ SA-SIP-30 Median (IQR)	Investigated	Language version
Mills [66]	MS	317–318[c]	46.4 (10.6)–46.8 (11.3)[c]	14.2 (9.4)–16.0 (9.7)[c]	EDSS 0.0–4.0: 214[h] 4.5–6.5: 196[h] 7.0–7.5: 136[h] 8.0–9.5: 80[h] Unknown: 9[h]	NFI-MS	English
Mills [67]	MS	415	Not reported	Not reported	Not reported	MFIS MFIS C-5/MFIS P-8	English
Penner [68]	MS	309	43.4 (9.95)	Not reported	EDSS 3.4 (1.63)[d]	FSMC FSS MFIS	Not reported
Rendas–Baum [69]	MS	184	50.9 (10.5)	Not reported	EDSS 6 (0–9)[a]	FIS	Not reported
Reske [70]	MS	20	39.1[n]	9.0 (9.3)	EDSS 3.2 (1.9)[d]	FSS	German
Rietberg [71]	MS	43	48.7 (7.0)	14.3 (9.2)	EDSS 3.5 (1–6.5)[a]	CIS-20R FSS MFIS	Dutch
Schwartz [72]	MS	40	Not reported	Not reported	Not reported	FAI SF-36-V	English
Smith [73]	Stroke	80	74.1 (6.6)	7.6 (5.4)[o]	SA-SIP-30 72.8 (31.5)[p] 77.9 (26.0)[q] 82.1 (29.0)[r] 36.3 (30.6)[s]	FAS	Dutch
Twiss [74]	MS	911	36.5 (8.4)	4.8 (5.2)	EDSS 0.0–1.5: 400[h] 2.0–2.5: 262[h] 3.0–3.5: 135[h] >4: 105[h] Unknown: [h]9	U-FIS	Australian-English Canadian-English Canadian-French French German Italian Spanish UK-English US-English

Table 2 continued

References	Patient characteristics					Questionnaire	
	Population	N	Age Years Mean (SD)	Disease duration Years Mean (SD)	Disease severity EDSS/S&E/ SA-SIP-30 Median (IQR)	Investigated	Language version
Valko [75]	MS	188	45.0 (13.0)	11.07 (9.79)	EDSS 3.61 (2.26)[d]	FSS	German
	Stroke	235	63 (14)	1.21 (0.62)	Not reported		

[a] Expressed as median (Range)

[b] Expressed as median (IQR)

[c] Range of different (sub)samples

[d] Expressed as mean (SD)

[e] Female

[f] Male

[g] During ‘off’ phase

[h] Expressed as numbers: EDSS categorized scores

[i] Expressed as mean (Range)

[j] Based on a total sample of $N = 142$

[k] Inpatients, expressed in days

[l] Outpatients, expressed in days

[m] Range

[n] SD Not reported

[o] Expressed in months

[p] Expressed as percentage of total score body care and movement subscale

[q] Expressed as percentage of total score mobility subscale

[r] Expressed as percentage of total score ambulation subscale

[s] Expressed as percentage of total score alertness behaviour subscale

summarized in Table 4. Most studies investigated reliability and construct validity, whereas results on measurement error and responsiveness were often not reported.

Eight out of 31 studies that investigated hypothesis testing [41, 43, 50, 51, 61, 62, 64, 66] formulated a priori hypothesis about the expected direction or magnitude of the correlation between the investigated questionnaires. Seven studies [39, 40, 54, 59, 61, 70, 75] that translated a questionnaire scored poor methodological quality because the translated questionnaires were not pre-tested in a small sample to check interpretation, cultural relevance and ease of comprehension of the translation.

All studies [53, 56, 69, 71, 74] that reported on responsiveness scored poor methodological quality.

Overall quality of measurement properties

Table 5 presents the overall quality of the measurement properties per self-report questionnaire, accompanied by the level of evidence.

Reliability

The EMIF-SEP and FSMC showed moderate evidence for adequate internal consistency in patients with MS (Cronbach’s $\alpha = 0.82$–0.93) [3, 68] and the D-FIS in patients with PD (Cronbach’s $\alpha = 0.93$) [61]. Limited evidence for adequate internal consistency was found for the D-FIS and FSS in patients with MS (Cronbach’s $\alpha = 0.91$–0.93) [41, 46], the FACIT-F and FSS in patients with PD (Cronbach’s $\alpha = 0.90$–0.94) [49, 50], and the MFSI-G, POMS-F and SF-36-V (V2.0) in patients with stroke (Cronbach’s $\alpha = 0.76$–0.93) [63].

Moderate evidence was found for adequate reliability for the FSS, MFIS and U-FIS in patients with MS (CC or ICC $= 0.73$–0.93) [39, 43, 52, 54, 64, 71]. Limited evidence for adequate reliability was found for the FAS, MFSI-G and POMS-F in patients with stroke (ICC $= 0.74$–0.77) [63] and the FACIT-F in patients with PD (ICC $= 0.84$–0.85) [50]. Reliability of the PFS-16 (5) was found not adequate (limited evidence, CC $= 0.63$) [42].

Table 3 Characteristics of included questionnaires

Questionnaire	Construct assessed	Recall period	Dimensions (number of items)	Response options (range)	Range of scores	Time to administer	Ease of scoring
CIS-20R	Impact of fatigue Fatigue severity	Last 2 weeks	Subjective experience of fatigue (8) Reduction in motivation (4) Reduction in activity (3) Reduction in concentration (5) Total (20)	7-point Likert (1–7)	20–140 (Best–worst)	Not reported	Easy
D-FIS	Impact of fatigue	Last day	One dimension Total (8)	5-point Likert (0–4)	0–32 (Best–worst)	Not reported	Easy
EMIF-SEP	Impact of fatigue	Last month	Cognitive (10) Physical (13) Psychological (4) Social (13) Total (40)	4-point Likert (1–4)	0–100[a] (Best–worst)	Not reported	Difficult[a]
FACIT-F	Impact of fatigue Fatigue severity	Last week	One dimension Total (13)	5-point Likert (0–4)	0–52 (Worst–best)	Not reported	Easy
FAI	Impact of fatigue Fatigue severity	Last 2 weeks	Psychological consequences[b] Severity[b] Situation—specific[b] Response to rest[b] Total (29)	7-point Likert (1–7)	29–203 (Best–worst)	Not reported	Easy
FAS	Impact of fatigue	Usually…	One dimension Total (10)	5-point Likert (1–5)	10–50 (Best–worst)	Not reported	Easy
FIS	Impact of fatigue	Last month	Cognitive (10) Physical (10) Social (20) Total (40)	5-point Likert (0–4)	0–160 (Best–worst)	Not reported	Easy
FSMC	Impact of fatigue Fatigue severity Factors influencing fatigue	In general…	Cognitive (10) Motor (10) Total (20)	5-point Likert (1–5)	20–100 (Best–worst)	Not reported	Easy
FSS	Impact of fatigue Fatigue severity	Not specified	One dimension Total (9)	7-point Likert (1–7)	1–7[c] (Best–worst)	Not reported	Moderate[c]
FSS-7	Impact of fatigue Fatigue severity	Not specified	One dimension Total (7)	7-point Likert (1–7)	1–7[c] (Best–worst)	Not reported	Moderate[c]
FSS-5	Impact of fatigue	Not specified	One dimension Total (5)	7-point Likert (1–7)	0–100[d] (Best–worst)	Not reported	Moderate[d] Easy[e]

Table 3 continued

Questionnaire	Construct assessed	Recall period	Dimensions (number of items)	Response options (range)	Range of scores	Time to administer	Ease of scoring
MFI	Impact of fatigue	Lately…	General (4) Physical (4) Reduced activity (4) Reduced motivation (4) Mental (4) Total (20)	5-point Likert (1–5)	20–100 (Best–worst)	Not reported	Easy
MFIS	Impact of fatigue	Last month	Cognitive (10) Physical (9) Social (2) Total (21)	5-point Likert (0–4)	0–84 (Best–worst)	Not reported	Easy
MFIS C-5/ MFIS P-8	Impact of fatigue	Last month	Cognitive (5) Physical (8) Total (13)	5-point Likert (0–4)	0–52 (Best–worst)	Not reported	Easy
MFSI-G	Fatigue severity	Last week	One dimension Total (6)	5-point Likert (0–4)	0–24 (Best–worst)	Not reported	Easy
MFSS	Factors influencing fatigue	Not specified	One dimension Total (6)	7-point Likert (1–7)	1–7[c] (Best–worst)	Not reported	Moderate[c]
NFI-MS	Fatigue severity Factors influencing fatigue	Last 2 weeks	Abnormal nocturnal sleep (5) Cognitive (4) Physical (8) Relief by rest (6) Summary scale (10) Total (33)	4-point Likert (0–3)	0–99[e] (Best–worst)	Not reported	Moderate[d] Easy[e]
NHP-E	Impact of fatigue Fatigue severity	Not specified	One dimensional Total (3)	Adjectival (Weighted score per item)	0–100 (Best–worst)	Not reported	Easy
PFS-16 (2)	Impact of fatigue	Last 2 weeks	One dimension Total (16)	2-point Likert (0–1)	0–16 (Best–worst)	Not reported	Easy
PFS-16 (5)	Impact of fatigue	Last 2 weeks	One dimension Total (16)	5-point Likert (1–5)	1–5[c] (Best–worst)	Not reported	Moderate[c]
POMS-F	Fatigue severity	Last week	One dimension Total (6)	5-point Likert (0–4)	0–24 (Best–worst)	Not reported	Easy
PS-F	Impact of fatigue	Last month	One dimension Total (1)	6-point Likert (0–5)	0–5 (Best–worst)	Not reported	Easy
RFS	Fatigue severity	Last 2 weeks	One dimension Total (1)	11-point Likert (0–10)	0–10 (Best–worst)	Not reported	Easy
SF-36-V	Fatigue severity	Last month	One dimension Total (4)	6-point Likert (1–6)	4–24 (Worst–best)	Not reported	Easy
SF-36-V (V2.0)	Fatigue severity	Last month	One dimension Total (4)	5-point Likert (1–5)	4–20 (Worst–best)	Not reported	Easy

Table 3 continued

Questionnaire	Construct assessed	Recall period	Dimensions (number of items)	Response options (range)	Range of scores	Time to administer	Ease of scoring
SOFI	Fatigue severity	Last 6 months	Lack of energy (4) Lack of motivation (4) Physical discomfort (4) Physical exertion (4) Sleepiness (4) Total (20)	7-point Likert (0–6)	0–30[f] (Best–worst)	Not reported	Moderate[f]
U-FIS	Impact of fatigue	Last week	One dimension Total (22)	4-point Likert (0–3)	0–66 (Best–worst)	Not reported	Easy
VAS-1	Impact of fatigue	Not specified	One dimension Total (1)	100 mm VAS	0–100[g] (Best–worst)	Not reported	Moderate[g]
VAS-2	Impact of fatigue	Not specified	One dimension Total (1)	100 mm VAS	0–100[g] (Best–worst)	Not reported	Moderate[g]
VAS-3	Impact of fatigue	Not specified	One dimension Total (1)	100 mm VAS	0–100[g] (Best–worst)	Not reported	Moderate[g]
WEIMUS	Impact of fatigue	Last 2 weeks	Cognitive (9) Physical (8) Total (17)	5-point Likert (0–4)	0–68 (Best–worst)	Not reported	Easy

[a] Adjusted total score on 0–100 scale

[b] Not reported

[c] Average of total summed items

[d] Ordinal-interval (Rasch) transformation

[e] Summed raw (ordinal) score

[f] Summed total of averaged domain scores

[g] Visual analogue scale

Measurement error was investigated for the CIS-20R, D-FIS, FAS, FSS, MFIS, MFSI-G, POMS-F and SF-36-V (V2.0), but only one study on the D-FIS used in patients with MS [41] reported details about the MIC. There was limited evidence for adequate measurement error of the D-FIS in patients with MS (SEM = 3.18 and MIC = 3.65) [41].

Validity

Content validity was investigated for the FAS, FIS, FSMC, MFSI-G, NFI-MS, PFS-16 (2), PFS-16 (5), POMS-F, SF-36-V (V2.0) and U-FIS. Moderate evidence was found for adequate content validity of the U-FIS in patients with MS [43, 64]. Limited evidence for adequate content validity was found for the FSMC and NFI-MS in patients with MS [66, 68], for the PFS-16 (2) and PFS-16 (5) in patients with PD [42], and for the FAS, MFSI-G, POMS-F and SF-36-V (V2.0) in patients with stroke [63].

Moderate evidence for adequate structural validity was found for the EMIF-SEP, FSMC (% total explained variance = 61.4–61.5) [3, 68] and U-FIS [43] in patients with MS and for the PFS-16 (5) in patients with PD (% total explained variance = 63.2–64.0) [42]. Four studies that applied IRT methods to assess structural validity demonstrated misfits for items in the FSS and MFIS in patients with MS [58, 65, 67] and in the FACIT-F and FSS in patients with PD [50]. Based on these analyses, new versions for the FSS (FSS-7, FSS-5) [58, 65] and for the MFIS (MFIS C-5/MFIS P-8) [67] were introduced.

Moderate evidence for convergent validity was found for the MFIS (CC = 0.54–0.89 with CIS-20R, FSMC, FSS, PS-F, WEIMUS, WEIMUS Cognitive subscale, WEIMUS Physical subscale) [46, 54, 60, 68, 71], U-FIS (CC = 0.48–0.86 with NHP-E) [43, 64] and NHP-E (CC = 0.48–0.86 with U-FIS) [43, 64] in patients with MS, and for the FSS (CC = 0.62–0.84 with FACIT-F, NHP-E, PFS-16 (5)) [49, 50] and PFS-16 (5)

Table 4 Methodological quality and investigated measurement properties per study

Reference	Population	Investigated measurement properties							
		Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity[a]	Responsiveness
Armutlu [39]	MS	Poor	Fair				Fair	Poor	
Armutlu [40]	MS	Poor	Fair				Fair	Poor	
Benito–León [41]	MS	Fair	Fair	Fair			Fair		
Brown [42]	PD	Good	Fair[b] Poor[c]		Fair	Good	Fair		
Debouverie [3]	MS	Good	Fair			Good		Fair	
Doward [43]	MS	Good[d]	Fair		Fair	Good[d]	Fair	Poor	
Fisk [44]	MS	Poor			Poor		Poor		
Flachenecker [45]	MS						Poor		
Flachenecker [46]	MS	Fair[e] Poor[f]	Poor			Fair[g] Poor[h]	Fair		
Flachenecker [47]	MS		Poor				Poor		
Flensner [48]	MS	Poor					Fair	Fair	
Grace [49]	PD	Fair[b] Poor[i]					Fair		
Hagell [50]	PD	Fair	Fair			Fair[j] Good[k]	Fair		
Johansson [51]	MS	Fair				Fair	Fair		
Kim [52]	MS		Fair						
Kos [53]	MS		Poor				Poor	Poor	Poor
Kos [54]	MS	Fair	Fair			Fair	Fair	Poor	
Kos [55]	MS		Fair				Poor		
Krupp [56]	MS	Poor	Poor			Poor	Poor		Poor
Kummer [57]	PD	Fair[b] Poor[c]						Fair	
Lerdal [58]	MS					Good			
Losonci [59]	MS	Poor	Poor				Poor	Poor	
Marrie [60]	MS						Fair		
Martínez-Martín [61]	PD	Good		Poor		Fair	Fair	Poor	
Mathiowetz [62]	MS		Fair				Fair		
Mead [63]	Stroke	Fair	Fair	Fair	Fair		Fair		
Meads [64]	MS	Poor	Fair		Fair	Poor	Fair		
Mills [65]	MS					Good			
Mills [66]	MS		Fair		Fair	Fair	Fair		
Mills [67]	MS					Good			
Penner [68]	MS	Good	Fair		Fair	Good	Fair		
Rendas-Baum [69]	MS								Poor
Reske [70]	MS	Poor	Poor			Poor	Poor	Poor	
Rietberg [71]	MS		Fair	Fair			Fair	Poor	Poor

Table 4 continued

Reference	Population	Investigated measurement properties							
		Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity[a]	Responsiveness
Schwartz [72]	MS	Fair	Fair			Fair	Poor		
Smith [73]	Stroke	Fair	Poor				Fair		
Twiss [74]	MS	Poor					Fair		Poor
Valko [75]	MS Stroke	Poor					Poor	Poor	

[a] Only items for translation scored

[b] PFS-16 (5)

[c] PFS-16 (2)

[d] Based on Swedish subsample

[e] FSS, MFSS

[f] MFIS, WEIMUS

[g] FSS, MFIS, MFSS

[h] WEIMUS

[i] FSS

[j] CTT

[k] IRT

(CC = 0.71–0.84 with FSS, RFS) [42, 49] in patients with PD.

In 13 studies [3, 39, 40, 43, 48, 53, 54, 57, 59, 61, 70, 71, 75], questionnaires were translated. None of these studies investigated cross-cultural validity by means of confirmatory factor analysis or differential item functioning (DIF).

Responsiveness

Five studies [53, 56, 69, 71, 74] reported on responsiveness. None of these studies presented details about the correlation coefficient between change scores in the investigated questionnaires with change in an external anchor. Therefore, responsiveness was scored unknown for these questionnaires.

Interpretability

Clinically relevant differences in scores between subgroups were reported for the FIS [48], FSS [45], U-FIS [43, 64, 74] and WEIMUS [47] in patients with MS, and for the FACIT-F [50], FSS [50] and PFS-16 (5) [57] in patients with PD.

No floor or ceiling effects were found for the D-FIS [41], FSS [53], FSS-7 and FSS-5 [58], MFIS [53, 54], MFIS C-5/MFIS P-8 [67], NFI-MS [66] and U-FIS [74] in patients with MS. The SOFI showed a floor effect in patients with MS (on 12 of the 20 items, more than 25% of patients achieved the lowest possible score) [51]. The D-FIS [61], FACIT-F [50], FSS [50], PFS-16 (5) and PFS-16 (2) [57] showed no floor or ceiling effects in patients with PD.

Values for the MIC were reported for the D-FIS (MIC = 3.65) [41], FIS (MIC = 9.0–24.0) [69] and U-FIS (MIC = 2.4–7.0) [74] in patients with MS.

Discussion

To our knowledge, this review is the first that systematically appraised and summarized the evidence on the measurement properties of self-report fatigue questionnaires validated in patients with MS, PD or stroke, by taking the methodological quality of the included studies into account. Thirty-one questionnaires were evaluated. No multidimensional questionnaires were identified that were adequately validated in patients with PD or stroke. Moderate evidence was found for adequate internal consistency and structural validity of the FSMC and for adequate reliability and structural validity of the U-FIS in patients with MS. Therefore, we recommend the FSMC for the multidimensional, and the U-FIS for the unidimensional assessment of fatigue in patients with MS. The FACIT-F and FSS show promise for the assessment of fatigue in patients with PD, and the POMS-F for patients with stroke. However, reliability and validity should be confirmed in

Table 5 Data synthesis, levels of evidence and overall quality of measurement properties per questionnaire

Questionnaire	Population	Measurement properties							
		Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity	Responsiveness
CIS-20R	MS		+ Limited	? Unknown			− Limited		? Unknown
D-FIS	MS	+ Limited	+ Limited	+ Limited			− Limited	? Unknown	
	PD	+ Moderate		? Unknown		+ Limited	− Limited		
EMIF-SEP	MS	+ Moderate	+ Limited			+ Moderate		? Unknown	
FACIT-F	PD	+ Limited	+ Limited			− Moderate	+ Limited		
FAI	MS	+ Limited	− Limited			− Limited	? Unknown		
FAS	Stroke	± Conflicting	+ Limited	? Unknown	+ Limited		− Limited		
FIS	MS	? Unknown	± Conflicting		? Unknown		− Moderate	? Unknown	? Unknown
FSMC	MS	+ Moderate	+ Limited		+ Limited	+ Moderate	+ Limited		
FSS	MS	+ Limited	+ Moderate			− Strong	± Conflicting	? Unknown	? Unknown
	PD	+ Limited				− Moderate	+ Moderate		
	Stroke	? Unknown					? Unknown	? Unknown	
FSS-7	MS					+ Moderate			
FSS-5	MS					± Conflicting			
MFI	MS						− Limited		
	PD						− Limited		
MFIS	MS	− Limited	+ Moderate	? Unknown		± Conflicting	+ Moderate	? Unknown	? Unknown
MFIS C-5/ MFIS P-8	MS					+ Moderate			
MFSI-G	Stroke	+ Limited	+ Limited	? Unknown	+ Limited		− Limited		
MFSS	MS	− Limited	? Unknown			+ Limited	− Limited		
NFI-MS	MS		+ Limited		+ Limited	+ Limited	− Limited		
NHP-E	MS						+ Moderate		

Table 5 continued

Questionnaire	Population	Measurement properties							
		Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypothesis testing	Cross-cultural validity	Responsiveness
	PD						+ Limited		
PFS-16 (2)	PD	? Unknown	? Unknown		+ Limited			? Unknown	
PFS-16 (5)	PD	− Moderate	− Limited		+ Limited	+ Moderate	+ Moderate	? Unknown	
POMS-F	Stroke	+ Limited	+ Limited	? Unknown	+ Limited		+ Limited		
PS-F	MS	Not applicable				Not applicable	+ Limited		
RFS	PD	Not applicable				Not applicable	+ Limited		
SOFI	MS	− Limited				− Limited	− Limited		
SF-36-V	MS						+ Limited		
SF-36-V (V2.0)	Stroke	+ Limited	− Limited	? Unknown	+ Limited		− Limited		
U-FIS	MS	− Moderate	+ Moderate		+ Moderate	+ Moderate	+ Moderate	? Unknown	? Unknown
VAS-1	MS	Not applicable	− Limited			Not applicable	? Unknown		
VAS-2	MS	Not applicable	− Limited			Not applicable	? Unknown		
VAS-3	MS	Not applicable	− Limited			Not applicable	? Unknown		
WEIMUS	MS	? Unknown	? Unknown			? Unknown	− Limited		

+ Adequate, − Not adequate, ± Conflicting, ? Unknown

high-quality studies on the FACIT-F, FSS and POMS-F in these populations. Above recommendations should be considered with caution, given that studies investigating measurement error, responsiveness and interpretability are lacking. Second, as the level of evidence supporting the overall quality of most measurement properties was limited, future high-quality studies may change our recommendations.

Two reviews [8, 10] recommend on the use of a questionnaire. One review [10] suggested the FIS and MFIS in patients with MS. The other review [8] recommended the FSS for the unidimensional assessment of fatigue in patients with PD. Although not specifically validated in PD, the MFI was recommended for the multidimensional assessment of fatigue in patients with PD [8]. These recommendations are partially in line with our findings. However, taken the methodological quality of the studies included in our systematic review into account, most measurement properties of the FIS showed only unknown level of evidence. In addition, four studies [50, 58, 65, 67] that applied IRT methods to investigate structural validity demonstrated misfits for some items in the FSS and MFIS.

The inconsistent scores for hypothesis testing confirm that different questionnaires measure different aspects or constructs of fatigue. Unfortunately, details on the construct of fatigue measured by a questionnaire were often not reported. Furthermore, factors contributing to fatigue in patients with MS, PD or stroke are still not well known [2, 76, 77]. Translational research, bridging pre-clinical and clinical research [78], focused on physiological and clinical aspects contributing to peripheral and central fatigue [6], may provide input for more clearly defined concepts and dimensions of fatigue. As both fatigue and most clinical aspects contributing to fatigue fluctuate in time, associations between these factors may be more accurately reflected using longitudinal study designs with repeated measures in time [79]. Repeated measurement designs allow the investigation of the longitudinal construct validity of fatigue measures.

For now, we suggest that clinicians assessing fatigue carefully consider whether a questionnaire reflects the most relevant aspects of fatigue of their interest. Furthermore, a comprehensive evaluation of fatigue should be accompanied by the assessment of clinically related factors such as mood and sleep. Acknowledging that each fatigue questionnaire measures different aspects of fatigue, we recommend the simultaneous use of different questionnaires in research.

Interpretability is considered an important characteristic of a measurement scale [16], unfortunately, only a few studies reported details on clinically relevant differences in scores between subgroups [43, 45, 47, 48, 50, 57, 64, 74], floor and ceiling effects [41, 50, 51, 53, 54, 57, 58, 61, 66, 67, 74] and the MIC [41, 69, 74]. This makes it difficult to interpret scores and change scores on a fatigue questionnaire in both clinical practice and research.

Although it is believed that measurement properties are sample dependent [80], no major differences in measurement properties were found for questionnaires that were evaluated in more than one population. For example, all estimates of measurement properties for the D-FIS were consistent in patients with MS and PD. The FSS showed consistent scores for most measurement properties that were evaluated in patients with MS, PD and stroke. In addition, another review [8] concluded that the items of the disease-specific PFS-16 (5) did not differ much from other generic fatigue questionnaires and that it provided no clear advantages above a generic questionnaire for use in patients with PD. Furthermore, it is not clear whether manifestations of fatigue are different between neurological disorders [8]. These results suggest that generic fatigue questionnaires presented in this review can be used interchangeably in patients with MS, PD and stroke and favour a generic approach for the assessment of fatigue. In contrast, studies using IRT methods showed misfits on the FSS for four items in patients with MS [65], and for only one item in patients with PD [50]. This difference might have been caused by a difference in statistical power between both studies [65], but it is also possible that it was related to DIF in patients with MS and PD [65]. This emphasizes the importance of disease-specific validation for fatigue questionnaires used in patients with MS, PD and stroke. Abovementioned findings suggest that self-report fatigue questionnaires should contain a core set of items assessing generic aspects of fatigue, whereas some additional items are more disease specific. We therefore recommend the adaptation of existing questionnaires, incorporating a uniform section on general aspects of fatigue and a section with disease-specific items. Items to assess general aspects of fatigue may be derived from the recently developed Patient-Reported Outcomes Measurement Information System (PROMIS) fatigue item bank [81].

This systematic review has some limitations. First, only studies published in Dutch, English, French or German were included. This language restriction resulted in the exclusion of six articles [22, 28–30, 35, 38]; however, these studies evaluated a diversity of questionnaires and language versions, so it is not likely that this resulted in selection bias. Second, the COSMIN checklist has some items that require subjective judgment, which may lead to disagreement between raters. However, we tested the COSMIN checklist with all reviewers before assessing the methodological quality of the included studies, and one reviewer (RE) was involved in the assessment of all studies to improve consistency in rating across studies. Third, the quality criteria we applied for rating measurement properties heavily weighed on classical test theory (CTT). As a consequence, IRT methods were not considered for underpinning the structural validity of questionnaires. To overcome this incompleteness, we decided, post hoc, that any misfit in a questionnaire displayed by a study using IRT methods was judged as not adequate structural validity.

Conclusion

We recommend the FSMC and U-FIS for the assessment of fatigue in patients with MS. The FACIT-F and FSS show promise in patients with PD, and the POMS-F for patients with stroke. No multidimensional questionnaires were adequately validated in patients with PD or stroke. Future studies should focus on translational research in which assumed underlying physiological and clinical aspects contributing to fatigue are investigated longitudinally, as perceptions of fatigue often show fluctuations in time. Such studies may provide input for the development of the theoretical construct of self-report fatigue questionnaires. We suggest that existing questionnaires should be adapted to contain both a uniform section that reflects general aspects of fatigue, and a disease-specific section that contains items that are related with physiological and clinical aspects of underlying disease. Studies on responsiveness and the MIC of fatigue questionnaires in patients with MS, PD and stroke are needed, to establish whether an instrument can detect meaningful changes in clinical practice and research.

Acknowledgments The authors would like to thank R. Spijker (MSc), medical information specialist and Trial Search Coordinator from the Academic Medical Center and the Dutch Cochrane Centre, Amsterdam, The Netherlands for his valuable support regarding the design of the search strategy. *Financial disclosures:* This research project was supported by a grant from the Royal Dutch Society for Physical Therapy, Scientific College Physical Therapy, project number: BU003-2010, and the Stichting MS Research, project number 04-553 MS.

References

1. De Groot, M., Phillips, S., & Eskes, G. (2003). Fatigue associated with stroke and other neurologic conditions: Implications for stroke rehabilitation. *Archives of Physical Medicine and Rehabilitation, 84*(11), 1714–1720.
2. Friedman, J., Brown, R., Comella, C., Garber, C., Krupp, L., Lou, J., et al. (2007). Fatigue in Parkinson's disease: A review. *Movement Disorders, 22*(3), 297–308.
3. Debouverie, M., Pittion-Vouyovitch, S., Louis, S., & Guillemin, F. (2007). Validity of a French version of the fatigue impact scale in multiple sclerosis. *Multiple Sclerosis, 13*(8), 1026–1032.
4. Lou, J. (2009). Physical and mental fatigue in Parkinson's disease: epidemiology, pathophysiology and treatment. *Drugs and Aging, 26*(3), 195–208.
5. Smets, E., Garssen, B., Bonke, B., & De Haes, J. (1995). The multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychosomatic Research, 39*(3), 315–325.
6. Chaudhuri, A., & Behan, P. (2004). Fatigue in neurological disorders. *Lancet, 363*(9413), 978–988.
7. Dittner, A., Wessely, S., & Brown, R. (2004). The assessment of fatigue: A practical guide for clinicians and researchers. *Journal of Psychosomatic Research, 56*(2), 157–170.
8. Friedman, J., Alves, G., Hagell, P., Marinus, J., Marsh, L., Martínez-Martín, P., et al. (2010). Fatigue rating scales critique and recommendations by the movement disorders society task force on rating scales for Parkinson's disease. *Movement Disorders, 25*(7), 805–822.
9. Hewlett, S., Hehir, M., & Kirwan, J. (2007). Measuring fatigue in rheumatoid arthritis: A systematic review of scales in use. *Arthritis and Rheumatism, 57*(3), 429–439.
10. Kos, D., Kerckhofs, E., Ketelaer, P., Duportail, M., Nagels, G., D'hooghe, M., et al. (2004). Self-report assessment of fatigue in multiple sclerosis: A critical evaluation. *Occupational Therapy in Health Care, 17*(3–4), 45–62.
11. 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.
12. Mota, D., & Pimenta, C. (2006). Self-report instruments for fatigue assessment: A systematic review. *Research and Theory for Nursing Practice, 20*(1), 49–78.
13. 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.
14. Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., 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.
15. Terwee, C., Jansma, E., Riphagen, I., & de Vet, H. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research, 18*(8), 1115–1123.
16. Mokkink, L., Terwee, C., Patrick, D., Alonso, J., Stratford, P., Knol, D., 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.
17. Terwee, C., Mokkink, L., Knol, D., Ostelo, R., Bouter, L., & de Vet, H. (2011). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research.* doi: 10.1007/s11136-011-9960-1.
18. Terwee, C., Bot, S., de Boer, M., van der Windt, D., Knol, D., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42.
19. Van Tulder, M., Furlan, A., Bombardier, C., & Bouter, L. (2003). Updated method guidelines for systematic reviews in the Cochrane collaboration back review group. *Spine, 28*(12), 1290–1299.
20. Streiner, D., & Norman, G. (2003). *Health measurement scales, a practical guide to their development and use*. New York: Oxford University Press.
21. Mokkink, L., Terwee, C., Knol, D., Stratford, P., Alonso, J., Patrick, D., et al. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology, 10*, 22.
22. Balci, K., Armutlu, K., Fil, A., & Karabudak, R. (2008). If a special scale to measure subjective fatigue in multiple sclerosis was needed? *Fizyoterapi Rehabilitasyon, 19*(3), 188.
23. Chang, C., Cella, D., Clarke, S., Heinemann, A., von Roenn, J., & Harvey, R. (2003). Should symptoms be scaled for intensity, frequency, or both? *Palliative & Supportive Care, 1*(1), 51–60.
24. Chipchase, S., Lincoln, N., & Radford, K. (2003). Measuring fatigue in people with multiple sclerosis. *Disability and Rehabilitation, 25*(14), 778–784.
25. Debouverie, M., Pittion-Vouyovitch, S., & Guillemin, F. (2009). Reconsidering fatigue at the onset of multiple sclerosis. *Revue Neurologique (Paris), 165*(Suppl 4), 135–144.
26. Fisk, J., & Doble, S. (2002). Construction and validation of a fatigue impact scale for daily administration (D-FIS). *Quality of Life Research, 11*(3), 263–272.
27. Graham, J., Fisher, N., Granger, C., & Tomita, M. (2007). Test battery for evaluating fatigue in multiple sclerosis patients. *International Journal of MS Care, 9*(3), 118–125.
28. Gruszczak, A., Bartosik-Psujek, H., Pocinska, K., & Stelmasiak, Z. (2009). Validation analysis of selected psychometric features of Polish version of modified fatigue impact scale—preliminary findings. *Neurologia i Neurochirurgia Polska, 43*(2), 148–154.
29. Haase, V., Lacerda, S., de Paula Lima, E., de Deus Corrêa, T., de Brito, D., & Lana-Peixoto, M. (2004). Assessment of psychosocial functioning in multiple sclerosis: Psychometric characteristics of four self-report measures. *Arquivos de Neuro-psiquiatria, 62*(2-A), 282–291.
30. Iriarte, J., & de Castro, P. (1994). Proposal of a new scale for assessing fatigue in patients with multiple sclerosis. *Neurologia, 9*(3), 96–100.
31. Iriarte, J., Katsamakis, G., & de Castro, P. (1999). The fatigue descriptive scale (FDS): A useful tool to evaluate fatigue in multiple sclerosis. *Multiple Sclerosis, 5*(1), 10–16.

32. Jason, L., Ropacki, M., Santoro, N., Richman, J., Heatherly, W., Taylor, R., et al. (1997). A screening instrument for chronic fatigue syndrome: Reliability and validity. *Journal of chronic fatigue syndrome, 3*(1), 39–59.
33. Lynch, J., Mead, G., Greig, C., Young, A., Lewis, S., & Sharpe, M. (2007). Fatigue after stroke: The development and evaluation of a case definition. *Journal of Psychosomatic Research, 63*(5), 539–544.
34. Montreuil, M. (2000). Analysis of subjective complaints of fatigue in patients with multiple sclerosis. *Revue Neurologique (Paris), 156*(11), 1048–1050.
35. Pavan, K., Schmidt, K., Marangoni, B., Fernanda Mendes, M., Tilbery, C., & Lianza, S. (2007). Multiple sclerosis: cross-cultural adaptation and validation of the modified fatigue impact scale. *Arquivos de Neuro-psiquiatria, 65*(3-A), 669–673.
36. Taylor, R., Jason, L., & Torres, A. (2000). Fatigue rating scales: An empirical comparison. *Psychological Medicine, 30*(4), 849–856.
37. Tseng, B., Gajewski, B., & Kluding, P. (2010). Reliability, responsiveness, and validity of the visual analog fatigue scale to measure exertion fatigue in people with chronic stroke: a preliminary study. *Stroke Research and Treatment*. doi: 10.4061/2010/412964.
38. Wu, C., Liu, Z., Zhang, Y., Li, J., & Wang, D. (2008). Validity and reliability of Chinese version of fatigue impact scale in cerebral infarction patients. *Neural Regeneration Research, 3*(2), 177–181.
39. Armutlu, K., Korkmaz, N., Keser, I., Sümbüloglu, V., Akbiyik, D., Güney, Z., et al. (2007). The validity and reliability of the fatigue severity scale in Turkish multiple sclerosis patients. *International Journal of Rehabilitation Research, 30*(1), 81–85.
40. Armutlu, K., Keser, I., Korkmaz, N., Akbiyik, D., Sümbüloglu, V., Güney, Z., et al. (2007). Psychometric study of Turkish version of fatigue impact scale in multiple sclerosis patients. *Journal of the Neurological Sciences, 255*(1–2), 64–68.
41. Benito-León, J., Martínez-Martín, P., Frades, B., Martínez-Ginés, M., de Andrés, C., Meca-Lallana, J., et al. (2007). Impact of fatigue in multiple sclerosis: The fatigue impact scale for daily use (D-FIS). *Multiple Sclerosis, 13*(5), 645–651.
42. Brown, R., Dittner, A., Findley, L., & Wessely, S. (2005). The Parkinson fatigue scale. *Parkinsonism and Related Disorders, 11*(1), 49–55.
43. Doward, L., Meads, D., Fisk, J., Twiss, J., Hagell, P., Oprandi, N., et al. (2010). International development of the unidimensional fatigue impact scale (U-FIS). *Value in Health, 13*(4), 463–468.
44. Fisk, J., Ritvo, P., Ross, L., Haase, D., Marrie, T., & Schlech, W. (1994). Measuring the functional impact of fatigue: Initial validation of the fatigue impact scale. *Clinical Infectious Diseases, 18*(Suppl 1), 79–83.
45. Flachenecker, P., Kümpfel, T., Kallmann, B., Gottschalk, M., Grauer, O., Rieckmann, P., et al. (2002). Fatigue in multiple sclerosis: A comparison of different rating scales and correlation to clinical parameters. *Multiple Sclerosis, 8*(6), 523–526.
46. Flachenecker, P., Müller, G., König, H., Meissner, H., Toyka, K., & Rieckmann, P. (2006). Fatigue in multiple sclerosis: development and validation of the Würzburger fatigue inventory for MS. *Der Nervenarzt, 77*(2), 165–174.
47. Flachenecker, P., König, H., Meissner, H., Müller, G., & Rieckmann, P. (2008). Fatigue in multiple sclerosis: Validation of the WEIMUS scale ('Würzburger erschöpfungs-inventar bei multipler sklerose'). *Neurologie & Rehabilitation, 14*(6), 299–306.
48. Flensner, G., Ek, A., & Söderhamn, O. (2005). Reliability and validity of the Swedish version of the fatigue impact scale (FIS). *Scandinavian Journal of Occupational Therapy, 12*(4), 170–180.
49. Grace, J., Mendelsohn, A., & Friedman, J. (2007). A comparison of fatigue measures in Parkinson's disease. *Parkinsonism and Related Disorders, 13*(7), 443–445.
50. Hagell, P., Höglund, A., Reimer, J., Erikson, 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.
51. Johansson, S., Ytterberg, C., Back, B., Holmqvist, L., & von Koch, L. (2008). The Swedish occupational fatigue inventory in people with multiple sclerosis. *Journal of Rehabilitation Medicine, 40*(9), 737–743.
52. Kim, E., Lovera, J., Schaben, L., Melara, J., Bourdette, D., & Whitham, R. (2010). Novel method for measurement of fatigue in multiple sclerosis: Real-time digital fatigue score. *Journal of Rehabilitation and Development, 47*(5), 477–484.
53. Kos, D., Kerckhofs, E., Nagels, G., D'hooghe, M., Duquet, W., Duportail, M., et al. (2003). Assessing fatigue in multiple sclerosis: Dutch modified fatigue impact scale. *Acta Neurologica Belgica, 103*(4), 185–191.
54. Kos, D., Kerckhofs, E., Carrea, I., Verza, R., Ramos, M., & Jansa, J. (2005). Evaluation of the modified fatigue impact scale in four different European countries. *Multiple Sclerosis, 11*(1), 76–80.
55. Kos, D., Nagels, G., D'Hooghe, M., Duportail, M., & Kerckhofs, E. (2006). A rapid screening tool for fatigue impact in multiple sclerosis. *BMC Neurology, 6*, 27.
56. Krupp, L., LaRocca, G., Muir-Nash, J., & Steinberg, A. (1989). The fatigue severity scale: Application to patients with multiple sclerosis and systemic lupus erythematosus. *Archives of Neurology, 46*(10), 1121–1123.
57. Kummer, A., Scalzo, P., Cardoso, F., & Teixeira, A. (2010). Evaluation of fatigue in Parkinson's disease using the Brazilian version of Parkinson's fatigue scale. *Acta Neurologica Scandinavica*. doi: 10.1111/j.1600-0404.2010.01364.x.
58. Lerdal, A., Johansson, S., Kottorp, A., & von Koch, L. (2010). Psychometric properties of the fatigue severity scale: Rasch analyses of responses in a Norwegian and a Swedish MS cohort. *Multiple Sclerosis, 16*(6), 733–741.
59. Losonczi, E., Bencsik, K., Radja, C., Lencsés, G., Török, M., & Vécsei, L. (2010). Validation of the fatigue impact scale in Hungarian patients with multiple sclerosis. *Quality of Life Research, 20*(2), 301–306.
60. Marrie, R., Cutter, G., Tyry, T., Hadjimichael, O., Campagnolo, D., & Vollmer, T. (2005). Validation of the NARCOMS registry: Fatigue assessment. *Multiple Sclerosis, 11*(5), 583–584.
61. Martínez-Martín, P., Catalan, M., Benito-León, J., Ortega Moreno, A., Zamarbide, I., Cubo, E., et al. (2006). Impact of fatigue in Parkinson's disease: The fatigue impact scale for daily use (D-FIS). *Quality of Life Research, 15*(4), 597–606.
62. Mathiowetz, V. (2003). Test-retest reliability and convergent validity of the fatigue impact scale for persons with multiple sclerosis. *The American Journal of Occupational Therapy, 57*(4), 389–395.
63. 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.
64. Meads, D., Doward, L., McKenna, S., 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.
65. Mills, R., Young, C., Nicholas, R., Pallant, J., & Tennant, A. (2009). Rash analysis of the fatigue severity scale in multiple sclerosis. *Multiple Sclerosis, 15*(1), 81–87.
66. Mills, R., Young, C., Pallant, J., & Tennant, A. (2010). Development of a patient reported outcome scale for fatigue in multiple

sclerosis: The neurological fatigue index (NFI-MS). *Health and Quality of Life Outcomes, 8*, 22.

67. Mills, R., Young, C., Pallant, J., & 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.
68. Penner, I., Raselli, C., Stöcklin, M., Opwis, K., Kappos, L., & Calabrese, P. (2009). The fatigue scale for motor and cognitive functions (FSMC): Validation of a new instrument to assess multiple sclerosis-related fatigue. *Multiple Sclerosis, 15*(12), 1509–1517.
69. Rendas-Baum, R., Yang, M., Cattelin, F., Wallenstein, G., & Fisk, J. (2010). A novel approach to estimate the minimally important difference for the fatigue impact scale in multiple sclerosis patients. *Quality of Life Research, 19*(9), 1349–1358.
70. Reske, D., Pukrop, R., Scheinig, K., Haupt, W., & Petereit, H. (2006). Measuring fatigue in patients with multiple sclerosis with standardized methods in German-speaking areas. *Fortschritte der Neurologie, Psychiatrie, 74*(9), 497–502.
71. Rietberg, M., van Wegen, 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.
72. Schwartz, J., Jandorf, L., & Krupp, L. (1993). The measurement of fatigue: A new instrument. *Journal of Psychosomatic Research, 37*(7), 753–762.
73. Smith, O., van den Broek, K., Renkens, M., & Denollet, J. (2008). Comparison of fatigue levels in patients with stroke and patients with end-stage heart failure: Application of the fatigue assessment scale. *Journal of the American Geriatrics Society, 56*(10), 1915–1919.
74. Twiss, J., Doward, L., McKenna, S., & Eckert, B. (2010). Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS). *Health and Quality of Life Outcomes, 8*, 117.
75. Valko, P., Bassetti, C., Bloch, K., Held, U., & Baumann, C. (2008). Validation of the fatigue severity scale in a Swiss cohort. *Sleep, 31*(11), 1601–1607.
76. Kos, D., Kerckhofs, E., Nagels, G., D'hooghe, M., & Ilsbroukx, S. (2008). Origin of fatigue in multiple sclerosis: Review of the literature. *Neurorehabilitation and Neural Repair, 22*(1), 91–100.
77. Tseng, B., Billinger, S., Gajewski, B., & Kluding, P. (2010). Exertion fatigue and chronic fatigue are two distinct constructs in people post-stroke. *Stroke, 41*(12), 2908–2912.
78. Kwakkel, G. (2009). Towards integrative neurorehabilitation science. *Physiotherapy Research International, 14*(3), 137–146.
79. Elbers, R., van Wegen, E., Rochester, L., Hetherington, V., Nieuwboer, A., Willems, A., et al. (2009). Is impact of fatigue an independent factor associated with physical activity in patients with idiopathic Parkinson's disease? *Movement Disorders, 24*(10), 1512–1518.
80. McHorney, C., Ware, J., Lu, J., & Sherbourne, C. (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), 40–60.
81. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194.