Quality of Life Research (2019) 28:3375–3384
https://doi.org/10.1007/s11136-019-02289-4

Psychometric properties of the PROMIS® Fatigue Short Form 7a among adults with myalgic encephalomyelitis/chronic fatigue syndrome

Manshu Yang[1,2] · **San Keller**[2] · **Jin-Mann S. Lin**[3]

Accepted: 27 August 2019 / Published online: 10 September 2019

Abstract

Purpose To evaluate the psychometric properties of the Patient-Reported Outcome Measurement Information System® Fatigue Short Form 7a (PROMIS F-SF) among people with Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS).

Methods Analyses were conducted using data from the Multi-Site Clinical Assessment of ME/CFS study, which recruited participants from seven ME/CFS specialty clinics across the US. Baseline and follow-up data from ME/CFS participants and healthy controls were used. Ceiling/Floor effects, internal consistency reliability, differential item functioning (DIF), known-groups validity, and responsiveness were examined.

Results The final sample comprised 549 ME/CFS participants at baseline, 386 of whom also had follow-up. At baseline, the sample mean of PROMIS F-SF T-score was 68.6 (US general population mean T-score of 50 and standard deviation of 10). The PROMIS F-SF demonstrated good internal consistency reliability (Cronbach's $\alpha = 0.84$) and minimal floor/ceiling effects. No DIF was detected by age or sex for any item. This instrument also showed good known-groups validity with medium-to-large effect sizes ($\eta^2 = 0.08$–0.69), with a monotonic increase of the fatigue T-score across ME/CFS participant groups with low, medium, and high functional impairment as measured by three different variables ($p < 0.01$), and with significantly higher fatigue T-scores among ME/CFS participants than healthy controls ($p < 0.0001$). Acceptable responsiveness was found with small-to-medium effect sizes (Guyatt's Responsiveness Statistic = 0.28–0.54).

Conclusions Study findings support the reliability and validity of PROMIS F-SF as a measure of fatigue for ME/CFS and lend support to the drug development tool submission for qualifying this measure to evaluate therapeutic effect in ME/CFS clinical trials.

Keywords Fatigue · Myalgic encephalomyelitis/chronic fatigue syndrome · Ceiling/floor effects · Internal consistency reliability · Differential item functioning · Known-groups validity · Responsiveness

Introduction

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is a debilitating long-term illness [1, 2]. ME/CFS affects multiple body systems and is characterized by at least 6 months of reduced ability to perform usual activities, accompanied by incapacitating fatigue experienced as profound exhaustion. Other symptoms include dizziness and problems with sleep, concentration, and memory. A distinctive characteristic of the illness is post-exertional malaise, which is a worsening of ME/CFS symptoms following physical or mental exertion that may require an extended recovery period [3]. The consequences of ME/CFS are devastating for individuals and their families and costly for society: those with moderate-to-severe disease are mostly housebound and may have lost many of their social roles; while those with very severe disease are mostly bedbound and cannot visit a provider for therapy [2]. Several population-based studies estimated

✉ Manshu Yang
myang@uri.edu

[1] Department of Psychology, University of Rhode Island, 142 Flagg Road, Kingston, RI 02881, USA

[2] American Institutes for Research, Chapel Hill, NC, USA

[3] Centers for Disease Control and Prevention, Atlanta, GA, USA

the prevalence of ME/CFS among the United States (US) population at more than 1 million [4–6]. Costs associated with ME/CFS are estimated at $18–$51 billion annually in the US ($9–$14 billion in medical costs and $9–$37 billion in lost productivity) with about one-quarter of medical expenses paid directly out-of-pocket [7–9].

Despite the significant impact on individuals, families, and society, there still are no FDA-approved treatments for ME/CFS. Clinical research approved by FDA requires valid, fit-for-purpose assessments to document change due to intervention (e.g., when used as primary or secondary endpoints); yet, such assessments for ME/CFS are not currently available. The variety of research and clinical case definitions for ME/CFS [1, 10–12] consistently recognize the importance of fatigue to the experience of the illness. Thus, the use of a fatigue measure as a drug development tool (DDT) for therapy for ME/CFS is important; yet, the choice of a fatigue measure for ME/CFS is challenged by lack of consensus regarding how fatigue should be assessed for patients with ME/CFS. We sought to meet this challenge by leveraging recent advances in the science of measuring symptoms of chronic illness.

The Patient-Reported Outcome Measurement Information System® (PROMIS®) is the product of a massive effort begun in 2004 and based on modern measurement theory, to address the need to develop precise, consensus measures of health outcome, including fatigue, suitable for use across chronic diseases. The PROMIS Fatigue Short Form 7a (PROMIS F-SF) is a seven-item questionnaire derived from the 95-item PROMIS Fatigue item bank [13] that assesses both the experience of fatigue and the interference of fatigue on daily activities over the past week [14]. Compared with other fatigue measures used in ME/CFS research, such as the Multidimensional Fatigue Inventory [15] and the Chalder Fatigue Scale [16], the key advantage of the PROMIS F-SF is its use of item response theory (IRT) to increase the precision and interpretability of scores while reducing respondents' burden. Specifically, IRT could improve the ability of scores to discriminate between various levels of fatigue. Moreover, IRT produces standardized scores that are readily comparable across different studies or patient populations, which help bring clinical meaning to the fatigue score for ME/CFS patients. The PROMIS F-SF can achieve evaluations of fatigue within a few minutes [17], which is particularly important to respondents with extremely limited energy, such as ME/CFS patients. However, its psychometric properties when used with ME/CFS patients, specifically, have not been documented. A concern about the use of a fatigue assessment as a measure of outcome for ME/CFS is that given the nature of the disease, there would be a restriction in the range of possible scores such that most patients would have the highest fatigue score possible. This could lead to poor evidence of reliability and validity of the scores.

In the present study, we evaluated the psychometric properties of the PROMIS F-SF to describe the fatigue of people suffering from ME/CFS using data from ME/CFS specialty clinics in the US. We aimed to obtain information that could be used in an application to the FDA to qualify the PROMIS F-SF as a DDT for use in clinical trials of therapy for ME/CFS. As described in FDA's guidance for clinical outcome assessment (COA) qualification [18], a patient-reported outcome instrument needs to be evaluated within the intended patient population in terms of test–retest reliability, internal consistency, content validity, construct validity (e.g., known-groups validity), and responsiveness. The current study provided information on all the required measurement properties except content validity and test–retest reliability.

Methods

Data source and study sample

Data were obtained from the *M*ulti-Site *C*linical *A*ssessment of *M*yalgic Encephalomyelitis/Chronic Fatigue Syndrome (MCAM) study [19]. The study was approved by the Institutional Review Boards of the Centers for Disease Control and Prevention, Open Medicine Institute Consortium, Mount Sinai Beth Israel, and Nova Southeastern University. One of the objectives for the MCAM study was to improve how ME/CFS symptoms could be measured. In brief, MCAM was conducted in multiple stages with a rolling cohort design. Not all participants were enrolled in the same stage; therefore, baseline data could come from any stage. Participants were recruited from seven ME/CFS specialty clinics across seven US states. The MCAM study relied on ME/CFS expert clinicians to determine patient eligibility using their clinical expertise with the condition. Enrolled patients were 18–70 years of age who had been diagnosed with CFS, ME, or post-infectious fatigue or who were managed as were other ME/CFS patients in the clinical practice.

Of 549 participants who completed the baseline assessment, 386 also completed the follow-up assessment approximately 10–14 months after baseline. Participants did not receive any intervention between their baseline and follow-up assessments. We used the baseline data from ME/CFS participants for most of the analyses, and data from participants who completed both baseline and follow-up assessments for evaluating responsiveness of the PROMIS F-SF over time. In addition, we used data from 304 healthy controls to examine known-groups validity.

Measures

PROMIS F-SF

The PROMIS F-SF was one of the instruments administered to participants of the MCAM study, via web-based platforms at five clinics and paper forms at the other two clinics. Prior research showed PROMIS scores to be robust to mode effects in comparisons of electronic to paper and pencil modes [20]. PROMIS F-SF includes seven items with response options on a 5-point Likert scale, ranging from 1 = "never" to 5 = "always." One item, "How often did you have enough energy to exercise strenuously," is reverse scored so that higher scores indicate greater fatigue. The PROMIS fatigue item bank was developed using a mixed-method approach and IRT methods [21]. Scores were on a T-score metric (mean = 50, standard deviation, SD = 10), so that one can compare the score to the national norm (e.g., 50 referred as the mean T-score of the US general population), matching the marginal distributions of gender, age, race, and education in the 2000 US Census [22].

In the validity analysis, we also included the following patient characteristics: Hours spent in vertical/horizontal activities per day, and physical health as measured by the SF-36 Health Survey (SF-36). We considered these three variables as proxy measures of functional impairment due to ME/CFS and used them to define groups of participants when evaluating known-groups validity. Data for these survey questions were collected at the same time as data for the PROMIS F-SF. The information for age and sex was abstracted from medical records and used in the differential item functioning (DIF) analysis.

Analyses

All IRT analyses were conducted using the IRTPRO software Version 4.2 [23] and other analyses using SAS 9.4 [24].

Descriptive statistics

ME/CFS patients often experience unusual fatigue that the general population does not experience. Therefore, generic fatigue measures might be more likely than condition-specific measures to have a restricted range of measurement among ME/CFS patients. In other words, ME/CFS patients may be more likely than patients with other chronic conditions, to select the response option that indicates the greatest fatigue for all items in the measure. For each item from the PROMIS F-SF, we calculated univariate descriptive statistics including the mean, standard deviation, and frequency distribution. In addition, we examined the proportion of ME/CFS participants at the lowest or highest possible score for the entire measure. We adopted a standard of less than 15% of respondents at either the highest or lowest fatigue score as our standard for evaluating the suitability of the PROMIS F-SF for ME/CFS [25, 26].

IRT scoring

We applied IRT models to score the PROMIS F-SF. T-scores were calculated for both ME/CFS participants and healthy controls at baseline and at follow-up.

Reliability of scores

We estimated the reliability of PROMIS F-SF raw sum scores by calculating internal consistency coefficients (Cronbach's alpha) and by calculating the amount of measurement error in the T-score under IRT. It is generally recommended to use scales with reliability coefficients of 0.70 and above for group-level analyses and 0.90 and above for tracking individual patients [27, 28]. These two coefficients are equivalent to a T-score standard error of measurement (SEM) of 5.5 and 3.2, respectively. Previous research has shown IRT-based reliability statistics for the PROMIS F-SF scores which exceed 0.90 and T-score standard errors < 3.2 for people with a wide range of fatigue scores [13]. We hypothesized that, consistent with what has been observed in patients with other chronic diseases [29, 30], *the PROMIS F-SF scores of the ME/CFS participants would have reliability estimates exceeding those recommended for group-level comparisons (i.e., Cronbach's alpha > 0.70 and the average standard error for the T-score < 5.5).*

DIF analysis

DIF signifies that the item provides biased measurement across population subgroups. Evidence of DIF occurs when the likelihoods of endorsing an item response category are different across respondent subgroups that are matched on the underlying trait being measured. Such a difference suggests that the variable used to define population subgroups, other than the trait being measured, had an influence on the item responses.

We examined the possibility of DIF by sex and age for each PROMIS F-SF item using the Wald test [31, 32]. For age, we compared the measurement properties of PROMIS F-SF items across three groups: 18–39, 40–59, and 60 or above. Patients ages 40–59 were used as the reference group since ME/CFS is more prevalent in those ages [5, 33]. This categorization allowed us to have a sufficiently large number of participants in each group to examine DIF. For the Wald test, a non-significant χ^2 value indicates no detectable DIF. We used the Benjamin–Hochberg [34, 35] procedure to control for the multiple comparisons involved in checking DIF for each item by sex and age. *We hypothesized that there* *would be no evidence of DIF for PROMIS F-SF items in this sample of ME/CFS participants.*

Known-groups validity

Known-groups validity of the PROMIS F-SF was evaluated by comparing T-scores for groups that are known to differ in their fatigue levels. *We hypothesized that ME/CFS participants with higher levels of functional impairment would have PROMIS F-SF scores indicating greater fatigue*; *and that ME/CFS participants would exhibit greater fatigue than healthy controls.* Initially, ME/CFS participants were classified into "low", "medium", and "high" level of functional impairment, based on each of the following variables:

- *Hours spent in vertical activities per day* Participants reported the average time with feet on the floor (e.g., sitting, standing, or walking). Reported hours could be up to 15 and fewer hours indicate more ME/CFS functional impairment. The median of vertical hours reported by ME/CFS participants and healthy controls was 7 (range 0–15) and 14 (range 0–15), respectively.
- *Hours spent in horizontal activities per day* Participants reported the average time with feet up (e.g., resting in recliner with feet up, napping, sleeping in bed). Reported hours could be up to 15 and more hours indicate more ME/CFS functional impairment. The median of horizontal hours reported by ME/CFS participants and healthy controls was 14 (range 0–15), and 8 (range 0–15), respectively.
- *Physical Health* Overall physical health was measured by Physical Component Summary (PCS) T-scores of SF-36 (lower scores indicate more ME/CFS functional impairment).

Details about the variables for vertical/horizontal hours can be accessed in the supplement materials [19]. For each variable, we used two tertiles to split data into three approximately equal-sized groups of ME/CFS participants, so that each group has adequate sample size for validity analyses. Since over 46% of ME/CFS participants were at the maximum of 15 h of horizontal activities per day, we were unable to divide participants into three equal-sized groups using that variable. Instead, we divided participants into two groups based on horizontal activity level (15 h vs. < 15 h) to ensure adequate sample size in each group.

Analysis of variance (ANOVA) was used to examine mean differences in fatigue T-scores among known-groups defined by the three aforementioned variables for ME/CFS participants, and between ME/CFS participants and healthy controls. We used the Tukey–Kramer method [36] for multiple comparison adjustment among known-groups. Known-groups validity was considered acceptable when the difference in fatigue scores across groups was observed at a significance level of 0.05. We reported means and standard deviations for the fatigue scores by group along with effect sizes, and test statistics. Using the ANOVA results, we computed η^2 by dividing the sum of squares for the known-groups effect by the total sum of squares. We used the η^2 statistic as an effect-size measure and interpreted such effects as small (0.01), medium (0.09), and large (0.25), following convention [37, 38].

Responsiveness

Responsiveness represents an instrument's ability to detect changes over time when true changes exist. *We hypothesized that ME/CFS participants with "improved", "stable", and "worsened" health would show a significant decrease, no significant changes, and a significant increase in their PROMIS F-SF fatigue scores, respectively.* We initially categorized ME/CFS participants into three groups of "improved", "stable," and "worsened" using the following measures of functional status:

- *Hours vertical* Change in number of hours of vertical activities per day from baseline to follow-up: improved = "≥ 1-h increase in vertical activities"; stable = "within 1-h"; worsened = "≥ 1-h decrease in vertical activities".
- *Hours horizontal* Change in number of hours of horizontal activities per day from baseline to follow-up: improved = "≥ 1-h decrease in horizontal activities"; stable = "within 1-h"; worsened = "≥ 1-h increase in horizontal activities".
- *Physical health* Change in SF-36 PCS T-score from baseline to follow-up: improved = "score increase > 5″; stable = "score increase or decrease ≤ 5"; worsened = "score decrease > 5".

For vertical/horizontal activities, there was no established cutoff for the minimal clinically important difference (MCID). Because time spent in these activities was reported in integer hours, we considered patients as stable if the change from baseline to follow-up was within an hour. For SF-36 PCS T-scores, we identified a change of 5 as the MCID and considered patients as stable if the change from baseline to follow-up was ≤ 5. Previously published MCIDs of SF-36 PCS across different studies range from 2.5 to about 7 [39–41]. To our knowledge, the MCID of SF-36 PCS has not been studied for ME/CFS. Therefore, we used the half standard deviation approach [42] and chose an MCID of 5 for assessing the responsiveness among ME/CFS participants.

About 50% of ME/CFS participants reported the worst possible functional status at baseline on any of the three aforementioned criterion variables. Among those participants, it was impossible to distinguish those whose functional status declined from those whose functional status remained stable: some of the participants who appeared to be stable (e.g., reporting 15 h of horizontal activities at both baseline and follow-up) might have experienced a worsening in their functioning that was undetectable. Thus, we combined the "stable" and "worsened" groups into a group of "not improved."

We used ANOVA to examine if changes in fatigue scores significantly differed among groups that we defined as "improved," "stable", and "worsened" or between the "improved" and "not improved" groups. We report means and standard deviations of fatigue change scores for each group. We calculated Guyatt's responsiveness statistic (GRS) as an effect size comparing the "improved" groups to the "stable," "worsened," or "not improved" groups. The GRS is defined as the mean change in fatigue scores for the target group (i.e., "improved") divided by the standard deviation of the comparison group (e.g., "not improved") [43] and is interpreted as small (≥ 0.2 and < 0.5), medium (≥ 0.5 and < 0.8), and large (≥ 0.8) [37].

Results

The majority of ME/CFS participants were female (74.3%), white (92.5%), and not currently working (69.6%) (Table 1). Their mean age was 48.1 years with the illness duration of 13.5 years.

Item descriptive statistics

Table 2 summarizes the item responses of ME/CFS participants for each PROMIS F-SF item. The percentage of participants responding "5" (representing the highest level of fatigue) ranged from 6.4 to 72.9% across the seven items, indicating that these items cover a wide range of fatigue level.

Table 2 also shows the number and percentage of ME/CFS participants who selected the greatest-fatigue category for all items in the measure or selected the least-fatigue category for all items in the measure, along with the raw sum score means and standard deviations. Only a very small proportion of participants were at the greatest fatigue score (1.8%) and almost no participants were at the lowest fatigue score (0.4%), suggesting that floor/ceiling effect was not a concern.

When using the IRT approach, the mean T-score of the PROMIS F-SF at 68.6 in ME/CFS participants was nearly

Table 1 Sample characteristics at baseline

	ME/CFS ($n=549$)[a]		Healthy control ($n=304$)[a]	
	n	%	*n*	%
Female sex	408	74.3	200	65.8
Race				
White	508	92.5	176	57.9
Black/African American	11	2.0	18	5.9
All others	25	4.6	76	25.0
Marital status				
Married/committed	293	53.4	151	49.7
Previously married	92	16.8	54	17.8
Never married	150	27.3	91	29.9
Employment				
Full-time	84	15.3	162	53.3
Part-time	60	10.9	54	17.8
Not working	382	69.6	77	25.3
	Mean	SD	Mean	SD
Age, years	48.1	12.8	43.8	14.9
Illness duration, years	13.5	9.7	NA	NA

SD standard deviation

[a]Frequency numbers do not add up to the total number of participants and percentages do not add up to 100% because of missing data

Table 2 Item-level descriptive statistics and measure-level raw sum score distributions for PROMIS F-SF, based on ME/CFS participant data at baseline ($n = 549$) from the MCAM study

Item	Mean	SD	Response of "1" (lowest level of fatigue)		Response of "5" (highest level of fatigue)	
			n	%	*n*	%
PROMIS1: feel tired	4.54	0.68	3	0.5	339	61.7
PROMIS2: extreme exhaustion	3.70	0.93	13	2.4	99	18.0
PROMIS3: out of energy	4.16	0.79	4	0.7	192	35.0
PROMIS4: limit work	4.39	0.84	7	1.3	303	55.2
PROMIS5: think clearly	3.60	0.87	15	2.7	58	10.6
PROMIS6: bath/shower	2.98	1.13	75	13.7	35	6.4
PROMIS7: strenuous exercise	4.51	1.00	23	4.2	400	72.9
			Participants at the lowest possible score of fatigue		Participants at the highest possible score of fatigue	
PROMIS F-SF (raw sum score 7–35)	27.6	4.7	2	0.4	10	1.8

SD standard deviation

two standard deviations above the national norm, supporting the validity of the PROMIS F-SF for ME/CFS.

Reliability

The PROMIS F-SF showed high internal consistency (Cronbach's $\alpha = 0.84$) among ME/CFS participants. When using the IRT approach, we found the average standard error of fatigue T-score was 2.9 and well below the hypothesized threshold of 5.5 that corresponds to a reliability of 0.7, indicating high precision of the PROMIS F-SF scores for ME/CFS.

DIF

Table 3 shows results from the Wald test for DIF detection by sex and by age. Before using the Benjamini–Hochberg correction for multiplicity, one PROMIS F-SF item ("think clearly") showed possible DIF by sex, with a p value < 0.05. However, after correction for multiplicity, none of the items exhibited significant DIF by sex. None of the items exhibited significant DIF by age, even before correcting for multiplicity.

Known-groups validity

Results in Table 4 show that the omnibus hypothesis of no differences among known groups was rejected with $p < 0.0001$, providing supportive evidence for the validity of the PROMIS F-SF in ME/CFS. Furthermore, we found significant differences for all pairwise comparisons between groups with a monotonic increase in mean scores across low, medium, and high functional impairment groups of ME/CFS participants. ME/CFS participants had significantly higher mean scores than healthy controls. Effect sizes, η^2, were medium for known-groups defined by vertical activities and

Table 3 Differential item functioning (DIF) statistics by sex and age, based on baseline ME/CFS participant data ($n = 549$) from the MCAM study

Label	DIF by sex Male vs. female			DIF by age 18–39 vs. 40–59			DIF by age 60 + vs. 40–59		
	χ^2	d*f*	*p*	χ^2	d*f*	*p*	χ^2	d*f*	*p*
PROMIS F-SF									
PROMIS1: feel tired	1.9	5	0.860	4.1	4	0.394	4.5	4	0.344
PROMIS2: extreme exhaustion	1.7	5	0.889	1.1	5	0.958	5.3	5	0.379
PROMIS3: out of energy	1.5	5	0.914	0.4	4	0.985	4.8	4	0.308
PROMIS4: limit work	3.1	5	0.688	1.2	4	0.882	7.1	4	0.133
PROMIS5: think clearly	14.9	5	0.011	5.4	5	0.365	2.8	5	0.730
PROMIS6: bath/shower	6.9	5	0.233	6.3	5	0.283	3.8	5	0.580
PROMIS7: strenuous exercise	8.7	5	0.121	9.6	5	0.087	6.7	5	0.247

Table 4 Mean fatigue scores by three indicators of functional impairment level as well as between ME/CFS participants and healthy controls from the MCAM study

Known groups	*n*	PROMIS F-SF			
		Mean	SD	*F*	η^{2a}
ME/CFS functional impairment level defined by hours of vertical activities per day[b]					
Low (≥ 10 h)	190	65.5	8.4	34.4	0.115
Medium (≥ 5 and < 10 h)	181	68.9	5.9		
High (< 5 h)	164	71.5	5.7		
ME/CFS functional impairment level defined by hours of horizontal activities per day[c]					
Lower (< 15 h)	286	66.6	8.0	45.8	0.079
Higher (15 h)	250	70.7	5.5		
ME/CFS functional impairment level defined by SF-36 PCS score[d]					
Low (T-score ≥ 28.78)	180	63.9	8.0	88.8	0.247
Medium (20.49 ≤ T-score < 28.78)	184	69.3	5.6		
High (T-score < 20.49)	179	72.7	4.8		
ME/CFS participants vs. healthy controls[c]					
ME/CFS participants	549	68.6	7.2	1915.0	0.692
Healthy controls	304	44.8	8.2		

SD standard deviation

[a] η^2 is an effect-size measure and was computed by dividing the sum of squares for the known-groups effect by the total sum of squares

[b] Overall differences among the three groups were significant at $p < 0.0001$; group differences of low vs. medium and low vs. high functional impairment were significant at $p < 0.0001$; group differences of medium vs. high functional impairment were significant at $p < 0.01$

[c] Differences were significant at p < 0.0001

[d] Group differences, overall and between any two functional impairment levels, were significant at $p < 0.0001$

horizontal activities, and large for known-groups defined by SF-36 PCS scores and when comparing ME/CFS participants to healthy controls.

Responsiveness

The omnibus hypothesis of no differences among groups of ME/CFS participants defined by whether they improved, remained stable, or worsened was rejected with $p < 0.05$. In pairwise comparisons between the groups, however, no significant difference was found between the "stable" and "worsened" groups. This is probably due to: (1) the inability to differentiate between "stable" and "worsened" for a large proportion of participants who remained at the maximum hours of horizontal activity from baseline to follow-up, and (2) the small sample size for the "worsened" group based on SF-36 PCS change score (11%, $n = 43$). Due to space limitations, detailed results comparing the "improved", "stable", and "worsened" groups are omitted here.

In Table 5, we show the results comparing change for "improved" vs. "not improved" groups. Decreases in fatigue scores were significantly larger in the "improved" group than in the "not improved" group ($p < 0.05$). Guyatt's responsiveness statistics were small to medium. These results provide supportive evidence of responsiveness for PROMIS F-SF in ME/CFS.

Discussion

The PROMIS F-SF showed strong reliability and validity to assess fatigue outcome for ME/CFS participants by demonstrating good internal consistency, minimal to almost no floor/ceiling effects at the total score level, good known-groups validity with medium-to-large effect sizes, and acceptable responsiveness to change with small-to-medium

Table 5 Mean changes in fatigue scores between "improved" and "not improved" groups defined by three anchor measures from the MCAM study

Change status	*n*	PROMIS F-SF				
		Mean	SD	*F*	*p*	Guyatt's responsiveness statistic
Change in hours of vertical activities per day, from baseline to follow-up						
Improved	135	− 1.73	5.96	8.3	0.004	− 0.318
Not improved	240	0.04	5.57			
Change in hours of horizontal activities per day, from baseline to follow-up						
Improved	100	− 1.72	6.52	5.1	0.024	− 0.280
Not improved	275	− 0.20	5.42			
Change in SF-36 PCS score, from baseline to follow-up						
Improved	98	− 2.70	6.30	19.0	< .0001	− 0.536
Not improved	276	0.19	5.41			

SD standard deviation

effect sizes. Consistent with our hypotheses, the PROMIS F-SF had internal consistency reliability well above the acceptable threshold of 0.70. As expected, the fatigue T-score had a monotonic increase across ME/CFS participant groups with low, medium, and high functional impairment, and ME/CFS participants had significantly higher fatigue scores than healthy controls. Moreover, as expected, ME/CFS participants who experienced improvement in their functional status over time showed a significant decrease in their PROMIS F-SF fatigue scores compared with those experiencing no improvement. No DIF was detected by age or sex, suggesting that all items provide unbiased measurement across these population subgroups. All findings were in support of qualifying the PROMIS F-SF as a DDT for use in clinical trials of therapy for ME/CFS. To meet FDA's requirement for COA qualification, we will further examine the content validity and test–retest reliability of the PROMIS F-SF in future research.

Since ME/CFS participants commonly have much higher fatigue levels than the general US population, we originally expected that quite a few participants, although no more than 15%, would select the greatest-fatigue category for all items in the PROMIS F-SF. However, results suggest that the PROMIS F-SF covered a broad range of fatigue levels and exhibited minimal restriction on the measurement range within this high-fatigue patient population. The PROMIS F-SF is a generic (i.e., not condition-specific) measure that has been tested in various patient populations (e.g., multiple sclerosis, fibromyalgia, sickle cell disease, cardiometabolic risk, pregnancy) [44, 45]. It provides a good opportunity for researchers and clinicians to use a single instrument that allows comparisons within ME/CFS participants with varying levels of functioning impairment as well as comparisons of ME/CFS participants and patients with other fatiguing illnesses.

There are a few limitations to our research which should be considered when interpreting results from this study, and some of them suggest future research directions. First, all participants from the MCAM study were receiving tertiary care, thus did not match the composition of the broad U.S. ME/CFS population. On the other hand, ME/CFS participants receiving tertiary care from a clinic are likely to be the population recruited for clinical trials of therapy in ME/CFS. Although we believe our large sample size recruited from seven clinics across seven states provides a useful population for validity analysis, future studies using more diverse samples should be conducted to evaluate the stability of parameter estimates.

Second, the known-group validity analysis for responsiveness to change was compromised by restricted range of measurement of functional status. That is, a large proportion of ME/CFS participants began the study with the worst possible functional impairment. For such participants, we were unable to distinguish those who were truly stable from those who experienced an undetectable worsening in functional status. Other ME/CFS severity indicators with minimal floor/ceiling effects need to be identified to better define participants with stable functional status over time. Such indicators would facilitate further analyses on responsiveness and the evaluation of test–retest reliability among stable participants.

Third, some researchers may argue that, compared to a condition-specific measure, the generic PROMIS F-SF does not include items reflecting unique fatigue symptoms in ME/CFS, thus reducing its content validity in this patient population. Patients with chronic fatigue did take part in focus groups conducted to evaluate the content of the PROMIS Fatigue measure, but the data from those focus groups were not tagged by diagnosis, so it is not possible to distinguish the comments of participants with ME/CFS from those of participants with other chronic conditions [46]. Cognitive debriefing studies of the PROMIS F-SF could be conducted with ME/CFS patients to obtain additional content validity evidence for those items in that specific patient population.

Despite these limitations, the results of this study contribute unique information about the reliability and validity of the PROMIS F-SF in general and when used in patients with ME/CFS in particular. This information will be useful to those selecting measures of fatigue for ME/CFS going forward.

Conclusion

Study findings support the reliability and validity of PROMIS F-SF as a measure of fatigue for ME/CFS and lend support to the drug development tool submission for qualifying this measure to evaluate therapeutic effect in ME/CFS clinical trials, which is an important step toward developing FDA-approved treatment for this debilitating illness.

Acknowledgements We would like to thank the Multi-Site Clinical Assessment of ME/CFS (MCAM) study group who made the data sets available through Research Collaboration Agreement. The group included: Division of High-Consequence, Pathogens and Pathology, Centers for Disease Control and Prevention, Atlanta, Georgia: Elizabeth R. Unger, Britany Helton, Yang Chen, and Monica Cornelius; Open Medicine Institute Consortium, Mountain View, CA: Andreas Kogelnik, Catt Phan, Joan Danver, Lucinda Bateman, Jennifer Bland, Charles Lapp, Wendy Springs, Richard Podell, Trisha Fitzpatrick, Daniel Peterson, and Marco Maynard; Institute for Neuro Immune Medicine, Miami, Florida: Nancy Klimas, Elizabeth Balbin, Precious Leaks-Gutierrez, and Shuntae Parnell; and Mount Sinai Beth Israel, New York, New York: Benjamin Natelson and Diana Vu.

Disclaimer The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Funding Manshu Yang (MY) and San Keller (SK) were supported by a National Institutes of Health (NIH) Grant Award (U2CCA186878). Jin-Mann S. Lin (JML) was supported by the Centers for Disease Control and Prevention (CDC).

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

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

References

1. Clayton, E. W. (2015). Beyond myalgic encephalomyelitis/chronic fatigue syndrome: An IOM report on redefining an illness. *JAMA, 313*(11), 1101–1102. https://doi.org/10.1001/jama.2015.1346.
2. United States Food and Drug Administration. (2013). *The voice of the patient: Chronic fatigue syndrome and myalgic encephalomyelitis*. Bethesda: United States Food and Drug Administration.
3. Afari, N., & Buchwald, D. (2003). Chronic fatigue syndrome: A review. *American Journal of Psychiatry, 160*(2), 221–236. https://doi.org/10.1176/appi.ajp.160.2.221.
4. Jason, L. A., Richman, J. A., Rademaker, A. W., Jordan, K. M., Plioplys, A. V., Taylor, R. R., et al. (1999). A community-based study of chronic fatigue syndrome. *Archives of Internal Medicine, 159*(18), 2129–2137.
5. Reyes, M., Nisenbaum, R., Hoaglin, D. C., et al. (2003). Prevalence and incidence of chronic fatigue syndrome in Wichita, Kansas. *Archives of Internal Medicine, 163*(13), 1530–1536. https://doi.org/10.1001/archinte.163.13.1530.
6. Reeves, W. C., Jones, J. F., Maloney, E., Heim, C., Hoaglin, D. C., Boneva, R. S., et al. (2007). Prevalence of chronic fatigue syndrome in metropolitan, urban, and rural Georgia. *Popul Health Metr, 5,* 5. https://doi.org/10.1186/1478-7954-5-5.
7. Lin, J. S., Resch, S. C., Brimmer, D. J., Johnson, A., Kennedy, S., Burstein, N., et al. (2011). The economic impact of chronic fatigue syndrome in Georgia: Direct and indirect costs. *Cost Effectiveness and Resource Allocation, 9*(1), 1. https://doi.org/10.1186/1478-7547-9-1.
8. Reynolds, K. J., Vernon, S. D., Bouchery, E., & Reeves, W. C. (2004). The economic impact of chronic fatigue syndrome. *Cost Effectiveness and Resource Allocation, 2,* 4. https://doi.org/10.1186/1478-7547-2-4.
9. Jason, L. A., Benton, M. C., Valentine, L., Johnson, A., & Torres-Harding, S. (2008). The economic impact of ME/CFS: Individual and societal costs. *Dynamic Medicine, 7,* 6. https://doi.org/10.1186/1476-5918-7-6.
10. Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., & Komaroff, A. (1994). The chronic fatigue syndrome: A comprehensive approach to its definition and study International. Chronic Fatigue Syndrome Study Group. *Annals of Internal Medicine, 121*(12), 953–959.
11. Carruthers, B. M., Jain, A. K., De Meirleir, K. L., Peterson, D. L., Klimas, N. G., Lerner, A. M., et al. (2003). Myalgic encephalomyelitis/chronic fatigue syndrome: Clinical working case definition, diagnostic and treatment protocols. *Journla of Chronic Fatigue Syndrome, 11*(1), 7–115.
12. Carruthers, B. M., van de Sande, M. I., De Meirleir, K. L., Klimas, N. G., Broderick, G., Mitchell, T., et al. (2011). Myalgic encephalomyelitis: International consensus criteria. *Journal of Internal Medicine, 270*(4), 327–338. https://doi.org/10.1111/j.1365-2796.2011.02428.x.
13. Lai, J.-S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation, 92*(10.0), S20–S27. https://doi.org/10.1016/j.apmr.2010.08.033.
14. National Institutes of Health (2007). PROMIS domain framework/definitions. http://www.nihpromis.org/measures/domainframework.
15. Smets, E. M. A., Garssen, B., Bonke, B., & De Haes, J. C. J. M. (1995). The Multidimensional Fatigue Inventory (MFI) Psychometric Qualities of an Instrument to Assess Fatigue. *Journal of Psychosomatic Research, 39*(3), 315–325. https://doi.org/10.1016/0022-3999(94)00125-O.
16. Chalder, T., Berelowitz, G., Pawlikowska, T., Watts, L., Wessely, S., Wright, D., et al. (1993). Development of a fatigue scale. *Journal of Psychosomatic Research, 37*(2), 147–153.
17. Khanna, D., Maranian, P., Rothrock, N., Cella, D., Gershon, R., Khanna, P. P., et al. (2012). Feasibility and construct validity of PROMIS and legacy Instruments in an Academic Scleroderma Clinic—Analysis from the UCLA Scleroderma Quality of Life Study. *Value Health, 15*(1), 128–134. https://doi.org/10.1016/j.jval.2011.08.006.
18. US Department of Health and Human Services Food and Drug Administration. (2009). Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims.
19. Unger, E. R., Lin, J.-M. S., Tian, H., Natelson, B. H., Lange, G., Vu, D., et al. (2017). Multi-site clinical assessment of myalgic encephalomyelitis/chronic fatigue syndrome (MCAM): Design and Implementation of a prospective/retrospective rolling cohort study. *American Journal of Epidemiology, 185*(8), 617–626. https://doi.org/10.1093/aje/kwx029.
20. Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E., Jr. (2014). Method of administration of PROMIS scales did not significantly impact score level, reliability, or validity. *Journal of Clinical Epidemiology, 67*(1), 108–113. https://doi.org/10.1016/j.jclinepi.2013.07.016.
21. Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Rotterdam: Psychology Press.
22. Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the PROMIS Internet Panel. *Journal of Clinical Epidemiology, 63*(11), 1169–1178. https://doi.org/10.1016/j.jclinepi.2009.11.021.
23. Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling*. Lincolnwood, IL: Scientific Software International.
24. SAS Institute Inc. (2002–2014). *SAS® 9.4 Help and Documentation*. Cary, NC: SAS Institute Inc.
25. Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2009). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behav Res, 43*(3), 476–496. https://doi.org/10.1080/00273170802285941.
26. Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). *Quality criteria were proposed for measurement properties of health status questionnaires., 60*(1), 34–42.

27. Nunnally, J., & Bernstein, I. (1978). Psychometric Theory McGraw-Hill New York Google Scholar.
28. Lohr, K N J Qo L R. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life, 11*(3), 193–205.
29. Bartlett, S. J., Orbai, A.-M., Duncan, T., DeLeon, E., Ruffing, V., Clegg-Smith, K., et al. (2015). Reliability and validity of selected PROMIS measures in people with rheumatoid arthritis. *PLoS ONE, 10*(9), e0138543.
30. Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. J. J. O. C. E. (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology, 64*(5), 507–516.
31. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
32. Langer, M. M. (2008). *A Reexamination of Lord's Wald Test for Differential Item Functioning Using Item Response Theory and Modern Error Estimation*. Doctoral Dissertation, University of North Carolina at Chapel Hill.
33. Rusu, C., Gee, M. E., Lagacé, C., & Parlor, M. (2015). Chronic fatigue syndrome and fibromyalgia in Canada: Prevalence and associations with six health status indicators. *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice, 35*(1), 3–11.
34. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300.
35. Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77–83. https://doi.org/10.3102/10769986027001077.
36. Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*(2), 99–114. https://doi.org/10.2307/3001913.
37. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
38. Miles, J., & Shevlin, M. (2001). *Applying regression & correlation: A guide for students and researchers* (1st ed.). Thousand Oaks, CA: Sage.
39. Badhiwala, J. H., Witiw, C. D., Nassiri, F., Akbar, M. A., Jaja, B., Wilson, J. R., et al. (2018). Minimum clinically important difference in SF-36 scores for use in degenerative cervical myelopathy. *Spine*. https://doi.org/10.1097/brs.0000000000002684.
40. Swigris, J. J., Brown, K. K., Behr, J., du Bois, R. M., King, T. E., Raghu, G., et al. (2010). The SF-36 and SGRQ: Validity and first look at minimum important differences in IPF. *Respiratory Medicine, 104*(2), 296–304. https://doi.org/10.1016/j.rmed.2009.09.006.
41. Ward, M. M., Guthrie, L. C., & Alba, M. I. (2014). Clinically important chances in short form-36 scales for use in rheumatoid arthritis clinical trials: The impact of low responsiveness. *Arthritis Care & Research, 66*(12), 1783–1789. https://doi.org/10.1002/acr.22392.
42. Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41*(5), 582–592. https://doi.org/10.1097/01.mlr.0000062554.74615.4c.
43. Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: Assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases, 40*(2), 171–178. https://doi.org/10.1016/0021-9681(87)90069-5.
44. Cook, K. F., Bamer, A. M., Roddey, T. S., Kraft, G. H., Kim, J., & Amtmann, D. (2012). A PROMIS fatigue short form for use by individuals who have multiple sclerosis. *Quality of Life Research, 21*(6), 1021–1030. https://doi.org/10.1007/s11136-011-0011-8.
45. Ameringer, S., Elswick, R. K., Jr., Menzies, V., Robins, J. L., Starkweather, A., Walter, J., et al. (2016). Psychometric evaluation of the patient-reported outcomes measurement information system fatigue-short form across diverse populations. *Nursing Research, 65*(4), 279–289. https://doi.org/10.1097/nnr.0000000000000162.
46. DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(5 Suppl 1), S12–S21. https://doi.org/10.1097/01.mlr.0000254567.79743.e2.