Neurol Ther (2023) 12:1649–1668
https://doi.org/10.1007/s40120-023-00501-9

ORIGINAL RESEARCH

Measuring Fatigue in Multiple Sclerosis: There may be Trouble Ahead

James Close · Jo Vandercappellen · Miriam King · Jeremy Hobart

Received: November 11, 2022 / Accepted: May 17, 2023 / Published online: June 23, 2023

ABSTRACT

Introduction: Poorly developed patient-reported outcome measures (PROs) risk type-II errors (i.e. false negatives) in clinical trials, resulting in erroneous failure to achieve trial endpoints. Validity is a fundamental requirement of fit-for-purpose PROs, with the main determinant of validity being the PROs items, i.e. content validity. Here, we sought to identify fatigue PRO instruments used in multiple sclerosis (MS) studies and to assess the extent to which their development satisfied current content validity standards.

Methods: We searched Embase® and Medline® for MS studies using fatigue-based PROs. Abstracts were screened, PROs identified, and their relevant development papers assessed against seven Consensus Standards for Measurement Instruments (COSMIN) criteria for content development.

Results: From 3814 abstracts, 18 fatigue PROs met our inclusion criteria. Most PROs did not satisfy at least one COSMIN content validity standard. Frequent omissions during PRO development include: clearly defined constructs; conceptual frameworks; qualitative research in representative samples; and literature reviews. PRO development quality has improved significantly since FDA guidance was published ($U = 10.0$, $p = 0.02$). However, scatterplots and correlations between PRO COSMIN scores and citation frequency (rho $= -0.62$) and clinical trials usage (rho $= +0.18$) implied that PRO quality is unrelated to choice. COSMIN scores implied that the Fatigue Symptoms and Impact Questionnaire—Relapsing Multiple Sclerosis (FSIQ-RMS) and Neurological Fatigue Index—Multiple Sclerosis (NFI-MS) had the strongest evidence for adequate content validity.

Conclusion: Most existing fatigue PROs do not meet COSMIN content validity requirements. Although two PROs scored well on aggregate (NFI-MS and FSIQ-RMS), our subsequent evaluation of the item sets that generated their scores implied that both PROs have weaker content validity than COSMIN suggests. This indicates that COSMIN criteria require further development, and raises significant concerns about how we have measured one of the most common and burdensome MS symptoms. A detailed head-to-head psychometric evaluation is needed to determine the impact of different PRO

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40120-023-00501-9.

J. Close · J. Hobart (✉)
Peninsula Schools of Medicine and Dentistry, University of Plymouth, Plymouth, UK
e-mail: jeremy.hobart@plymouth.ac.uk

J. Vandercappellen · M. King
Novartis Pharma AG, Basel, Switzerland

development qualities and the implications of the problems implied by our analyses, on measurement performance.

PLAIN LANGUAGE SUMMARY

In MS clinical trials, impacts such as fatigue, walking ability, and quality of life, are measured using questionnaires—called patient-reported outcome measures—completed by people living with MS. The quality of these measures is fundamentally important. If poor quality patient-reported outcome measures are used, treatment benefits are easily missed or underestimated.

We studied the quality of 18 fatigue patient-reported outcome measures previously used in MS studies. Specifically, we studied how the questionnaire questions were developed and scored them against recognised quality control standards. In general, the patient-reported outcome measures were poor. Only two scored reasonably well. One common weakness was that people living with MS were not involved during patient-reported outcome measure development. We also conducted novel examinations that went beyond the quality control standards. These test how well the questions relate back to the MS impacts they claim to measure. We found even the two best patient-reported outcome measures were poor.

Our study had two findings. First, patient-reported outcome measures of MS fatigue are poor. Second, current standards for testing patient-reported outcome measure development are too easy to satisfy, overestimate patient-reported outcome measure quality, and need updating. Therefore, the ways we measure MS fatigue, one of the most common and burdensome MS symptoms, are scientifically weak.

Keywords: Multiple sclerosis; Patient reported outcomes; Fatigue; Content validity; Fatigue measurement; Measurement; COSMIN criteria

Key summary points

Weak measurement, from any cause, risks type-II errors (i.e. false negatives)

Content validity limitations are an important source of PRO measurement weakness

We examined how MS fatigue PROs adhere to content validity standards

Quality has improved over time, but the fatigue PROs examined had poor content validity. One likely outcome is type-II errors in clinical trials

COSMIN content validity criteria need to be more specific and stringent

DIGITAL FEATURES

This article is published with digital features, including a video abstract to facilitate understanding of the article. To view digital features for this article go to https://doi.org/10.6084/m9.figshare.22812440.

INTRODUCTION

Patient-reported outcomes (PROs) play an increasingly prominent role in clinical trials, drug approval and reimbursement. Any PRO measurement shortcomings risk type-II errors, threatening treatment development and licensing. The negative implications are pervasive and damaging, not just for trial outcomes and available treatment options for health care providers, but, more importantly, patient well-being is compromised [1].

Over the last decade, scientific and regulatory criteria for PRO development have evolved. There has been a shift from primarily emphasising psychometric properties to incorporating aspects of content validity; i.e. the extent to which PROs adequately reflect a defined

measurement construct [2]. The importance of content validity cannot be underestimated, but, we believe, remains underappreciated and misunderstood.

At face value, the notion of content validity is beguilingly simple: it is the extent to which a set of items fairly represent the construct they purport to measure. However, more careful consideration clarifies content validity's fundamental importance and complexity, and helps to explain why it is cited as the most important PRO measurement property [3], and a pre-requisite for any statistical ("psychometric") examinations [4].

In PRO measurement, the responses to a set of questions (items) are combined to derive a score. This score is intended to quantify a health concept or variable; for example, fatigue. Therefore, the items link the concept to the score. If the score derived from an item set is to be a valid indicator of the concept, the concept must be clearly defined and broken down to its relevant components and subcomponents. This process is known as construct definition, conceptualisation, and conceptual framework development. When this process is not explicit, validity is compromised to an extent that cannot be determined or quantified; in other words, unquantifiable type-II errors are liable to occur.

There is another critical step on the route to achieving content validity that comes after conceptual framework development. It is the articulation of the subcomponents, proposed to be scored, as sets of items. For example, motor, cognitive and psychosocial impacts of fatigue. Again, the subcomponents should be defined and broken down into their parts, so that the link between the items and the subcomponent are explicit. Also, item wording requires careful attention so that it aligns with the subcomponent, the other items in the set, and articulates both the concept and the measurement aspect (e.g. frequency, intensity, severity). Wording should be as unambiguous as possible. When all the steps are carefully attended to, the link from the concept, via its components and subcomponents, through the items to the score, is clear and content validity can be considered achieved.

This description indicates that beguiling simple content validity comprises multiple aspects, which, when incomplete, compromises PRO validity, causing uncertain measurement, with type-II errors as the ultimate result. There are two additional complexities. First, there is no external method of "proving" content validity. Second, achieving content validity does not guarantee that the scores generated by an item set satisfy requirements for measurement. This is the related but independent domain of psychometric (statistical) measurement performance testing, which is only meaningful when content validity is established.

While content validity cannot be proven, guidance aids its achievement and evaluation. The US Food and Drug Administartion (FDA) advise that PRO content validity should be underpinned by well-defined concepts of interests, contexts of use, and conceptualisation, with an item development process involving members of the target population for which the instrument is being designed [4]. The FDA principles are integrated into updated Consensus Standards for Measurement Instruments (COSMIN) PRO development guidance [3, 5]. When these updated guidelines have been applied to PRO content development in diverse clinical contexts [5–10], including upper limb function [5], consistently low-quality content validity evidence has resulted. General standards of PRO content validity have been described as "questionable" [8] and "worrisome" [6, 7].

We aimed to establish the extent to which fatigue PROs used in multiple sclerosis (MS) clinical trials satisfy COSMIN's content validity recommendations. We chose fatigue as this is one of the most common and burdensome symptoms for people living with MS (PLwMS) [11, 12], and there has been limited progress over time in our understanding of it and of its management. This could reflect a measurement problem.

METHODS

Overview

We searched for existing fatigue PRO instruments used in MS studies and assessed these against the COSMIN content validity criteria [3]; specifically, elements related to the conceptual basis of the instruments, namely, the PRO construct, conceptual framework, target population, context of use, development sample, qualitative work, and use of literature reviews in the instrument development process. While our final item (concerning appropriate literature searches) is not explicitly defined within the COSMIN criteria, FDA guidance [4] states that content development can involve both qualitative work and literature reviews. We did not evaluate psychometric criteria, as adequate content validity is a pre-requisite for meaningful psychometric comparisons and interpretations, and poor content development is not negated by a strong psychometric profile.

Literature review

Embase® and Medline® were searched for English language publications up to 20 October 2021. Our search terms were: "multiple sclerosis" AND "fatigue" AND ("instrument" OR "patient reported outcome" OR "patient-reported outcome" OR "questionnaire"). Abstracts were screened to identify fatigue PRO instruments, and relevant PRO development papers retrieved.

Instruments were retained for further assessment if they were (1) MS-specific fatigue PROs or (2) non-disease-specific (i.e. generic) fatigue PROs. We excluded fatigue items embedded within broader instruments and fatigue PROs specific to other diseases. Most PROs had a single associated development paper. When relevant, linked papers were retrieved in line with COSMIN's recommendations for using 'indirect evidence' and 'other additional information' when assessing PRO content validity [3].

Table 1 COSMIN checklist items

COSMIN items
Item 1: Is a clear description provided of the construct to be measured?
Item 2: Is the origin of the construct clear: was a theory, conceptual framework or disease model used or clear rationale provided to define the construct to be measured?
Item 3: Is a clear description provided of the target population for which the PRO measure (PROM) was developed?
Item 4: Is a clear description provided of the context of use?
Item 5: Was the PROM development study performed in a sample representing the target population for which the PROM was developed?
Item 6: Was an appropriate qualitative data collection method used to identify relevant items for a new PROM?
Finally, we added a further item to the checklist (derived from FDA criteria):
Item 6a: Were appropriate literature searches conducted to identify relevant items for a new PROM?

Quality analysis

Table 1 details the COSMIN standards [3, 13] against which our extracted PROs were assessed. Consistent with previous publications [5–7], we focus solely on checklist items associated with PRO content (and which map directly to FDA guidance [4, 5]). Extracted information on each of the seven content development domains from the PRO development papers was rated using a 4-point scale; "good" (3), "adequate" (2), "doubtful" (a), "poor/none" (0). Instruments were independently rated by two reviewers, and discrepancies were discussed and resolved. We also catalogued each PRO's number of items, item wording, response format, recall period, and administration format.

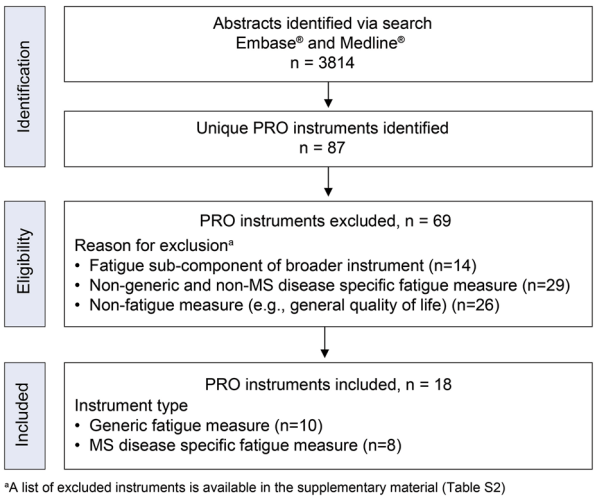

Fig. 1 PRISMA flow diagram illustrating literature searches, inclusion/exclusion criteria, and identification of 18 fatigue PRO instruments

Changes in quality of instrument development over time

To determine if PRO development quality has changed over time, we rank-ordered the PROs by development year and plotted total COSMIN scores over time. We compared average COSMIN scores of PROs developed before and after the FDA guidance was published. We examined scatterplots and computed the correlation between total COSMIN score and annual PubMed frequency of PRO development papers (a proxy of instrument use). We searched clinicaltrials.gov to determine the number of clinical trials using each PRO, and examined the correlation between total COSMIN score and the annual number of studies using PROs. Statistical analyses were conducted in JASP v.0.16.2 (https://jasp-stats.org/).

Ethics

This article is based on previously conducted studies and does not contain any new studies with human participants or animals. Ethics committee approval was therefore unnecessary.

RESULTS

Literature review

Searches identified 3814 abstracts containing 87 unique PROs. Figure 1 and Supplementary Table S2 show why 69 PROs were excluded.

PRO descriptions

We retained 18 PROs for quality analysis, comprising 10 generic fatigue measures and 8 MS-specific measures. Table 2 presents descriptive information, including the original development references, linked papers, and instrument characteristics.

Quality analysis

Table 3 provides the content development quality assessment for each instrument. Table 4 summarises the overall findings across all 18 PROs. Table S1 provides text extracts from the development papers that informed our scoring. Below, we summarise the findings for each assessment criterion.

Fatigue construct description

COSMIN guidance states that the construct description 'should be clear enough to judge whether the items of a PROM are relevant for the construct and whether the construct is comprehensively covered by the items' [3]. Only one instrument [Neurological Fatigue Index—Multiple Sclerosis (NFI-MS)] was rated 'good', where the definition was elaborated in a separate publication [14, 15]. Table S3 provides all fatigue definitions provided by the PRO authors, enabling their comparison. Half [9/18 (50%)] were rated 'doubtful'. Seven (39%) were rated 'poor/none'. For example, the Fatigue Severity Scale (FSS) authors say the instrument was designed to 'measure fatigue severity' without first providing a definition of fatigue, simply stating that fatigue 'has been notoriously difficult to define' [16].

Table 2 Properties of fatigue PRO instruments

PRO instrument	Reference (*linked publication*)		Citations/ year[a]	Trial use/ year[b]	Intended user group	Number of Items	Response categories	Recall period
	Year	Author						
Fatigue Severity Scale (FSS)	1989	Krupp [16]	42[c]	99	Medical and neurological disorders	9	1 – strongly disagree to 7 – strongly agree	NS
Multiple Sclerosis-Specific Fatigue Severity Scale (MFSS)	1989	Krupp [16]	42[c]	0	MS	6	1 – strongly disagree to 7 – strongly agree	NS
Visual Analogue Scale for Fatigue (VAS-F)	1991	Lee [20]	9	2	Healthy subjects and patients who complain of poor sleep	18	100 mm VAS scale per item	How you feel right now
Chalder Fatigue Questionnaire/ Scale (CFQ)	1993	Chalder [28]	21	3	Hospital and community populations	14	'Better than usual', 'no more than usual', 'worse than usual', 'much worse than usual'	NS
Fatigue Assessment Instrument/ Inventory (FAI)	1993 (*1989*)	Schwartz [29] (*Krupp* [16])	4	1	Patients with fatigue as a major presenting symptom	29	1—Completely disagree to 7—Completely agree	Past 2 weeks
Fatigue Impact Scale (FIS)	1994 (*1984*) (*1989*)	Fisk [18] (*Freal* [30]) (*Krupp* [16])	11	157	Wide range of patient groups	40	0—no problem to 4—extreme problem	Past month
Multidimensional Fatigue Inventory (MFI)	1995	Smets [31]	30	0	Patients (NS)	20	1—yes that is true to 5—no, that is not true	Lately
Modified Fatigue Impact Scale (MFIS)	1997 (*1994*)	Ritvo [32] (*Fisk* [18])	N/A	138	MS	21	0—Never to 4—Almost always	Past 4 weeks

Table 2 continued

PRO instrument	Reference (*linked publication*)		Citations/ year[a]	Trial use/ year[b]	Intended user group	Number of Items	Response categories	Recall period
	Year	Author						
Fatigue Descriptive Scale (FDS)	1999 (*1994*)	Iriarte [33] (*Iriarte* [34])	1	0	Clinically definite MS	12	Variable, depending on item	NS
Schedule of Fatigue and Anergia for General Practice (SOFA-GP)	2000	Hadzi-Pavlovic [35]	1	0	Community and primary care settings	10	1—None or a little to 4—Most of the time	Past few weeks
Iowa Fatigue Scale (IFS)	2003	Hartz [19]	1	0	General primary care patients	11	1—Not at all to 5—Extremely	Past month
Fatigue Assessment Scale (FAS)	2004	Michielsen [36]	N/A	0	General population	10	1—Never to 5—Always	How you usually feel
Fatigue Scale for Motor and Cognitive Functions (FSMC)	2009	Penner [37]	9	23	Fatigued MS patients	20	5-Point Likert scale from 'Does not apply at all' to 'Applies completely'	Normal day-to-day life
Unidimensional Fatigue Impact Scale (U-FIS)	2009 (*1994*)	Meads [38] (*Fisk* [18])	1	1	MS	22	0—Never to 3—all the time	Past 1 week
Neurological Fatigue Index—Multiple Sclerosis (NFI-MS)	2010 (*2008*)	Mills [15] (*Mills* [14])	3	1	MS	10	0—Strongly disagree to 3—Strongly agree	Past 2 weeks
PROMIS Fatigue MS	2012 (*2007*) (*2010*) (*2011*)	Cook [39] (*De Walt* [40]) (*Riley* [41]) (*Lai* [42])	2	0	MS	8	Never, Rarely, Sometimes, Often, or Always	Past 4 weeks

Table 2 continued

PRO instrument	Reference (*linked publication*)		Citations/ year[a]	Trial use/ year[b]	Intended user group	Number of Items	Response categories	Recall period
	Year	Author						
Fatigue Symptoms and Impacts Questionnaire -Relapsing Remitting Multiple Sclerosis (FSIQ-RMS)	2019	Hudgens [17]	1	3	RRMS	20	Variable, depending on item	Variable, depending on item (24 h, past 7 days)
Short Fatigue Questionnaire (SFQ)	2020 (*1994*)	Penson [43] (*Vercoulen* [44])	3	0	Clinical and research settings	4	1—Yes, that is true to 7—no, that is not true	NS

Italicised references denote linked publications

MS multiple sclerosis, *N/A* not available, *NS* details not specified in development publication, *RRMS* relapsing remitting MS

[a]Total number of citations according to PubMed divided by the number of years since publication of development paper

[b]Total number of trials using instrument according to clinicaltrials.gov divided by the number of years since publication of development paper

[c]The FSS and MFSS have the same number of citations per year because they have the same development paper

Table 3 Quality analysis of content development for 18 fatigue PRO instruments

PRO instrument	Construct[a]	Conceptual framework	Target population	Context of use	Development sample	Qualitative work	Literature review	Total COSMIN score[b]
FSS	0	0	0	1	0	0	0	1
MFSS	0	0	0	1	0	0	0	1
VAS-F	1	0	2	3	1	0	0	7
CFQ	1	1	1	3	0	0	0	6
FAI	0	0	3	3	0	1	0	7
FIS	0	2	2	1	1	1	0	7
MFI	1	2	1	3	0	0	0	7
MFIS	1	1	3	3	1	1	0	10
FDS	2	1	2	3	0	0	0	8
SOFA-GP	0	2	2	3	0	0	0	7
IFS	0	1	3	3	0	0	0	7
FAS	1	1	2	1	0	0	0	5
FSMC	1	2	2	3	1	2	0	11
U-FIS	1	2	1	1	3	3	0	11
NFI-MS	3	3	2	2	2	3	0	15
PROMIS Fatigue MS	1	1	2	2	2	3	3	14
FSIQ-RMS	0	3	3	3	3	3	2	17
SFQ	1	0	0	3	0	0	0	4

3 = Good, *2* = Adequate, *1* = Doubtful, *0* = Poor/none

[a]Definitions of fatigue from PRO instrument development papers are available in Table S3

[b]Ordinarily, COSMIN assessment is not scored, and instead utilises a 'worst score counts' approach. However, we nonetheless produced a 'total score', thus enabling us to evaluate how PRO development standards have changed over time (see Fig. 2); these scores should not be interpreted as a 'recommendation' for higher scoring PROs

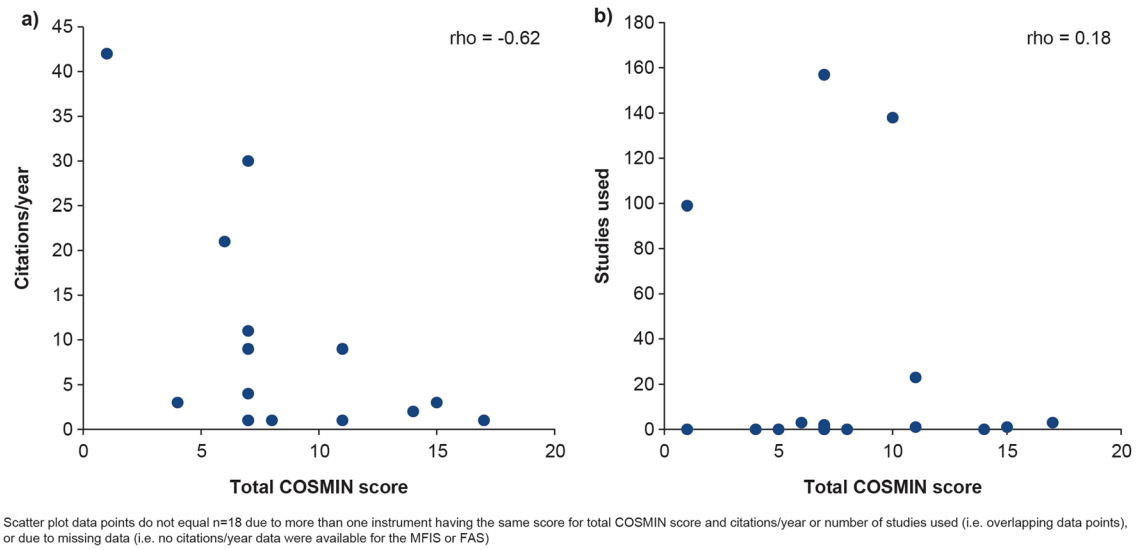

Fig. 2 Association between total COSMIN score and **a** citations/year and **b** studies using instrument

Conceptual/theoretical framework

COSMIN guidelines require the origin of the target construct to be 'based on a theory, conceptual framework, or disease model'. Here, only the NFI-MS and the Fatigue Symptoms and Impact Questionnaire—Relapsing Multiple Sclerosis (FSIQ-RMS) received a 'good' rating. The developers produced conceptualisations of fatigue in MS using semi-structured qualitative interviews with PLwMS, developing a framework of themes/subtheme, for which they evidence widespread endorsement [14]. These themes were used as a substrate for PRO development [15, 17].

Five instruments (28%) provided *some* conceptual basis (of sorts) for item selection and were rated 'adequate' rather than 'good', often because the link between the conceptual underpinning and item selection was not as comprehensive as COSMIN criteria require. For example, in the development paper for the Fatigue Impact Scale (FIS), the authors state that the 'measure was designed as a specific health status measure according to the taxonomy of Guyatt et al.', that they 'adopted the viewpoint expressed by the Canadian MS Research Group that 'measuring the effect of fatigue on activities … is more sensitive than simply asking patients to rate fatigue', and that 'Items for the FIS were selected on the basis of existing fatigue questionnaires' [18]. Despite the developers of the FIS clearly giving thought to the conceptual underpinnings of their instrument, it is unclear exactly how the conceptual basis drove item selection for this instrument.

Six instruments (33%) were rated as 'doubtful' and five (28%) as 'poor/none', where the development publications either made cursory reference to a conceptual framework, with limited information on how the framework drove item selection, or had no apparent conceptual underpinning whatsoever.

Target population

COSMIN recommends instrument developers provide a clear description of the target population for which the PRO was developed, including details about disease types, characteristics, and demographics. Additionally, if the instrument was developed for use across multiple populations, each should be clearly described. Most development papers provided target population descriptions that were rated as 'good' [4/18 (22%)] or 'adequate' [8/18 (44%)]. While the described target population must be clearly specified, it can still receive a good rating even if it covers a potentially broad population. For example, the development paper for the

Iowa Fatigue Scale (IFS) (rated as 'good') simply states that the 'development and testing of the IFS was performed on general patients in primary care and is designed to be used for that group of patients' [19]. Three papers contained target population descriptions rated as 'doubtful' quality. The Chalder Fatigue Questionnaire/Scale (CFQ) development publication, for example, describes the instrument as 'a short scale which can be used in both hospital and community populations', with no details given about the types of hospital patients or whether the intended community population is healthy or not. Development publications for three PROs [FSS, Multiple Sclerosis-Specific Fatigue Severity Scale (MFSS), Short Fatigue Questionnaire (SFQ)] did not specify a target population.

Context of use

PRO developers should be clear for which application the instrument was developed [3], such as for discriminative, evaluative, or predictive applications. Context of use may also refer to a specific setting for which the instrument was developed (e.g. for use in a hospital or at home) or a specific administration mode (e.g. paper or computer-administered). Across all COSMIN criteria, context of use had the highest number of instruments ranked as 'good' [11/18 (61%)]. The Visual Analogue Scale for Fatigue (VAS-F) is an example. The developers suggest 'potential uses including assessments of fatigue before and after clinical interventions as an indication of the effectiveness of therapy' [20]. Two instruments received 'adequate' ratings, and five were rated 'doubtful'. For example, the developers of the Unidimensional Fatigue Impact Scale (U-FIS, rated 'doubtful') suggest that their instrument may be 'valuable for future studies interested in MS-related fatigue'. While this gives a general indication of the potential context of use, it is not clear what type of studies are being referred to nor in which specific population of PLwMS.

Representative development sample

Health concepts can be context-dependent, in degree or nature. Therefore, both COSMIN and FDA recommend that PRO items are generated from qualitative work using samples representative of those in which the instrument will be used [3, 4], and should include a diversity of patients with different characteristics to cover the breadth of the concept of interest. Of the instruments assessed, only two (11%) rated 'good' (FSIQ-RMS and U-FIS). The development papers for both PROs provided clear descriptions of their qualitative research samples, which represented the intended target population for the finalised instrument. Two instruments were rated as 'adequate' (NFI-MS and PROMIS Fatigue MS), and four as 'doubtful'. The remaining ten instruments (56%) lacked a representative development sample, generally because of an absence of any documented qualitative research.

Qualitative work to generate items

COSMIN and the FDA require clarity about the process of item generation, including details about all qualitative work undertaken [3, 4]. Four instruments (22%) rated as 'good' (U-FIS, NFI-MS, PROMIS Fatigue MS, and FSIQ-RMS). They provided clear descriptions of their research process and how findings guided item selection. However, over half of the instruments (56%; 10/18) lacked any qualitative work in their development. Three instruments (17%) were rated as 'doubtful', as their development papers simply stated that interviews were conducted. They failed to describe how interview findings informed item generation or selection.

Literature reviews

The FDA recommends conducting literature reviews as part of the iterative PRO development process, to help identify measurement domains and items [4]. Only two instruments (FSIQ-RMS and PROMIS Fatigue MS) provided some description for how literature searches guided aspects of development. Literature reviews were not conducted during the development process for most PROs (16/18 [89%]).

Summary of COSMIN scoring and change in quality over time

Table 4 shows that 'target population' and 'context of use' had the highest number of

Table 4 Summary of content development quality analysis for fatigue PRO instruments

	Construct, *n* (%)	Conceptual framework, *n* (%)	Target population, *n* (%)	Context of use, *n* (%)	Development sample, *n* (%)	Qualitative work, *n* (%)	Literature review, *n* (%)
Good (3)	1 (6%)	2 (11%)	4 (22%)	11 (61%)	2 (11%)	4 (22%)	1 (6%)
Adequate (2)	1 (6%)	5 (28%)	8 (44%)	2 (11%)	2 (11%)	1 (6%)	1 (6%)
Doubtful (1)	9 (50%)	6 (33%)	3 (17%)	5 (28%)	4 (22%)	3 (17%)	0
Poor/none (0)	7 (39%)	5 (28%)	3 (17%)	0	10 (56%)	10 (56%)	16 (89%)

Proportions in each column may not sum to 100% as a result of rounding
PROs patient-reported outcomes

'good' and 'adequate' ratings. Conversely, 'construct definition', 'use of a guiding conceptual framework', 'qualitative work with an associated well-defined sample', and 'literature reviews to inform item selection' had the highest number of 'doubtful' and 'poor/none' ratings.

Overall, the reasons for poorer ratings may be the lack of guidance [3, 4]. Total COSMIN scores remained relatively stable until the mid-2000s ($n = 12$, median score = 7, of max 21), and have increased significantly since ($n = 6$, median score = 12.5, $U = 10.0$, $p = 0.02$), with the exception of the SFQ. This coincides with the FDA PRO guidance publications of 2006 (draft, [21]) and 2009 (final, [4]). The five highest scoring instruments (Fatigue Scale for Motor and Cognitive Functions [FSMC], U-FIS, NFI-MS, PROMIS Fatigue MS, and FSIQ-RMS) were all published since 2009.

Despite improvements in development quality, no instrument achieved a 'good' rating across all criteria. Furthermore, there was a negative, moderate-magnitude, correlation between total COSMIN score and the annual frequency of development paper PubMed citation (Rho = − 0.62), implying PRO selection is inversely related to development quality. Figure 3a (the associated scatterplot) shows that the moderate negative magnitude correlation is driven by high use of the FSS/MFSS (when this is excluded, rho = − 0.42). We think it is more correct to say there is no relationship between COSMIN score and use. It is notable that the five mostly highly scoring fatigue PROs have not had much use. Finally, there was a low association (Rho = 0.18) between total COSMIN score and the annual frequency of PRO use in clinical trials, according to clinicaltrials.gov. Figure 3b (the associated scatterplot) shows three PROs (FSS, FIS, and MFIS) have been heavily deployed, while the remainder have been very rarely used. Again, fatigue PRO use in clinical trials has not been driven by development quality, as measured using COSMIN criteria and scoring.

DISCUSSION

When PROs are selected for studies, we assume that their measured effects will adequately approximate the actual, but unmeasurable, effects. Three main PRO requirements are needed to satisfy this assumption. First, PROs must be valid indicators of the constructs they intend to measure. Second, PROs must satisfy statistical measurement criteria. Third, PROs must be to able adequately detect change when it occurs. Our study concerns validity, the most

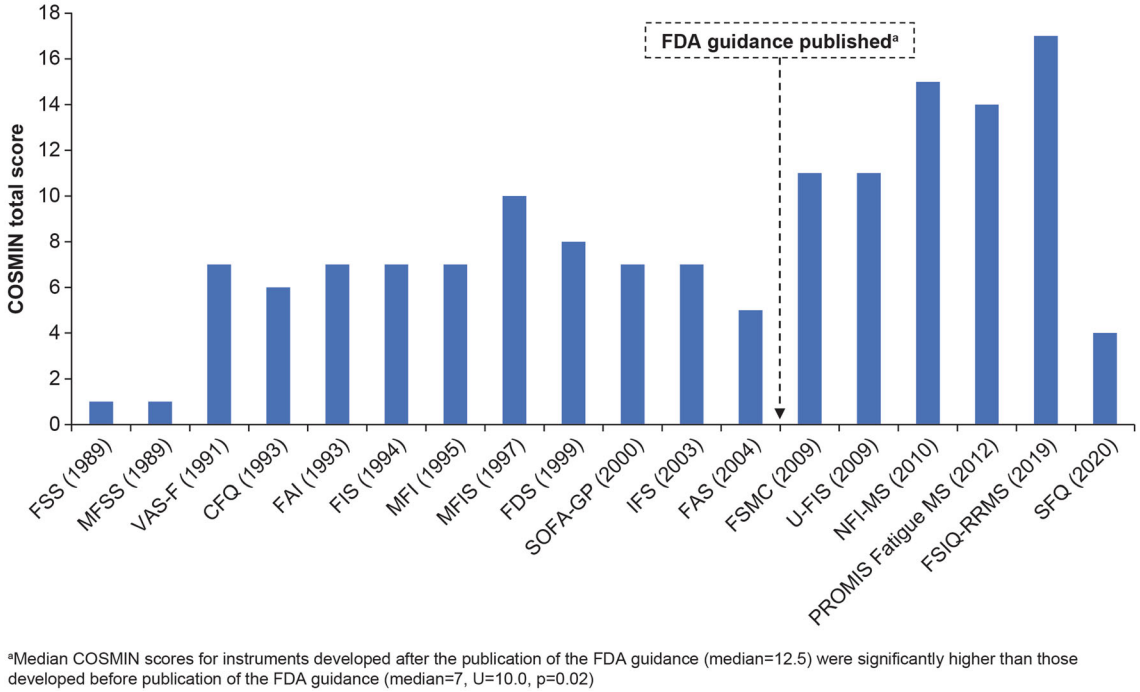

Fig. 3 Fatigue PRO instrument development quality over time

fundamental of the three requirements and a prerequisite for the others to be interpreted meaningfully. When PROs lack validity, type-II errors, with potentially pervasive implications, result. We believe this justifies a very critical approach to evaluating PROs.

Historically, content validity has always been identified as a measurement requirement, but its fundamental importance was not emphasised [22, 23]. This may have been because health measurement borrowed its methods from educational testing and measurement, where the educational curriculum set the framework for testing, thereby enabling content validity to be determined relatively easily. Then, the mainstay of validity testing was psychometric (statistical) examinations of PRO scores (convergent and discriminant construct validity, group differences and hypothesis testing). These tests were exposed as providing only circumstantial validity evidence 40 years ago [24, 25]. That message, however, seemed to go largely unnoticed until FDA published their guidance, emphasising content validity's central and fundamental importance in health measurement [4, 21, 26].

Our COSMIN-based appraisal of 18 fatigue PROs, selected from a pool of 87 instruments, showed considerable variability in quality. Five PROs achieved relatively high ratings across most categories (FSMC, U-FIS, NFI-MS, PROMIS Fatigue MS, and FSIQ-RMS), but none fully satisfied COSMIN criteria. Consistent weaknesses were for construct definitions, qualitative work, use of development samples representative of the target population, and use of literature reviews. Only the NFI-MS rated 'Good' for having a well-defined construct as well as the use of a conceptual framework. Our findings are concerning, as these are fundamental requirements for achieving valid PRO content development, and preclude definitive recommendations about which PROs might be most suitable for certain contexts.

PRO development chronology explains some of their variability in developmental quality. Median COSMIN scores increased significantly after FDA guidance was published [4, 21],

implying an influential intervention. However, the scatterplots between total COSMIN score and annual citations, and between total COSMIN score and annual frequency in clinical trials, imply that PRO quality may not drive use in studies nor selection for clinical trials. This means that studies influencing the care of PLwMS and our research directions have almost certainly been misleading, with potentially damaging implications for the quality of life for PLwMS. This also underscores the value of formalising PRO selection strategies and information dissemination.

While COSMIN criteria provide a useful framework to assess PRO content validity, and represent an important step in PRO quality control, we believe that they require further development to facilitate robust objective assessment. For example, COSMIN criteria require construct definitions, theories or conceptual frameworks to provide clear origins for the concept of interest, and the use of appropriate qualitative data collection methods. However, improved clarity is needed for exactly what constitutes an adequate construct definition, a sufficiently high-quality conceptual framework, and an adequate-quality qualitative work.

Given the increasingly important and influential role PROs play in clinical trials and patient care, we also feel that there are several areas where COSMIN criteria could have increased clarity and set a more rigorous benchmark. In some circumstances, "Good" ratings can be easily achieved and content validity therefore overestimated. Examples are the IFS for Target population ('development and testing of the IFS was performed on general patients in primary care, and is designed to be used for that group of patients' [19]), and VAS-F for Context of use ('potential uses including assessments of fatigue before and after clinical interventions as an indication of the effectiveness of therapy' [20]). According to current COSMIN guidance, both situations can be given good ratings because reviewers can assume 'that the PROM is applicable in all patients ' in primary care or in whom an intervention is being studied (see guidance, p.18, [13]). From our perspective, such guidance risks PRO misuse. We recommend that under such circumstances users should be encouraged to examine the performance of a PRO empirically in a sample representative of their content of use.

Another important consideration is the relative importance, and order, of COSMIN criteria. Logically, clear construct definitions and conceptual frameworks are prerequisites before other criteria are assessed. We found that these criteria had the highest number of 'doubtful' and 'poor/none' ratings. This implies that the use of COSMIN total or aggregate scores can give misleading impressions of scale development quality.

We suggest there is one specific area where COSMIN requires significant development in its guidance: evaluating the articulation of the scored concepts and components (subscale) as item sets. This is a critical step, separate from concept definitions and conceptual framework development, as the items link the concept/component of interest and the score. We demonstrate the importance of this step by examining—as an additional, subsequent evaluation prompted by our concerns about COSMIN criteria's completeness—the relevant item sets of the two highest COSMIN scoring fatigue PROs, NFI-MS and FSIQ-RMS.

Table 5 shows the item sets that generated the NFI-MS and FSIQ-RMS scores. The relevant question is, to what extent do the item sets represent the constructs measured by the subscales? A careful examination highlights our concerns. There are no definitions for the measured components, which makes it impossible to judge the item sets that generated the scores. Moreover, many of the items are non-specific and, as such, they are confounded. Often the relationships between item sets and underlying conceptual frameworks are ambiguous. For example, while the initial (57-item) set for the NFI-MS was explicitly derived from the thematic framework, the original items themselves are not provided. This also makes it difficult to judge the full item set, in terms of its representation of the purported measurement concept(s). This 57-item set was then reduced to the 4 subscales of the finalised instrument, and a 10-item summary subscale derived from two of the four subscales. This item reduction process

Table 5 Scored item sets for (a) the FSIQ-RMS and (b) the NFI-MS

(a)

PRO	Fatigue symptoms and impact questionnaire relapsing multiple sclerosis (FSIQ-RMS)				
Section	SYMPTOMS (Sect. 1)	IMPACTS (Sect. 2)			
SCORE	SYMPTOMS scale	Impacts scale	Physical Impacts subscale	Cognitive/Emotional Impacts subscale	Coping Impacts subscale
No. of items	7	13[a]	5	5	5
Author's fatigue definition	Often described as a feeling of extreme mental or physical exhaustion, MS-related fatigue has far-reaching effects on quality of life, employment, and productivity, imposing limitations independently of MS-related physical disability and depression				
FSIQ-RMS item	How physically tired did you feel? (S1, 1[b])	Difficulty running errands (S2, 1)	Difficulty running errands (S2, 1)	Difficulties communicating clearly (S2, 2)	Difficulties motivating yourself to do routine daily activities (S2, 4)
	How mentally tired did you feel? (S1, 2)	Difficulties communicating clearly (S2, 2)	Difficulty doing indoor household chores (S2, 5)	Difficulties thinking clearly (S2, 3)	Difficulty taking part in social activities (S2, 8)[c]
	How physically weak did you feel? (S1, 3)	Difficulties thinking clearly (S2, 3)	Difficulty walking (S2, 6)	Difficulty maintaining relationships with people close to you (S2, 7)	Frequency of taking a nap (S2, 11)
	Rate your energy level (S1, 4)	Difficulties motivating yourself to do routine daily activities (S2, 4)	Difficulty taking part in social activities[c] (S2, 8)	How frustrated (S2, 9)	Frequency of needing to take a break (S2, 12)
	How worn out did you feel? (S1, 5)	Difficulty doing …indoor household chores (S2, 5)	Frequency of rearranging plans (S2, 13)[c]	How often forgetful (S2, 10)	Frequency of rearranging plans (S2, 13)[c]
	How sleepy did you feel? (S1, 6)	Difficulty walking (S2, 6)			
	How worn out did you feel at rest? (S1, 7)	Difficulty maintaining relationships with people close to you (S2, 7)			
		Difficulty taking part in social activities[c] (S2, 8)			
		How frustrated (S2, 9)			
		How often forgetful (S2, 10)			
		Frequency of taking a nap (S2, 11)			
		Frequency of needing to take a break (S2, 12)			
		Frequency of rearranging plans (S2, 13)[c]			

Table 5 continued

(b)

PRO	**Neurological fatigue index multiple sclerosis (NFI-MS)**				
Score	**Summary subscale**[d]	**Physical subscale**	**Cognitive subscale**	**Relief by Diurnal Sleep or Rest subscale**	**Abnormal Nocturnal Sleep or Sleepiness subscale**
No. of items	10	8	4	6	5
Author's fatigue definition	Fatigue has been defined, as a result of qualitative analysis, as a: 'reversible motor and cognitive impairment with reduced motivation, and a desire to rest, either appearing spontaneously or brought on separately by mental or physical activity, humidity, acute infection and food ingestion. It was relieved by daytime sleep or rest without sleep. It could occur at any time but was usually worse in the afternoon'. In MS, fatigue could be daily, had usually been present for years and had greater severity than any pre-morbid fatigue				
NFI-MS item	I can become tired easily [1]	I can become tired easily [1]	Sometimes I really have to concentrate on what are usually simple things [9]	I need to rest in the day [13]	I get a feeling as if I have not slept for a couple of nights [19]
	Sometimes I lose my body strength [2]	Sometimes I lose my body strength [2]	I have problems with my speech when I'm tired [10]	I need to sleep in the day [14]	I yawn a lot [20]
	My limbs can become very heavy [3]	My limbs can become very heavy [3]	My coordination gets worse as the day goes on [11]	Sleep in the day can really help me [15]	I sometimes wake in the night for no reason [21]
	My body can't keep up with what I want to do [4]	My body can't keep up with what I want to do [4]	Mental effort really takes it out of me 12]	Resting allows me to carry on [16]	When I awake in the morning I feel unrefreshed [22]
	The longer I do something the more difficult it becomes [5]	The longer I do something the more difficult it becomes [5]		I try to get everything done in the morning [17]	Often in the morning, I don't feel like getting out of bed [23]
	Sometimes I have no option but to simply stop what I've been doing [6]	Sometimes I have no option but to simply stop what I've been doing [6]		I try to rest or sleep beforehand, if I know I've got to do something that requires a lot of effort [18]	
	I usually get tired on most days [7]	I usually get tired on most days [7]			
	Sometimes I really have to concentrate on what are usually simple things [8]	I can become weak even if I've not been doing anything [8]			

Table 5 continued

(b)					
PRO	**Neurological fatigue index multiple sclerosis (NFI-MS)**				
Score	**Summary subscale[d]**	**Physical subscale**	**Cognitive subscale**	**Relief by Diurnal Sleep or Rest subscale**	**Abnormal Nocturnal Sleep or Sleepiness subscale**
	My coordination gets worse as the day goes on [11]				
	Mental effort really takes it out of me [12]				

[a]There are $k = 13$ impacts items, and three $k = 5$ impacts subscales because $k = 2$ items are scored in two subscales

[b]In parentheses: S = Sect. 1 or 2, *number* = item number in sequence order

[c]Item that is scored in 2 different subscales

[d]NFI-MS has $k = 23$ items, the Summary subscale is a selection of 10 of these items

was driven by statistical/psychometric criteria, rather than being predicated on any conceptual groundings. This makes it challenging to evaluate item concept coverage of the original conceptualisations.

There are some surprising item groupings. For example, the 4-item NFI-MS cognitive scale contains an item on coordination. Conceptually, however, we might expect this item to be in the physical subscale. The FSIQ-RMS's 5-item physical subscale, and the 5-item coping subscale, both contain the same two items. This is conceptually questionable, as such 'multidimensional' items (according to the development publication [17]) cause measurement overlap, thereby reducing subscale validity. If the item sets of all the PROs are examined closely, other concerns are evident. In essence, we think these PROs have weaker content validity, and COSMIN scores alone are unable to establish content validity of finalised item sets.

We can identify four explanations for how suboptimal item sets arise. First, there is a general under-recognition of the important stage of articulating subscales as items. It is not emphasised in any guidance we know of. Second, there is an absence of subscale definitions to guide item generation, drafting and selection. Third, during PRO development, statistical methods are commonly used to group items into scales. This groups items on their statistical, rather than their conceptual, relationships. Fourth, we think there is a dissociation between the conceptual framework development and the qualitative work. It is common for patient statements from qualitative work to be used as items, often verbatim, without careful consideration of the relationship to the concept. All four explanations threaten content validity, risking type-II errors.

Our study had limitations. We identified PROs from abstract screening and may have missed relevant PROs. We did not assess the Quality-of-Life in Neurological Disorders (Neuro-QoL) fatigue scale. This was excluded during screening as a set of fatigue items within a broader instrument [27]. A brief examination of the Neuro-QoL fatigue PRO's 19-item set highlights many of the content validity issues that we have raised. We did not assess the

psychometric properties of the PROs. Our focus was their content validity, a prerequisite for psychometric evaluations [3, 4]. However, we recognise the need for head-to-head comparisons of these PROs to determine the impact of differing development quality on measurement performance.

CONCLUSIONS

PRO instruments must be valid to minimise type-II errors and to prevent misleading study results. Item content is the main determinant of PRO validity. Here, we demonstrate that fatigue PROs used in MS research have weak content validity. While the FDA recommendations have seen an apparent increase in quality, and have spawned other guidance documents, we think they are currently too vague and lenient for the measurement rigour required by today's clinical trials. An area we feel is particularly underemphasised is the articulation of a concept by a set of items that generate scores. A comparison of measurement PRO performance is required.

We recognise our work is critical of the field. We also recognise that the work done by others on fatigue measurement, and on content validity methods and assessment, is important and necessary to underpin developments and progress. However, we believe our level of critique and debate is required to ensure measured effects approximate real effects.

ACKNOWLEDGEMENTS

Funding. This review, as well as the journal's Rapid Service Fee, was funded by Novartis Pharma AG, Basel, Switzerland.

Medical Writing and/or Editorial Assistance. Medical writing support was provided by David McMinn, PhD of Novartis CONEXTS, Dublin, Ireland, and was funded by Novartis Pharma AG, Basel, Switzerland.

Author Contributions. James Close, Miriam King, Jo Vandercappellen, and Jeremy Hobart all meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article. All named authors helped to conceive the study concept, contributed to the reviewing and revision of the manuscript, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Disclosures. James Close has nothing to disclose. Jo Vandercappellen and Miriam King were employees of Novartis Pharma AG at the time of manuscript preparation. Jeremy Hobart has received consulting fees, honoraria, support to attend meetings or research support from Acorda, Asubio, Bayer Schering, Biogen Idec, F. Hoffmann-La Roche, Genzyme, Merck Serono, Novartis, Oxford PharmaGenesis, and Teva.

Compliance with Ethics Guidelines. This article is based on previously conducted studies and does not contain any new studies with human participants or animals performed by any of the authors.

Data Availability. There are no data, per se, associated with this manuscript. Information extracted from the articles identified from the search is available in the supplementary material (Table S1).

REFERENCES

1. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. 2007;6(12):1094–105.
2. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–45.
3. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159–70.
4. Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. 2009.
5. Close J, Baines K, Burke L, Hobart J. Measuring upper limb function in MS: which existing patient reported outcomes are fit for purpose? eNeurologicalSci. 2020;19:100237.
6. Chiarotto A, Ostelo RW, Boers M, Terwee CB. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. J Clin Epidemiol. 2018;95:73–93.
7. Chiarotto A, Terwee CB, Kamper SJ, Boers M, Ostelo RW. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review. J Clin Epidemiol. 2018;102:23–37.
8. Carlton J, Powell PA, Project HCG. Measuring carer quality of life in Duchenne muscular dystrophy: a systematic review of the reliability and validity of self-report instruments using COSMIN. Health Qual Life Outcomes. 2022;20(1):57
9. Pellekooren S, Ostelo R, Pool A, van Tulder M, Jansma E, Chiarotto A. Content validity of patient-reported outcome measures of satisfaction with primary care for musculoskeletal complaints: a systematic review. J Orthop Sports Phys Ther. 2021;51(3):94–102.
10. Jerofke-Owen T, Garnier-Villarreal M, Fial A, Tobiano G. Systematic review of psychometric properties of instruments measuring patient preferences for engagement in health care. J Adv Nurs. 2020. https://doi.org/10.1111/jan.14402.
11. Cook KF, Bamer AM, Roddey TS, Kraft GH, Kim J, Amtmann D. Multiple sclerosis and fatigue: understanding the patient's needs. Phys Med Rehabil Clin N Am. 2013;24(4):653–61.
12. Sellitto G, Morelli A, Bassano S, Conte A, Baione V, Galeoto G, et al. Outcome measures for physical fatigue in individuals with multiple sclerosis: a systematic review. Expert Rev Pharmacoecon Outcomes Res. 2021;21(4):625–46.
13. Terwee CB, Prinsen CAC, Chiarotto A, de Vet HC, Bouter LM, Alonso J, et al. COSMIN methodology for assessing the content validity of PROMs User manual. version 1.0. 2018. https://cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf.
14. Mills RJ, Young CA. A medical definition of fatigue in multiple sclerosis. QJM. 2008;101(1):49–60.
15. Mills RJ, Young CA, Pallant JF, Tennant A. Development of a patient reported outcome scale for fatigue in multiple sclerosis: the Neurological Fatigue Index (NFI-MS). Health Qual Life Outcomes. 2010;8:22.
16. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol. 1989;46(10):1121–3.
17. Hudgens S, Schuler R, Stokes J, Eremenco S, Hunsche E, Leist TP. Development and validation of the FSIQ-RMS: a new patient-reported questionnaire to assess symptoms and impacts of fatigue in relapsing multiple sclerosis. Value Health. 2019;22(4):453–66.
18. Fisk JD, Ritvo PG, Ross L, Haase DA, Marrie TJ, Schlech WF. Measuring the functional impact of fatigue: initial validation of the fatigue impact scale. Clin Infect Dis. 1994;18(Suppl 1):S79-83.
19. Hartz A, Bentler S, Watson D. Measuring fatigue severity in primary care patients. J Psychosom Res. 2003;54(6):515–21.
20. Lee KA, Hicks G, Nino-Murcia G. Validity and reliability of a scale to assess fatigue. Psychiatry Res. 1991;36(3):291–8.
21. Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support

labeling claims: draft guidance. Health Qual Life Outcomes. 2006;4:79.

22. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther. 1996;18(5): 979–92.

23. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 4th ed. Oxford University Press, Oxford; New York; xvii, pp 431. 2008

24. Stenner AJ, Smith M, Burdick DS. Toward a theory of construct definition. J Educ Meas. 1983;20: 305–16.

25. Stenner A, Smith M. Testing construct theories. Percept Mot Skills. 1982;55:415–26.

26. Food and Drug Administration. Roadmap to Patient-focused Outcome Measurement in Clinical Trials. https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/roadmap-patient-focused-outcome-measurement-clinical-trials-text-version. 2013.

27. Cella D, Lai JS, Nowinski CJ, Victorson D, Peterman A, Miller D, et al. Neuro-QOL: brief measures of health-related quality of life for clinical research in neurology. Neurology. 2012;78(23):1860–7.

28. Chalder T, Berelowitz G, Pawlikowska T, Watts L, Wessely S, Wright D, et al. Development of a fatigue scale. J Psychosom Res. 1993;37(2):147–53.

29. Schwartz JE, Jandorf L, Krupp LB. The measurement of fatigue: a new instrument. J Psychosom Res. 1993;37(7):753–62.

30. Freal JE, Kraft GH, Coryell JK. Symptomatic fatigue in multiple sclerosis. Arch Phys Med Rehabil. 1984;65(3):135–8.

31. Smets EM, Garssen B, Bonke B, De Haes JC. The Multidimensional fatigue inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res. 1995;39(3):315–25.

32. Ritvo PG, Fischer JS, Miller DM, Andrews H, Paty DW, LaRocca NG. The Consortium of Multiple Sclerosis Centers Health Services Research Subcommittee. MSQLI. Multiple Sclerosis Quality of Life Inventory: A User's Manual. 1997.

33. Iriarte J, Katsamakis G, de Castro P. The Fatigue Descriptive Scale (FDS): a useful tool to evaluate fatigue in multiple sclerosis. Mult Scler. 1999;5(1): 10–6.

34. Iriarte J, de Castro P. Proposal of a new scale for assessing fatigue in patients with multiple sclerosis. Neurologia. 1994;9(3):96–100.

35. Hadzi-Pavlovic D, Hickie IB, Wilson AJ, Davenport TA, Lloyd AR, Wakefield D. Screening for prolonged fatigue syndromes: validation of the SOFA scale. Soc Psychiatry Psychiatr Epidemiol. 2000;35(10):471–9.

36. Michielsen HJ, de Vries J, van Heck GL, van de Vijver FJR, Sijtsma K. Examination of the dimensionality of fatigue: The construction of the Fatigue Assessment Scale (FAS). Eur J Psychol Assess. 2004;20(1):39–48.

37. Penner IK, Raselli C, Stocklin M, Opwis K, Kappos L, Calabrese P. The Fatigue Scale for Motor and Cognitive Functions (FSMC): validation of a new instrument to assess multiple sclerosis-related fatigue. Mult Scler. 2009;15(12):1509–17.

38. Meads DM, Doward LC, McKenna SP, Fisk J, Twiss J, Eckert B. The development and validation of the Unidimensional Fatigue Impact Scale (U-FIS). Mult Scler. 2009;15(10):1228–38.

39. Cook KF, Bamer AM, Roddey TS, Kraft GH, Kim J, Amtmann D. A PROMIS fatigue short form for use by individuals who have multiple sclerosis. Qual Life Res. 2012;21(6):1021–30.

40. DeWalt DA, Rothrock N, Yount S, Stone AA, Group PC. Evaluation of item candidates: the PROMIS qualitative item review. Med Care. 2007;45(5 Suppl 1):S12-21

41. Riley WT, Rothrock N, Bruce B, Christodolou C, Cook K, Hahn EA, et al. Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: further evaluation of content validity in IRT-derived item banks. Qual Life Res. 2010;19(9):1311–21.

42. Lai JS, Cella D, Choi S, Junghaenel DU, Christodoulou C, Gershon R, et al. How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS fatigue item bank example. Arch Phys Med Rehabil. 2011;92(10 Suppl):S20–7.

43. Penson A, van Deuren S, Worm-Smeitink M, Bronkhorst E, van den Hoogen FHJ, van Engelen BGM, et al. Short fatigue questionnaire: screening for severe fatigue. J Psychosom Res. 2020;137: 110229.

44. Vercoulen JH, Swanink CM, Fennis JF, Galama JM, van der Meer JW, Bleijenberg G. Dimensional assessment of chronic fatigue syndrome. J Psychosom Res. 1994;38(5):383–92.