Rheumatol Ther (2021) 8:1287–1301
https://doi.org/10.1007/s40744-021-00338-4

ORIGINAL RESEARCH

Psychometric Analysis from EMBODY1 and 2 Clinical Trials to Help Select Suitable Fatigue PRO Scales for Future Systemic Lupus Erythematosus Studies

Sophie Cleanthous · Sabine Bongardt · Patrick Marquis · Christian Stach · Stefan Cano · Thomas Morel

Received: April 27, 2021 / Accepted: June 14, 2021 / Published online: July 9, 2021

ABSTRACT

Introduction: Fatigue is one of the most important symptoms reported by patients with systemic lupus erythematosus (SLE) and a key concept of interest in SLE clinical trials. Despite this, fatigue remains poorly understood and sub-optimally measured by existing patient-reported outcome (PRO) instruments and scales. Here, we psychometrically evaluated the measurement properties of three PRO scales that purport to measure fatigue, using data from two SLE clinical trials.

Methods: Data were pooled from two completed phase 3 SLE trials: EMBODY1 (NCT 01262365) and EMBODY2 (NCT01261793). FACIT-F, SF-36 Vitality and LupusQoL Fatigue data were selected for post hoc Rasch Measurement Theory psychometric analysis in two stages: (1) scale-to-sample targeting, thresholds for item response options, item fit statistics, and reliability; and (2) proposal and evaluation of pooled fatigue items based on the best-performing items. Responsiveness analyses on group-level (two effect size [ES] calculations and relative efficiency) and individual level (within person statistically significant difference), were conducted to compare original scales and pooled item sets.

Results: Scale-to-sample targeting was good for FACIT-F, but suboptimal for SF-36 Vitality and LupusQoL Fatigue. Thresholds for item response options were ordered for all three scales. Item misfit was found in all three scales (FACIT-F 10/13; SF-36 Vitality 4/4; LupusQoL Fatigue 1/4). Reliability statistics were good for FACIT-F (0.93) and LupusQoL Fatigue (0.80) but low for SF-36 Vitality (0.53). The pooled fatigue items improved some psychometric properties despite persisting misfit issues (2/10) and were more sensitive in detecting change at week 24 compared with un-pooled data (ES 0.41 vs. 0.26–0.25).

Conclusions: FACIT-F, SF-36 Vitality, and LupusQoL Fatigue were found to have important limitations in the EMBODY1 and EMBODY2 SLE clinical trials. Findings from pooled fatigue items support the need for further research to improve conceptual underpinnings of fatigue PROs and make them fit for purpose for drug development.

Keywords: Systemic lupus erythematosus; Autoimmune diseases; Quality of life

S. Cleanthous · S. Cano
Modus Outcomes, Letchworth Garden City, UK

S. Bongardt · C. Stach
UCB Pharma, Monheim, Germany

P. Marquis
Modus Outcomes, Newton, MA, USA

T. Morel (✉)
UCB Pharma, Allée de la Recherche 60, 1070 Anderlecht, Brussels, Belgium
e-mail: thomas.morel@ucb.com

Key Summary Points

Why carry out this study?

Fatigue is one of the most important symptoms reported by patients with systemic lupus erythematosus (SLE), yet is poorly understood and sub-optimally measured by existing patient-reported outcome (PRO) scales.

This study aimed to psychometrically evaluate the measurement properties of three PRO scales that purport to measure fatigue.

What was learned from the study?

Pooled, blinded data from two completed phase 3 SLE trials, EMBODY1 (NCT01262365) and EMBODY2 (NCT01261793) identified item misfit, suboptimal scale-to-sample targeting and low reliability in the psychometric analyses of the PRO scales assessed.

This highlights that FACIT-F, SF-36 Vitality, and LupusQoL Fatigue show limitations in SLE clinical trials.

These findings support the need for further patient-centered research to build an appropriate conceptualization of SLE fatigue to further support the development of a fit-for-purpose fatigue PRO scale for use in the context of SLE.

DIGITAL FEATURES

This article is published with digital features, including a summary slide, to facilitate understanding of the article. To view digital features for this article go to https://doi.org/10.6084/m9.figshare.14779596.

INTRODUCTION

Fatigue is one of the most common symptoms reported by patients affected with systemic lupus erythematosus (SLE) [1, 2], but despite this, it is poorly addressed by available treatments. As a concept, fatigue is complex, poorly understood, and sub-optimally measured, as the US Food and Drug Administration (FDA) acknowledged in its guidance on SLE [3]. Fatigue, however, is a key concept of interest (COI) [4–6] in industry-sponsored clinical trials [7–9]. Three legacy patient-reportedoutcome (PRO) scales are commonly used to measure fatigue. The Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F) scale [10], originally developed to assess fatigue associated with anemia in cancer, has subsequently been used across cancer groups [11, 12], general US populations [13], and in rheumatology [14], including SLE [15]. The Medical Outcomes Study Short-Form (SF-36) Vitality scale, which is part of a larger generic health-related quality of life (HRQoL) PRO instrument [16, 17], was developed on the basis of the RAND Health Insurance Experiment (HIE) [18] and is a PRO mainstay in SLE trials [6, 19–22]. Finally, the Lupus Quality of Life Questionnaire (LupusQoL) Fatigue scale, which is also part of a multiscale disease-specific HRQoL PRO instrument developed and validated in SLE [23], is also a common choice in SLE clinical trials (NCT01262365, NCT01261793) [7].

For PRO scales to be used to evaluate treatment benefit, they must first be shown to be fit for purpose [4, 24–26]. Current best-practice guidelines point to the need for a comprehensive and explicitly detailed COI (i.e., what the PRO scales aim to measure) in the specific context of use (i.e., the specific patient population in which the PRO scales will be used) [4, 27]. PRO scales should be well defined and reliable [25, 28, 29]. Although they are widely used, the published psychometric evidence supporting the FACIT-F, SF-36 Vitality, and LupusQoL Fatigue scales for use in SLE clinical trials is mixed. All scales have been found to be consistently reliable [23, 30–32], but evidence for validity (i.e., content, construct, and known groups) [23, 33–35] and ability to detect clinical

change [31, 36, 37] is less convincing. Additionally, it is important to indicate that all three scales produce a single score for overall fatigue [38], whereas other widely used fatigue PRO scales, including some used in SLE, distinguish between physical/motor and mental/cognitive manifestation of fatigue [39, 40].

The FACIT-F, SF-36 Vitality, and LupusQoL Fatigue scales were used as exploratory endpoints in EMBODY1 and EMBODY2 (ClinicalTrials.gov identifiers NCT01262365 and NCT01261793). These two identical, phase 3, multicenter, randomized, double-blind, placebo-controlled studies (with different geographic sites), assessed the efficacy and safety of epratuzumab in SLE. The study population included adult patients with moderately to severely active SLE who fulfilled the American College of Rheumatology (ACR) revised criteria for SLE [41, 42]. In line with the primary endpoint, no statistically significant differences were observed between the placebo and treatment groups for the PRO scales [7]. To better understand the psychometric performance of the FACIT-F, SF-36 Vitality, and LupusQoL Fatigue scales in these clinical trials, we present a post hoc psychometric analysis of the data using modern psychometric methods, including an exploratory analysis of pooled fatigue items to examine potential relative measurement benefits. This work reflects an exploratory exercise to examine the impact of having an item set that is psychometrically and conceptually more cohesive and clearer, which ultimately aims to inform the self-reported assessment of fatigue in SLE studies through 'fit for purpose' PRO scales.

METHODS

Study Population

This post hoc psychometric analysis was conducted on pooled, blinded baseline and week 24 FACIT-F, SF-36 Vitality, and LupusQoL Fatigue data from patients enrolled in the EMBODY1 and EMBODY2 clinical trials. All EMBODY1 and EMBODY2 patients had either moderate or severe SLE disease activity as defined by the BILAG-2004 [43] and SLE Disease Activity 2000 (SLEDAI-2K) indices [44]. The vast majority of patients were female (90%), with a mean age of 42 years in EMBODY1 and 41 years in EMBODY 2, while time since diagnosis ranged between 0 and 43 years with a median of 6 years. The study design and population are described in detail elsewhere [7].

Compliance with Ethics Guidelines

The study protocol, amendments, and patient informed consent were reviewed by a national, regional, or Independent Ethics Committee (IEC) or Institutional Review Board (IRB). This study was conducted in accordance with the current version of the applicable regulatory and International Conference on Harmonisation (ICH)-Good Clinical Practice (GCP) requirements, the ethical principles that have their origin in the principles of the Declaration of Helsinki, and the local laws of the countries involved.

Patient and Public Involvement

No patients or members of the public were involved in the design, conduct, reporting, or dissemination plans of this work.

PRO Instruments and Scales

The FACIT-F is a 13-item fatigue PRO scale with a 7-day recall period [10]. Items are scored on a five-point Likert-type response scale ranging from 0 to 4. All items are summed to create a single fatigue score with a range from 0 to 52, with higher values representing higher levels of fatigue. The SF-36 Vitality scale is one of eight sub-scales within the SF-36 PRO instrument comprising four items related to fatigue [16, 17]. Items are scored on a five-point Likert-type frequency scale ranging from all of the time to none of the time within a 4-week recall period, summed, and converted to norm-based 0–100 scores, with higher values representing more vitality (i.e., less fatigue). The LupusQoL Fatigue scale is one of eight scales comprising the LupusQoL PRO instrument, which is made up

of four fatigue items [23]. Items are scored on a five-point Likert-type frequency scale ranging from all of the time to never, within a 4-week recall period. A score from 0 to 100 is calculated for each domain scale by dividing the mean raw domain score by four and multiplying by 100, with higher scores representing less fatigue.

Rasch Measurement Theory

Psychometrics is an umbrella term for empirical evaluations of the measurement properties (e.g., reliability, validity, ability to detect change) of rating scales and tests [24], including PRO instruments and scales. Traditional psychometric methods have important limitations that are overcome by modern methods, such as Rasch Measurement Theory (RMT) [24, 29]. RMT analysis evaluates the extent to which the observed data fit predictions of the Rasch model, which in essence defines how a set of items should perform to generate reliable and valid measurements [45, 46]. The difference between expected and observed scores indicates the degree to which rigorous measurement is achieved [29, 45]. RMT analysis has three broad aims: (1) the evaluation of the scale-to-sample targeting; (2) the evaluation of the measurement continuum; and (3) the evaluation of the sample measurement.

RMT analyses, based on the unrestricted Rasch Model for polytomous ordered responses [10], were conducted cross-sectionally on baseline data from week 0. Responsiveness (i.e., ability to detect change) analyses were conducted on longitudinal data from weeks 0 and 24 of the EMBODY trials. The goal of these analyses was to compare the PRO scales on all the available pooled blinded data, as opposed to comparing treatment arms. RUMM2030 [47] was used to conduct the RMT and IBM SPSS 25.0 [48] was used for the responsiveness analyses. Responsiveness analyses were conducted on interval level 0–100 transformed scores computed on the basis of RMT-produced interval logit for total raw scores.

There were two stages of analysis: (1) evaluation of the measurement performance of the FACIT-F, SF-36 Vitality, and LupusQoL Fatigue scales; and (2) exploration of the potential measurement benefits of pooled fatigue items selected based on the best-performing items through an empirical post hoc analysis.

Stage 1: Measurement Performance Review of FACIT-F, SF-36 Vitality, and LupusQoL Fatigue Scales

There were four main areas of psychometric evaluation: (1) scale-to-sample targeting, (2) thresholds for item response options, (3) item-fit statistics, and (4) reliability. These are presented in more detail in Table 1 (columns 1 and 2) and elsewhere [3]. We examined group-level responsiveness by computing three standard indicators: two effect size calculations (Cohen's and standardized response mean) and relative efficiency (pairwise squared *t* values from paired samples *t* tests) [49, 50].

We assessed individual-level responsiveness by computing the significance of each person's change in each scale's score [29]. The standard error of the difference (SED; i.e., the size of the error associated with each person's change) was computed for each individual ($SED = \sqrt{((SE\ Time\ 1)^2 + (SE\ Time\ 2)^2)}$). The significance of change was then determined by dividing each person's change score by the SED. Significance of change values were categorized into five groups: (1) significant improvement = significant change ≤ -1.96; (2) non-significant improvement = $-1.95 <$ significant change < 0; (3) no change = significant change = 0; (4) non-significant worsening = $0 <$ significant change $< +1.95$; and (5) significant worsening = significant change $\geq +1.96$.

Stage 2: Construction and RMT Analysis of Pooled Fatigue Symptom Item Set

There were three steps in Stage 2: (1) review of findings from Stage 1 and the conceptual content of the FACIT-F, SF-36 Vitality, and LupusQoL Fatigue scales; (2) structuring and identifying a selection of items representing fatigue symptoms based on the empirical findings from Stage 1; and (3) analysis of the psychometric properties (as described in Stage 1) of the new pooled fatigue symptom item set and comparison against the original scales.

Table 1 Summary of analysis and findings

Analysis		Results*			
Question	**Summary description—full description of these methods is presented elsewhere [29]**	**FACIT-F (13 items)**	**SF-36 Vitality (4 items)**	**LupusQoL Fatigue (4 items)**	**Pooled Fatigue Symptoms Item Set (10 items)**
How adequate is the scale-to-sample targeting?	Items should be targeted to the SLE patient population. Targeting is examined by inspecting the spread of person locations (i.e., range of fatigue reported by the sample) and item locations (i.e., range of the fatigue measured by the items on a scale). There is no specific criterion, but more coverage (%) equates to better targeting	68	63	71	49
Do the response categories work as intended?	Successive response categories, for each item, should represent increasing levels of fatigue, as reflected by ordered of the category probability curves. Ideally, 100% of thresholds should be ordered	100	100	100	90
To what extent do the items work together to define a single measurement construct?	Statistical and graphical indicators of fit are investigated:				
	(1) Fit residuals summarize the difference between observed and expected responses to an item across all people and should ideally lie within the range − 2.5 and + 2.5	0	50	50	30
	(2) Chi-square values summarize the difference between observed and expected responses to an item for groups (or 'class intervals') and should be associated non-significant *P* values (after Bonferroni correction). Item characteristic curves display this graphically	77	0	75	80
Are participants in the sample separated by the scale items?	Person separation index (PSI) ranges from 0 (all error) to 1 (no error). Higher scores indicate higher reliability	0.93	0.53	0.80	0.88

*All results presented in % success except for person separation index (range 0–1)

RESULTS

Sample

Data from 1584 patients ($n = 793$ from EMBODY1 and $n = 791$ from EMBODY2) were used in these analyses. The sample (mean [SD] age, 42 [12] years; range, 18–64 years; 93% female; 75% white) included patients from a broad geographic distribution (36% USA, 41% EU, and 23% rest of world) with time since diagnosis ranging between 0 and 38 years (mean [SD], 10 [7] years). For the responsiveness analysis, available data from 1203 patients were used ($n = 605$ from EMBODY1 and $n = 598$ from EMBODY2).

Stage 1: Measurement Performance Review of FACIT-F, SF-36 Vitality, and LupusQoL Fatigue Scales

The FACIT-F demonstrated adequate targeting, as item thresholds covered 68% of the range of fatigue measured in the sample (Table 1, column 3) while showing some item bunching (Fig. 1). In contrast, the SF-36 and LupusQoL Fatigue scales demonstrated suboptimal targeting: despite covering 63 and 71% of the fatigue measured in the sample, respectively, both scales showed gaps on the continuum, indicating areas on the metric where no scale item matched the levels of fatigue reported in the sample (Fig. 1).

The response scales for all three scales worked as intended; however, all scales demonstrated some item-fit issues (Table 1, columns 3–6). The worst-fitting items were from the FACIT-F and SF-36 Vitality; both scales demonstrated underestimation of fatigue (Fig. 2), as scores at the lower end were higher than expected and lower than expected at the higher end. In contrast, the worst-fitting LupusQoL Fatigue item displayed overestimation of fatigue with the opposite pattern of observed scores (Fig. 2). Fatigue person separation indices (PSIs) ranged from 0.80 to 0.93, suggesting the sample was sufficiently separated by their items; this is in comparison with the SF-36 Vitality scale, which demonstrated a low PSI (0.53), suggesting low reliability (Table 1, columns 3–6).

At the group level, all three fatigue scales showed a significant improvement of fatigue scores at week 24 ($P < 0.001$) with small and medium effect sizes (ES) and standardized response means (SRMs; Table 2, columns 9–12). The FACIT-F scale was the most responsive (ES = 0.35; SRM = 0.39) and the LupusQoL Fatigue the least responsive of the three scales (ES = 0.26; SRM = 0.24). At the individual level, the three different fatigue scales yielded different results regarding the percentage of patients reaching various degrees of improvement, worsening or no change in their fatigue levels at week 24, especially for the significant improvement and no change categories (Fig. 3). As assessed with the FACIT-F, 27% of patients reached significant improvement of fatigue scores at week 24 as opposed to 10% and 14% when assessed with the SF-36 Vitality and the LupusQoL Fatigue scales, respectively.

Stage 2: Construction and RMT Analysis of Pooled Fatigue Items

Construction of Item Pool

The RMT findings were reviewed in reference to item content of the three unique scales. Targeting findings suggested that the range of fatigue captured by the SF-36 and LupusQoL items did not cover the range of fatigue issues displayed in the sample and indicated low reliability for the SF-36 Vitality scale. Findings further demonstrated cohesiveness issues questioning the legitimacy of the total scores, particularly within the FACIT-F and SF-36 Vitality scales. The item content of these scales was closely reviewed to consider whether the conceptual content of these scales might mirror these statistical misfit issues. Some candidate problematic items were identified; these were eliminated from the item pool, resulting in the final selection of the ten pooled fatigue items (Fig. 4).

Items were selected on the basis of content clarity, quality, and conceptual relevance in relation to other fatigue items, in that items associated with the potential impact of fatigue

on daily activities or emotional consequences, or items associated with cognitive issues were excluded. Subsequently, items demonstrating misfit were also excluded whether these fit issues were hypothesized to be related to the content of the item, such as items confounding the symptoms of fatigue with its impact (e.g., frustration and social activities) or test-design issues. For example, some items were conceptually relevant to fatigue symptoms but were still associated with strong evidence of statistical misfit, such as the SF-36 item 'Did you have a lot of energy?' (Fig. 4). This finding could be attributed to test-design issues and the fact this was a positively worded item within an item pool of negatively worded items, which may have caused errors in the selected responses.

RMT Analysis of Pooled Fatigue Items and Comparison with Original Scales

The reconceptualized Fatigue Symptoms scale demonstrated adequate targeting and good reliability with a PSI of 0.88; but some fit issues persisted for two items, while one item had marginal problems with the five-point response scale (Table 1). Although item thresholds covered less of an absolute range of the sample (49%) in comparison to the original three fatigue scales (Table 1), the reconceptualized Fatigue Symptoms scale showed an improved item continuum with fewer gaps in comparison to the SF-36 Vitality and LupusQoL Fatigue scales (Fig. 1). In terms of item fit, the reconceptualized Fatigue Symptoms scale showed an improvement of statistical fit, especially in comparison to the FACIT-F and the SF-36 Vitality scales (Table 1).

In the group-level responsiveness analysis, the reconceptualized Fatigue Symptoms scale showed a significant improvement of fatigue at week 24 ($P < 0.001$) in line with all original scales (Table 2), but the reconceptualized scale was also associated with the highest ES (0.41) and SRM (0.44). The reconceptualized scale also had the highest relative efficacy, suggesting it was the most sensitive scale for detecting change in fatigue (Table 2).

At the individual level, different results with regard to the percentage of patients reaching significant or not improvement or worsening or no change in their fatigue levels at week 24 when assessed with the reconceptualized Fatigue Symptoms scale (Fig. 3) yielded the highest percentage of patients reaching significant improvement (28%) in comparison to the original scales.

DISCUSSION

Our psychometric evaluation of the FACIT-F, SF-36 Vitality, and LupusQoL Fatigue scales in the context of the EMBODY clinical trials provided mixed findings, challenging the extent to which these PRO scales are fit to quantify fatigue in a valid and reliable way in SLE. The pooled fatigue items, comprising a selection of the best-performing and conceptually clearest items from the three original scales, improved but did not resolve the identified measurement issues. Importantly, the pooled fatigue items systematically enhanced sensitivity in detecting changes in fatigue levels. This pooled item set was not put forward to propose a new fatigue scale, but rather to examine the impact of an item set that is psychometrically and conceptually more cohesive and clearer. This item set, therefore, was used to further elaborate upon some of the limitations of the reviewed scales and to illustrate potential initial steps that could be used to develop a new fatigue PRO. This exercise demonstrated the importance and value of a scale's conceptual underpinnings and clarity in the psychometric item design.

Findings from the RMT analysis revealed various issues. The FACIT-F demonstrated adequate targeting, indicating the relevance of the FACIT-F items in the population under measurement. However, fit analyses challenged the legitimacy of the FACIT-F total score. Strong evidence of statistical misfit was identified, suggesting the potential presence of multiple underpinning concepts within the scale's content. The qualitative review of the FACIT-F item content further indicated that the items covered fatigue symptoms, as well as the functional and emotional impact of fatigue, supporting the multiple conceptual underpinnings of the scale.

The SF-36 Vitality and LupusQoL Fatigue scales demonstrated sub-optimal targeting, with

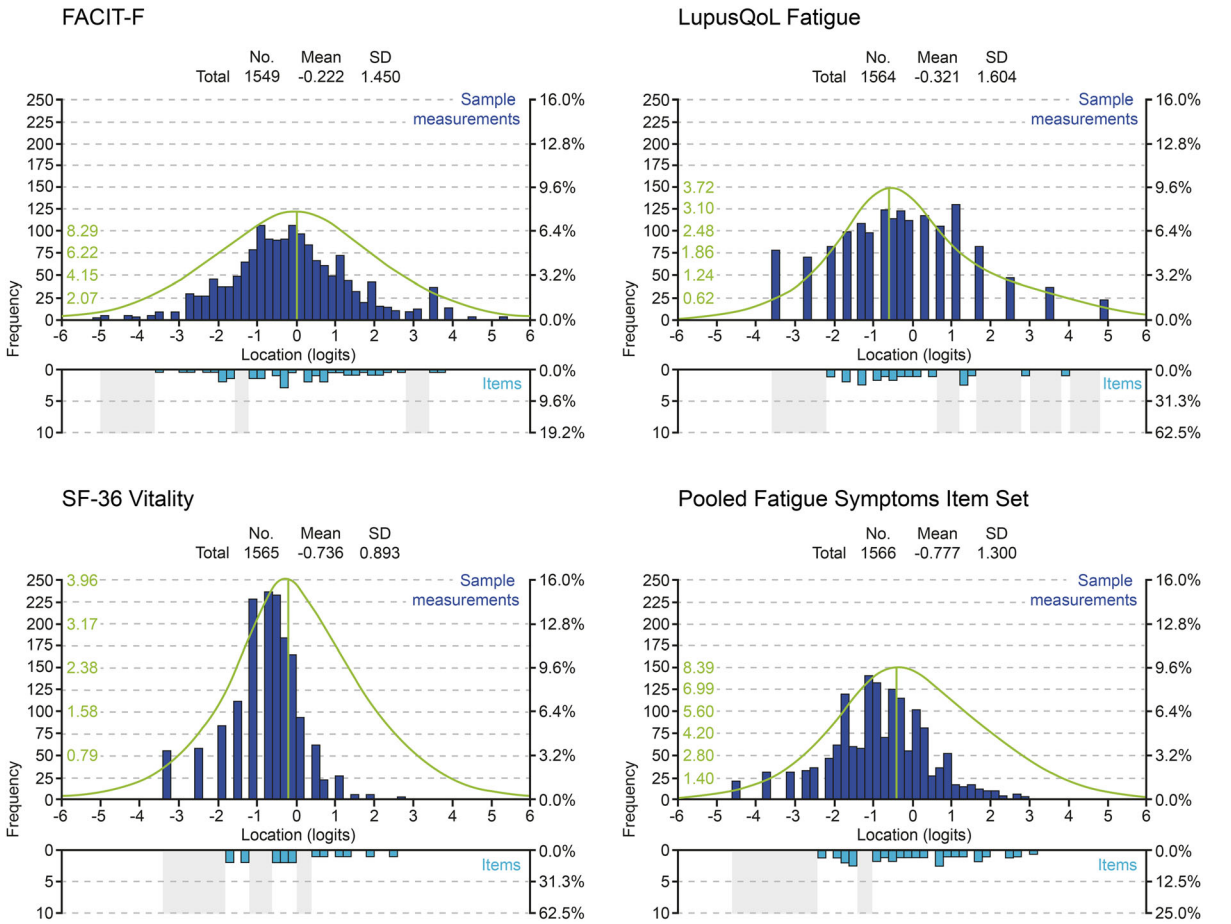

Fig. 1 Scale-to-sample targeting exemplars. The upper histograms (*dark blue bars*) represent the sample distribution for the scale total score whereas the lower histograms (*pale blue bars*) represent the scale item threshold distribution plotted on the same linear measurement continuum (higher scores reflect better outcomes/lower fatigue). The *green curve* represents an inverse function of the standard error associated with each person measurement (the peak of the curve indicating the best point of measurement). The *grey panels* on the lower histograms signify areas on the continuum with sample measurements but no corresponding item thresholds

findings indicating that the scales do not address all fatigue issues relevant in this population, accounting for the lack of precision associated with the scales' scores. Furthermore, fit analyses also indicated some issues with the scales' cohesiveness and reliability analysis, which challenged the SF-36 Vitality scale's ability to detect differences in the sample.

The pooled fatigue items were conceptually clearer and less ambiguous, and showed good psychometric properties including fit (especially compared to FACIT-F) and targeting (especially compared to SF-36 Vitality and LupusQoL Fatigue scales). In addition, although all of the scales demonstrated small to moderate improvements in fatigue scores at week 24, the pooled fatigue item set displayed the larger ES and SRMs at the group-level, and the highest percentage of 'significant improvers' on the individual level. The reconceptualized scale did not resolve all of the measurement issues. Of note, the FACIT-F demonstrated more optimal targeting, but this was probably due to its multidimensional content covering a wider range of HRQoL issues as opposed to it focusing on issues proximal to fatigue symptoms.

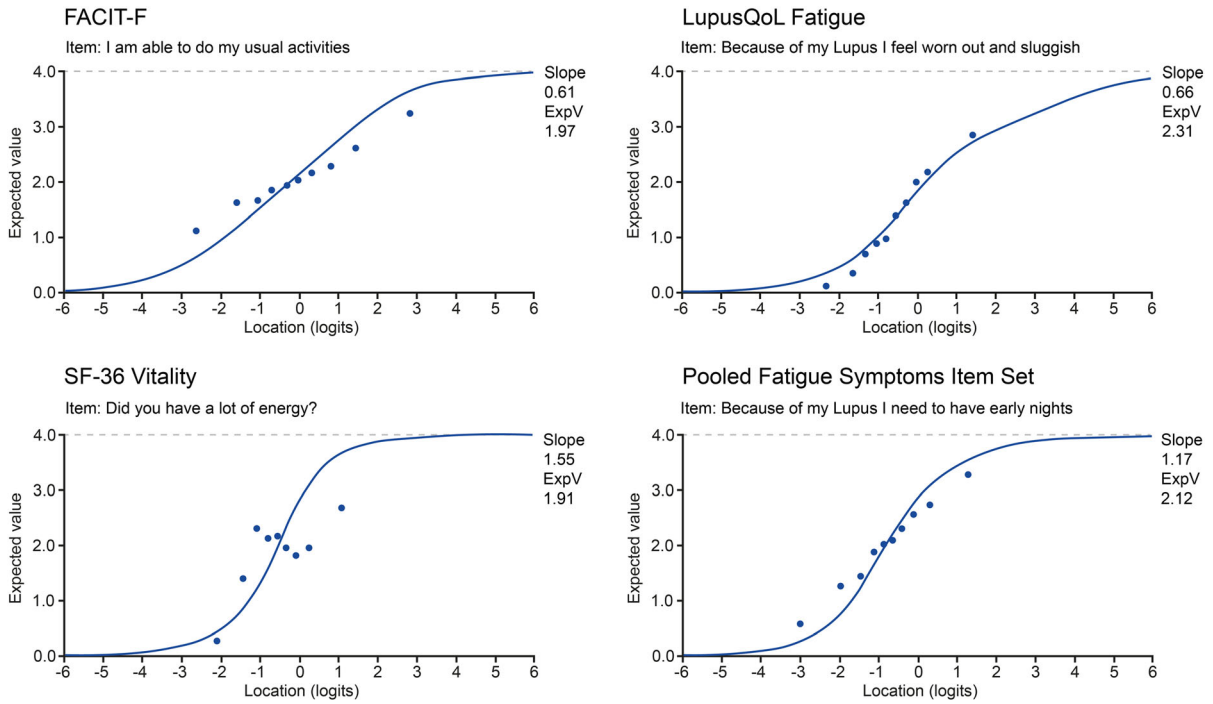

Fig. 2 Item characteristic curve (ICC) exemplars. The ICC plots the scores expected by the Rasch model for each individual item on the *y*-axis at each level of the fatigue measurement continuum (*x*-axis), with higher scores representing better outcomes/lower fatigue. The *blue dots* represent observed scores in each of the ten class intervals of the fatigue levels. The closer the blue dots (observed scores) lie to the curve (expected scores), the better the item fit for the item under investigation

However, the improvement of the pooled fatigue items sensitivity to measure clinical change, in comparison to the original scales, highlights the importance of a scale's conceptual underpinning.

Our psychometric analysis findings challenge the extent to which the three reviewed scales quantify fatigue in a reliable and valid way, and consequently call into question whether they should be used in high-stakes decision making in relation to SLE fatigue. Regardless of previously published quantitative psychometric evidence [23, 30–32], it is critical that a scale purporting to measure a clinical concept [24] is evaluated using both qualitative and psychometric methods, and specifically that the scale's content validity (Do the items reflect all relevant aspects of the COI?) and face validity (Do the items 'on their face' look like they measure the target COI?) are established.

It is important to acknowledge three limitations. First, the study constitutes a post hoc psychometric analysis of existing clinical trial data relating to a specific sample of patients with moderate to severe SLE. It would therefore be of value to replicate such analyses in further SLE samples to establish generalizability of findings. Second, and related to the first limitation, the EMBODY clinical trials were not designed for the purpose of this post hoc psychometrics analysis (e.g., sample size, power). However, the sample sizes (n = 158) would be considered adequate and power analysis is much less relevant for psychometric data analysis [51]. Thirdly, the responsiveness analyses were conducted on pooled blinded data preventing any comparisons between treatment arms from being made, but rather focusing on the relative sensitivity of the reviewed scales in detecting changes in fatigue levels. It is important to state that the three reviewed PRO scales

Table 2 Group-level responsiveness results

	N	Mean	SD	Mean change	SD change	*t*	*P*	RE	ES	SRM
FACIT-F (13 items)										
T1	1549	46.36	13.25	4.62	11.78	13.41	< .001	0.89	0.35	0.39
T2	1182	51.48	15.04							
SF-36 Vitality (4 items)										
T1	1565	45.57	13.18	4.12	15.15	9.39	< .001	0.62	0.31	0.27
T2	1200	50.06	16.87							
LupusQoL Fatigue (4 items)										
T1	1564	45.81	18.21	4.69	19.84	8.17	< .001	0.54	0.26	0.24
T2	1201	50.96	19.54							
Pooled Fatigue Symptoms Item Set (10 items)										
T1	1203	45.72	12.71	5.2	11.94	15.07	< .001	1.00	0.41	0.44
T2	1160	51.43	15.54							

T1 baseline, *T2* week 24, *ES* = effect size = mean change score divided by SD at T1, *SRM* = standardized response mean = mean change score divided by SD change, *RE* = relative efficiency = *t* statistic divided by largest *t* statistic value

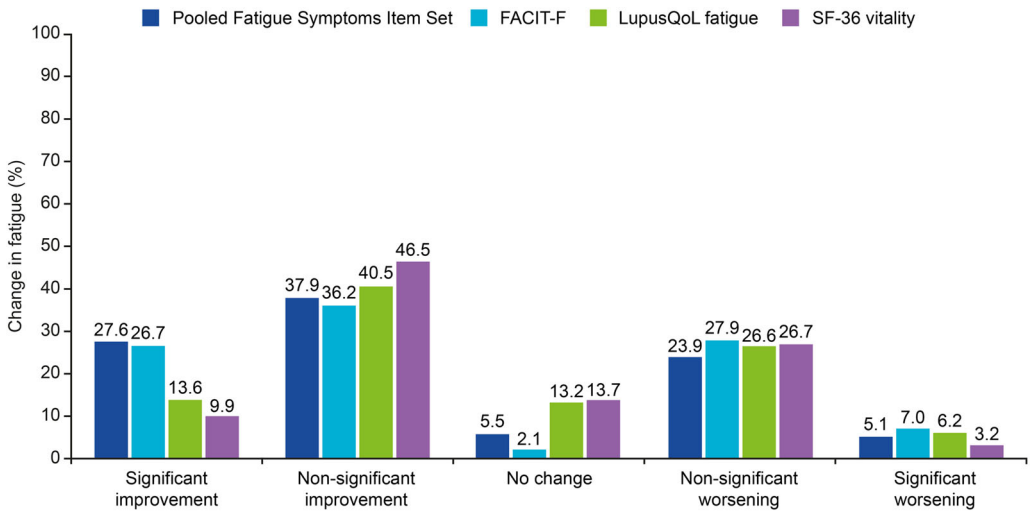

Fig. 3 Individual-level responsiveness. Percentage of patients displaying significant improvement, worsening, non-significant improvement, worsening or no change on the reviewed PRO instrument scales at week 24. Individual-level responsiveness was conducted using pooled blinded data in line with the methods described in Hobart & Cano 2009 (pages 151–152) [29]

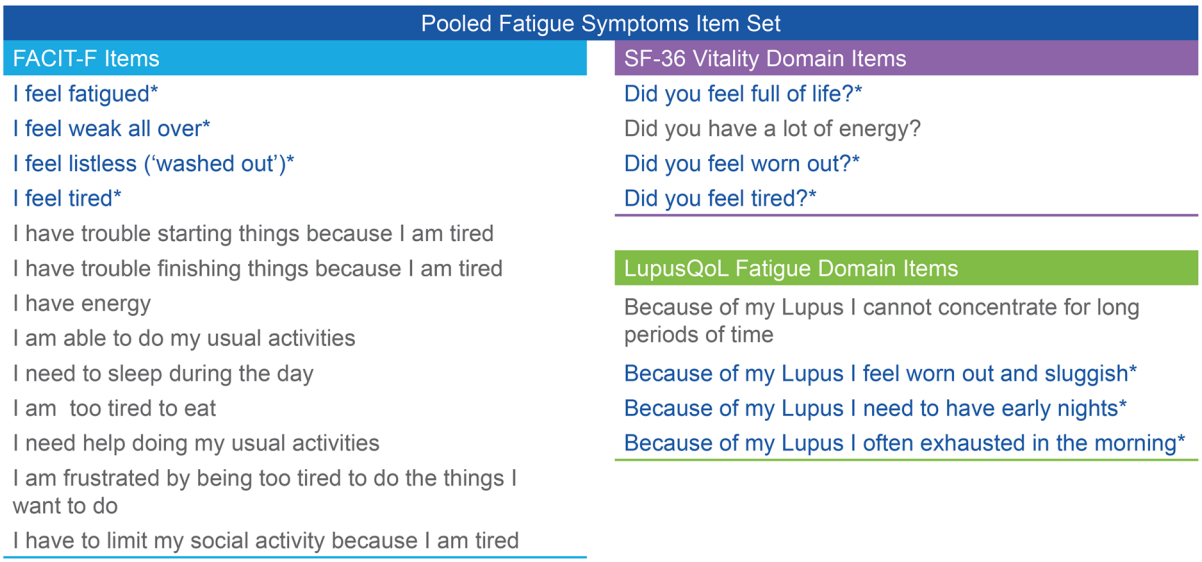

Fig. 4 Scale reconceptualization exemplar. Item-level content of the reviewed PRO instrument scales; *Items comprising the pooled Fatigue Symptoms Item Set

examined in this study were developed prior to regulatory guidelines, articulating the importance of clear definition and conceptualization of the construct under measurement in each context of use [4, 27]. Additionally, the FACIT-F and SF-36 were not developed specifically for use in SLE, while the SF-36 and LupusQoL, were not developed specifically to assess fatigue, as the reviewed scales constituted only one of multiple components within these PRO instruments. The conceptual underpinning of an item set used to quantify an underlying COI is of fundamental importance, particularly when it is used to make high-stake decisions affecting patients' treatment and care [4]. Without a clearly and comprehensively defined COI adequately reflected in the range of items within a scale leading to a standalone score, all subsequent quantitative psychometric evidence can be misleading [52].

CONCLUSIONS

Our study findings indicate shortcomings of the reviewed scales in quantifying fatigue, while the exploratory reconceptualized item set demonstrated the benefits of a concept-driven approach in improving the scales' measurement properties. Establishing a PRO scale that is fit for purpose to quantify fatigue in SLE will require thorough and robust exploration of the COI in the specific context of SLE, in order to create an appropriate conceptualization of fatigue to support a fatigue PRO scale content. As new treatments for SLE are developed and tested, developing a fit-for-purpose fatigue PRO for the SLE context of use will be vital for adequately quantifying patient fatigue, in order to evaluate potential treatments for one of the most important and relevant symptoms in SLE.

ACKNOWLEDGEMENTS

Funding. This study was funded by UCB Pharma, Belgium. The journal's Rapid Publication Fee was funded by UCB Pharma, Belgium.

Medical Writing, Editorial and other Assistance. The authors acknowledge Bengt Hoepken PhD, UCB Pharma, Germany, and Simone E. Auteri MSc EMS PhD, UCB Pharma, Italy, for critical review and Louise Barrett BSc, Modus Outcomes for supporting part of the analysis and write-up of this manuscript. The authors also acknowledge Sarah Jayne Clements, PhD,

Costello Medical, Cambridge, UK for editorial support in the development of this manuscript, funded by UCB Pharma. This study was funded by UCB Pharma, Belgium.

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Authors' Contributions. Substantial contributions to study conception and design: Sophie Cleanthous, Patrick Marquis, Stefan Cano, Thomas Moral, Christian Stach, and Sabine Bongardt; contributions to analysis and interpretation of the data: Sophie Cleanthous, Patrick Marquis, Stefan Cano, Thomas Morel, Christian Stach and Sabine Bongardt; drafting the article or revising it critically for important intellectual content: Sophie Cleanthous, Patrick Marquis, Stefan Cano, Thomas Morel, Christian Stach and Sabine Bongardt; final approval of the version of the article to be published: Sophie Cleanthous, Patrick Marquis, Stefan Cano, Thomas Morel, Christian Stach, and Sabine Bongardt.

Disclosures. Thomas Morel is an employee of UCB Pharma and owns UCB Pharma company stock awards. Christian Stach is an employee of UCB Pharma and owns UCB Pharma company stock awards. Sabine Bongardt is an employee of UCB Pharma and owns UCB Pharma company stock awards. Sophie Cleanthous, Patrick Marquis, and Stefan Cano are employees of Modus Outcomes, which received payment from UCB Pharma to conduct this research.

Compliance with Ethics Guidelines. The study protocol, amendments, and patient informed consent were reviewed by a national, regional, or Independent Ethics Committee (IEC) or Institutional Review Board (IRB). This study was conducted in accordance with the current version of the applicable regulatory and International Conference on Harmonisation (ICH)-Good Clinical Practice (GCP) requirements, the ethical principles that have their origin in the principles of the Declaration of Helsinki, and the local laws of the countries involved.

Data Availability. The datasets generated during and/or analyzed during the current study are not publicly available. Underlying data from this manuscript may be requested by qualified researchers 6 months after product approval in the US and/or Europe, or global development is discontinued, and 18 months after trial completion. Investigators may request access to anonymized IPD and redacted study documents which may include: raw datasets, analysis-ready datasets, study protocol, blank case report form, annotated case report form, statistical analysis plan, dataset specifications, and clinical study report. Prior to use of the data, proposals need to be approved by an independent review panel at www.Vivli.org and a signed data sharing agreement will need to be executed. All documents are available in English only, for a pre-specified time, typically 12 months, on a password protected portal.

REFERENCES

1. Lupus and Allied Diseases Association, Lupus Foundation of America, Lupus Research Alliance. Lupus: Patient Voices 2018. Available from: http://lupuspfdd.org/LupusPatientVoicesFINAL.pdf. Cited Jan 2021.

2. Cleanthous S, Tyagi M, Isenberg D, Newman S. What do we know about self-reported fatigue in systemic lupus erythematosus? Lupus. 2012;21(5):465–76.

3. Food and Drug Administration. Guidance for Industry. Systemic Lupus Erythematosus. 2010. Available from: https://www.fda.gov/media/71150/download. 19 Jan 2021.

4. Food and Drug Administration. Guidance for industry— - patient-reported outcome measures: use in medical product development to support labeling claims. 2009 [January 2021]. Available from: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf.

5. Food and Drug Administration. Roadmap to Patient-Focused Outcome Measurement in Clinical Trials 2013 [January 2021]. Available from: http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/UCM370174.pdf.

6. Strand V, Chu AD. Measuring outcomes in systemic lupus erythematosus clinical trials. Expert Rev Pharmacoecon Outcomes Res. 2011;11:455–68.

7. Clowse ME, Wallace DJ, Furie RA, Petri MA, Pike MC, Leszczyński P, et al. Efficacy and safety of epratuzumab in moderately to severely active systemic lupus erythematosus: results from two phase III randomized, double-blind, placebo-controlled trials. Arthritis Rheumatol. 2017;69(2):362–75.

8. Furie R, Petri MA, Strand V, Gladman DD, Zhong ZJ, Freimuth WW. Clinical, laboratory and health-related quality of life correlates of Systemic Lupus Erythematosus Responder Index response: a post hoc analysis of the phase 3 belimumab trials. Lupus Sci Med. 2014;1(1):e000031.

9. Khamashta M, Merrill JT, Werth VP, Furie R, Kalunian K, Illei GG, et al. Sifalimumab, an anti-interferon-α monoclonal antibody, in moderate to severe systemic lupus erythematosus: a randomised, double-blind, placebo-controlled study. Ann Rheum Dis. 2016;75(11):1909–16.

10. Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the functional assessment of cancer therapy (FACT) measurement system. J Pain Symptom Manag. 1997;13(2):63–74.

11. Berndt E, Kallich J, McDermott A, Xu X, Lee H, Glaspy J. Reductions in anaemia and fatigue are associated with improvements in productivity in cancer patients receiving chemotherapy. Pharmacoeconomics. 2005;23(5):505–14.

12. Ng AK, Li S, Recklitis C, Neuberg D, Chakrabarti S, Silver B, et al. A comparison between long-term survivors of Hodgkin's disease and their siblings on fatigue level and factors predicting for increased fatigue. Ann Oncol. 2005;16(12):1949–55.

13. Brucker PS, Yost K, Cashy J, Webster K, Cella D. General population and cancer patient norms for the functional assessment of cancer therapy-general (FACT-G). Eval Health Prof. 2005;28(2):192–211.

14. Mease PJ, Revicki DA, Szechinski J, Greenwald M, Kivitz A, Barile-Fabris L, et al. Improved health-related quality of life for patients with active rheumatoid arthritis receiving rituximab: results of the dose-ranging assessment: international clinical evaluation of rituximab in rheumatoid arthritis (DANCER) Trial. J Rheumatol. 2008;35(1):20–30.

15. Goligher EC, Pouchot J, Brant R, Kherani RB, Aviña-Zubieta JA, Lacaille D, et al. Minimal clinically important difference for 7 measures of fatigue in patients with systemic lupus erythematosus. J Rheumatol. 2008;35(4):635–42.

16. Ware JE, Kosinski M, Dewey JE. How to score version 2 of the SF-36 health survey. Lincoln: QualityMetric Incorporated; 2000.

17. Ware JE Jr, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. Med Care. 1992;30(6):473–83.

18. Ware JE, Robert BH, Davies AR, et al. Conceptualization and measurement of health for adults in the health insurance study: vol VIII overview. Santa Monica: RAND Corporation; 1979.

19. Strand V, Levy RA, Cervera R, Petri MA, Birch H, Freimuth WW, et al. Improvements in health-related quality of life with belimumab, a B -lymphocyte stimulator-specific inhibitor, in patients with autoantibody-positive systemic lupus erythematosus from the randomised controlled BLISS trials. Ann Rheum Dis. 2014;73(5):838–44.

20. Strand V, Petri M, Kalunian K, Gordon C, Wallace DJ, Hobbs K, et al. Epratuzumab for patients with moderate to severe flaring SLE: health-related quality-of-life outcomes and corticosteroid use in the randomized controlled ALLEVIATE trials and

extension study SL0006. Rheumatology (Oxford). 2014;53(3):502–11.

21. Medeiros MMC, Menezes APT, Silveira VA, Ferreira FNH, Lima GR, Ciconelli RM. Health-related quality of life in patients with systemic lupus erythematosus and its relationship with cyclophosphamide pulse therapy. Eur J Intern Med. 2008;19(2):122–8.

22. Grootscholten C, Snoek FJ, Bijl M, van Houwelingen HC, Derksen RHWM, Berden JHM, et al. Health-related quality of life and treatment burden in patients with proliferative lupus nephritis treated with cyclophosphamide or azathioprine/methylprednisolone in a randomized controlled trial. J Rheumatol. 2007;34(8):1699–707.

23. McElhone K, Abbott J, Shelmerdine J, Bruce IN, Ahmad Y, Gordon C, et al. Development and validation of a disease-specific health-related quality-of-life measure, the LupusQol, for adults with systemic lupus erythematosus. Arthritis Rheum. 2007;57(6):972–9.

24. Cano S, Hobart JC. The problem with health measurement. Patient Prefer Adherence. 2011;5:279–90.

25. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 1st ed. Oxford: Oxford University Press; 1987.

26. Stewart A, Ware J. Measuring functioning and well-being: the medical outcomes study approach. Durham: Duke University Press; 1992.

27. Food and Drug Administration. Qualification of Clinical Outcome Assessments (COAs). 2013 [January 2021]. Available from: https://www.urmc.rochester.edu/MediaLibraries/URMCMedia/neurology/documents/COAWheelSpokes_FDA.pdf.

28. Andrich D, Styles IM. Report on the psychometric analysis of the early development instrument (EDI) using the Rasch model. Perth: Murdoch University; 2004.

29. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess. 2009;13:1–200.

30. Conti F, Perricone C, Reboldi G, Gawlicki M, Bartosiewicz I, Pacucci VA, et al. Validation of a disease-specific health-related quality-of-life measure in adult Italian patients with systemic lupus erythematosus: LupusQoL-IT. Lupus. 2014;23:743–51.

31. Lai JS, Beaumont JL, Ogale S, Brunetta P, Cella D. Validation of the functional assessment of chronic illness therapy-fatigue scale in patients with moderately to severely active systemic lupus erythematosus, participating in a clinical trial. J Rheumatol. 2011;38(4):672–9.

32. Stoll T, Gordon C, Seifert B, Richardson K, Malik J, Bacon P, et al. Consistency and validity of patient administered assessment of quality of life by the MOS SF-36; its association with disease activity and damage in patients with systemic lupus erythematosus. J Rheumatol. 1997;24(8):1608–14.

33. Kosinski M, Gajria K, Fernandes AW, Cella D. Qualitative validation of the FACIT-fatigue scale in systemic lupus erythematosus. Lupus. 2013;22(5):422–30.

34. Jolly M, Pickard SA, Mikolaitis RA, Rodby RA, Sequeira W, Block JA. LupusQoL-US benchmarks for US patients with systemic lupus erythematosus. J Rheumatol. 2010;37(9):1828–33.

35. McElhone K, Castelino M, Abbott J, Bruce IN, Ahmad Y, Shelmerdine J, et al. The LupusQoL and associations with demographics and clinical measurements in patients with systemic lupus erythematosus. J Rheumatol. 2010;37(11):2273–9.

36. Devilliers H, Amoura Z, Besancenot JF, Bonnotte B, Pasquali JL, Wahl D, et al. Responsiveness of the 36-item 36-Item Short Form Health Survey and the Lupus Quality of Life questionnaire questionnaire in SLE. Rheumatology (Oxford). 2015;54(5):940–9.

37. Touma Z, Gladman DD, Ibanez D, Urowitz MB. Is there an advantage over SF-36 with a quality of life measure that is specific to systemic lupus erythematosus? J Rheumatol. 2011;38(9):1898–905.

38. Cleanthous S, Tyagi M, Isenberg D, Newman S. What do we know about self-reported fatigue in systemic lupus erythematosus? Lupus. 2012;21:465–76.

39. Arnaud L, Gavand PE, Voll R, Schwarting A, Maurier F, Blaison G, et al. Predictors of fatigue and severe fatigue in a large international cohort of patients with systemic lupus erythematosus and a systematic review of the literature. Rheumatology (Oxford). 2019;58(6):987–96.

40. DaCosta D, Drista M, Bernatsky S, et al. Dimensions of fatigue in systemic lupus erythematosus: relationship to disease status and behavioral and psychosocial factors. J Rheumatol. 2006;33(7):1282–7.

41. Tan EM, Cohen AS, Fries JF, Masi AT, Mcshane DJ, Rothfield NF, et al. The 1982 revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum. 1982;25(11):1271–7.

42. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification

of systemic lupus erythematosus. Arthritis Rheum. 1997;40(9):1725.

43. Yee CS, Farewell V, Isenberg DA, Griffiths B, Teh LS, Bruce IN, et al. The BILAG-2004 index is sensitive to change for assessment of SLE disease activity. Rheumatology (Oxford). 2009;48(6):691–5.

44. Gladman DD, Ibanez D, Urowitz MB. Systemic lupus erythematosus disease activity index 2000. J Rheumatol. 2002;29(2):288–91.

45. Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res. 2011;11(5):571–85.

46. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research; 1960.

47. Andrich D, Sheridan B. RUMM 2030. RUMM Laboratory Pty Ltd, Perth. 1997–2014.

48. Pallant J. SPSS survival manual. Maidenhead: Open University Press; 2010.

49. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum. 1985;28(5):542–7.

50. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care. 1990;28(7):632–8.

51. Hobart JC, Cano SJ, Warner TT, Thompson AJ. What sample sizes for reliability and validity studies in neurology? J Neurol. 2012;259(12):2681–94.

52. Hobart J, Cano S, Baron R, Thompson A, Schwid S, Zajicek J, et al. Achieving valid patient-reported outcomes measurement: a lesson from fatigue in multiple sclerosis. Mult Scler J. 2013;19(13):1773–83.