

Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the Modified Fatigue Impact Scale and the PROMIS Fatigue Short Form

Vanessa K. Noonan · Karon F. Cook ·
Alyssa M. Bamer · Seung W. Choi ·
Jiseon Kim · Dagmar Amtmann

Accepted: 5 October 2011 / Published online: 3 November 2011
© Springer Science+Business Media B.V. 2011

Abstract

Purpose To create cross-walk tables to associate scores for the Modified Fatigue Impact Scale (MFIS) with scores for the Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) in persons with Multiple Sclerosis (MS).

Methods Cross-walk tables were created using equipercentile linking and based on data collected at one time point in a longitudinal study of persons with MS ($N = 458$). Validation of the tables was conducted using data collected at a subsequent time point ($N = 444$). Deviations between estimates and actual scores were compared across levels of fatigue. The impact of sample size on the precision of sample mean estimates was evaluated using bootstrapping.

Results Correlations between deviations and fatigue level for the PROMIS Fatigue SF and MFIS were (-0.31) and (-0.30), respectively, indicating moderately greater deviations with lower fatigue scores. Estimated sample means were impacted by sample size.

Conclusions Cross-walk tables allow data from studies using different measures of fatigue to be combined to achieve larger sample sizes and to compare results. These

tables are valid for group-level analyses with sample sizes of 150 or greater.

Keywords Fatigue · Multiple sclerosis · Outcome assessment · Questionnaires

Introduction

Multiple Sclerosis (MS) is a chronic progressive health condition with an estimated global incidence of 2.5 per 100,000 individuals per year [1]. One of the most disabling sequelae of this health condition is fatigue, which is reported to affect 53–92% of persons with MS [2]. No cure for MS currently exists, and there is a need to evaluate new and existing disease managing therapies. Comparisons of treatment effectiveness with respect to symptoms such as pain and fatigue depend on the availability of clinically meaningful and psychometrically sound instruments.

A number of instruments have been used to measure fatigue in MS. In 1998, the Multiple Sclerosis Council for Clinical Practice Guidelines recommended the Modified Fatigue Impact Scale (MFIS) for use in clinical practice and research [3]. The MFIS is commonly used to assess physical and cognitive fatigue in MS. More recently, a dynamic fatigue instrument was developed as part of the National Institutes of Health's Patient Reported Outcome Measurement Information System (PROMIS) [4]. The PROMIS instruments are based on item banks that were calibrated using item response theory (IRT). The item banking approach allows for computer adaptive administration of items and the construction of different short forms to target specific measurement needs. An additional advantage of the PROMIS instruments is that their scores are on a T -score metric (mean = 50, standard

V. K. Noonan (✉) · K. F. Cook · A. M. Bamer ·
J. Kim · D. Amtmann
Department of Rehabilitation Medicine,
University of Washington,
Box 356490, Seattle, WA 98195, USA
e-mail: Vanessa.Noonan@vch.ca

S. W. Choi
Department of Medical Social Sciences,
Northwestern University Feinberg School of Medicine,
Chicago, IL 60611, USA

deviation = 10) that is centered on the general United States population, thus assisting in interpretation of scores [5].

Despite the advantages of the PROMIS instruments, a barrier to adoption of any new instrument is the need to associate findings from new studies with those of previous studies that used a different measure of the outcome. One remedy is to use “linking” methods to associate scores from one instrument to those of another instrument of the same or a similar construct [6]. Linking methods vary in both methodology and assumptions. Critical to the success of a linking strategy is choice of an appropriate method [7].

The objectives of this study were to: (1) identify an appropriate method for linking MFIS total score and the PROMIS Fatigue Short Form (SF) scores, (2) develop cross-walk tables that associate scores from the MFIS with scores from the PROMIS Fatigue SF, and (3) validate the linking results at a follow-up time point.

Materials and methods

Participants and study design

Data for the current study were collected as part of a longitudinal study of symptoms and quality of life indicators of persons with MS. Research participants were recruited through the Greater Washington chapter of the USA National Multiple Sclerosis Society (NMSS). Letters were sent to 7,806 persons from the NMSS mailing list. Eligibility criteria included being over the age of 18 and self-reporting having been diagnosed with MS by a physician. Of the 1,629 respondents, 1,597 were eligible and were mailed a paper survey. Reminder letters were sent to non-responders 3–6 weeks after the survey was mailed. There were 1,271 subjects in the first survey (Time 1) and a random subset of 562 subjects was invited to participate in the longitudinal study that required completing five additional surveys, with approximately 4 months between the repeated administrations. For the current study, data from the fifth and sixth time points were used. Type of MS was categorized using a self-report item previously validated [8].

The MFIS and PROMIS Fatigue SF were administered at both time points (single-group linking design) [7]. Cross-walk tables were developed using responses collected at the fifth time point (‘linking data’). The cross-walk tables were evaluated using data collected at the sixth time point (‘cross-validation data’). The fifth and sixth time points were selected since both instruments were only included at these times. The study was approved by the Human Subjects Division at the University of Washington.

Instrumentation

Modified Fatigue Impact Scale (MFIS)

Developed to measure perceived impact of fatigue on the lives of individuals with MS, the MFIS is a 21-item version of the Fatigue Impact Scale (FIS) [9, 10]. The MFIS yields a total score and three domain scores: physical (9 items), cognitive (10 items), and psychosocial (2 items). Items are rated on a 5-point scale ranging from ‘never’ to ‘almost always’. Higher scores indicate more fatigue [3].

There is evidence to support the validity and reliability of MFIS scores in persons with MS. In a sample of MS participants from four European countries, reported internal consistency values for the total score, physical, cognitive, and psychosocial subscales were, respectively, 0.92, 0.88, 0.92, and 0.65 [11]. Test–retest reliability using intra class correlation coefficients was 0.91 (99% confidence interval 0.86–0.94) [11]. No floor or ceiling effects were observed for the total score [11]. Evidence for the concurrent validity of MFIS scores has been established by comparing MFIS scores to the Fatigue Severity Scale score [11–13].

PROMIS Fatigue Short Form (PROMIS Fatigue SF)

In the PROMIS domain framework, fatigue is described as ranging from “mild subjective feelings of tiredness to an overwhelming, debilitating, and sustained sense of exhaustion that is likely to decrease one’s ability to carry out daily activities, including the ability to work effectively and to function at one’s usual level in family or social roles” [14] (page 1,318). The PROMIS fatigue item bank was developed using both qualitative and quantitative methods, including review of over 1,000 items from 80 existing fatigue scales [15]. In addition to the item bank, several short forms (SFs) have been developed [16]. The first of these, the PROMIS Fatigue SF (V1.0), consists of seven items selected by multidisciplinary panels of clinical experts (including physicians, nurses, pharmacists, and psychologists) to represent the content in the PROMIS Fatigue item bank [16]. Item responses are rated on a 5-point scale ranging from ‘never’ to ‘always’ and are summed and transformed to a *T*-score metric. Higher scores indicate more fatigue. PROMIS Fatigue SF scores have good precision across different levels of fatigue. More than 95% of the PROMIS samples were measured with a reliability greater than 0.9 [16].

Analyses

Analyses were conducted in three phases. In Phase One, we identified an appropriate linking method for associating scores from the MFIS and PROMIS Fatigue SF. In Phase

Two, we constructed cross-walk tables to associate scores from the MFIS (total score) and the PROMIS Fatigue SF. In Phase Three, we cross-validated the linking results with data collected at a follow-up time point.

Phase One: selection of linking method

From most to least restrictive in terms of statistical assumptions, there are three different approaches to linking scores from two instruments: equating, scale alignment and prediction [7]. A detailed description and comparison of linking methods are included in Table 1. The more restrictive the approach used, the closer the link between scores. For example, when restrictive assumptions of equating are met, the resulting scores are considered interchangeable. When scale alignment is used, scores between instruments are “associated”, but are not equivalent. Prediction also associates scores, but the lack of assumptions weakens the association between scores. Therefore, it is advantageous to use the most restrictive linking method justified by the data. There are many examples of how these linking methods have been used to create cross-walk tables between scores on health outcome measures [17–19].

As detailed in Table 1, *equating* scores of two measures requires several assumptions be met: (1 & 2) the linked instruments must measure the same construct and have equal reliability, (3) the linking functions that associate the scores of the instrument must be symmetrical (one is the

inverse of the other), (4) it should not matter which instrument a score is based on (equity), and (5) the linking function should be invariant in different sub-populations [7]. In health outcomes measurement, the assumptions of equating are seldom, if ever, met.

Scale alignment, an omnibus term for several similar methods, is less restrictive than equating [6]. There are several scale alignment methods, but the most relevant in the current context is *concordance* [7]. Dorans [20] recommends concordance be used when: (1) constructs are similar; (2) scores are highly correlated; and (3) scores are invariant across sub-populations (such as different types of MS) suggesting similar linking relationships across groups [21]. This method is appropriate when the objective is to link summary scores from two instruments that assess similar constructs but were developed using different test specifications [7].

The least restrictive method for associating scores from two instruments is *prediction*. Prediction is used when the linking assumptions of equating or scale alignment cannot be met [7]. Using the prediction method, scores on one instrument are estimated (predicted) based on prior information (e.g., scores on a different instrument) [7].

Phase One analyses

Because the assumptions of equating are so restrictive, our selection of a linking strategy was a choice between scale alignment and prediction. To assist in this decision, we

Table 1 Overview of linking methods [6,7,20]

Type of Linking	Definition	Assumptions
Equating	Establishes an effective equivalence between scores on two measures to allow scores from both to be used interchangeably.	<ol style="list-style-type: none"> 1. Equal (same) constructs measured in both measures. 2. Equal reliability (measurement errors) in both measures. 3. Symmetrical (function for linking scores of Y to the scores of X should be the inverse of the linking function for equating the scores of X to those of Y). 4. Equity (should not matter if a person is assessed by either one of the two measures that have been equated). 5. Population invariance (linking function used to link measures X and Y should be population invariant).
Scale alignment	Transforms scores from two different measures onto the same metric.	<p>All approaches to scale aligning meet assumption 3 (symmetrical) as described above. Two approaches relevant to health outcomes measurement are concordance and calibration which are described below.</p> <p><i>Concordance</i>: linking two measures developed according to different test specifications that measure similar constructs and have similar reliability estimates (e.g., linking two different fatigue measures).</p> <p><i>Calibration</i>: linking two measures developed using the same test specifications that measure the same construct and have dissimilar reliability (e.g., a new fatigue short form developed from the original long form measure).</p>
Prediction	Estimates a score from a measure using information from the respondent.	Does not require meeting any of the assumptions 1 to 5.

conducted analyses to evaluate each of the conditions recommended by Dorans [20] as necessary for scale alignment (concordance).

Similarity of constructs measured This was evaluated by textual analysis of the content of the MFIS and PROMIS Fatigue SF items. Recurring themes were identified and items were grouped according to these themes. The relative representation of each instrument was evaluated by content theme.

Strength of the empirical relationship between the scores. The relationship between MFIS total score and the PROMIS Fatigue SF score was examined by evaluating dimensionality of the scales and the correlation among the scores. Dimensionality was assessed by conducting factor analyses of item responses using MPLUS 5.21 software [22]. Scales such as the MFIS in which a summary score can be obtained make the assumption that item responses yield data that are essentially unidimensional. This also is an assumption of the IRT scored measures, such as those developed by PROMIS. If this assumption is not supported, it is difficult to interpret what the summary score means. To evaluate the fit of a unidimensional model, we used an approach recommended by Reeve et al. [23], in which first a confirmatory factor analysis (CFA) model is fit to evaluate the fit of a unidimensional model. The following statistical criteria were used to assess model fit: (1) root mean square error of approximation (RMSEA) <0.06 [23]; (2) comparative fit index (CFI) >0.95 [23]; (3) Tucker-Lewis Index (TLI) >0.95 [23]; (4) weighted root mean square residual (WRMSR) <1.0 [24]; and 5) average absolute residual correlations <0.10 [23]. If the data fail these criteria, an exploratory factor analytic (EFA) model is fit and results are compared including: (1) magnitude of eigenvalues (a minimum of 20% of the variability on the first factor); (2) ratio of the first and second factor (a ratio greater than 4 is desirable); (3) scree test, and (4) pattern of factor loadings [23]. The association between the two scores was evaluated using Pearson correlation coefficient. Dorans [20] suggested $r >0.866$ when concordance procedures are used.

Invariance of scores across sub-populations To evaluate linking invariance across sub-populations, we calculated standardized mean differences (SMDs) for subsample scores categorized by age (≤ 53 vs. >53 years); education level (less than college versus college diploma or university degree); type of MS (relapsing remitting versus all others); and duration of MS (≤ 12 vs. >12 years) [20]. In the current study, the SMD was estimated by subtracting the mean score for group 1 (e.g., relapsing remitting MS) from the mean score for group 2 (e.g., other types of MS) and dividing by the total group SD [20]. Estimates of SMD were calculated for both PROMIS Fatigue scores and for MFIS scores; a difference in SMDs greater than 0.11

suggests a need for sub-population-specific cross-walks [20].

Phase Two: linking the MFIS and PROMIS Fatigue SF

As reported below, Phase One analyses supported use of concordance to link scores. Equipercentile methods were used to identify pairs of raw scores that corresponded to the same percentile rank. For example, the raw score associated with the 50th percentile of MFIS scores was paired with the PROMIS Fatigue SF score that was at the 50th percentile and vice versa. For discrete variables like raw summed scores, the equipercentile linking functions may not be exactly symmetric because of rounding errors and differences in score ranges. Therefore, we identified two cross-walk tables, one for estimating MFIS total scores based on PROMIS Fatigue SF scores and the other for estimating PROMIS Fatigue SF scores based on MFIS total scores.

Phase Three: validating the cross-walk tables

The success of the cross-walk tables developed in Phase Two was evaluated using data from Time 6. Using the cross-walk developed in Time 5 and the data collected at Time 6, we projected Time 6 MFIS total scores (MFIS_{PROJ}) based on actual Time 6 PROMIS Fatigue SF scores. We projected Time 6 PROMIS Fatigue SF scores (PROMIS_{PROJ}) based on actual Time 6 MFIS total scores. We calculated deviations between projected and actual scores. MFIS_{DEV} was defined as MFIS_{PROJ} minus Time 6 MFIS; PROMIS_{DEV} was defined as PROMIS_{PROJ} minus Time 6 PROMIS Fatigue SF. The values of MFIS_{DEV} and PROMIS_{DEV} were examined over the score range to determine whether there were greater deviations between projected and actual values at different fatigue levels.

With larger sample sizes, the impact of deviations in individual scores may average out, but with smaller sample sizes, the cross-walking tables are less likely to closely approximate sample mean scores. Bootstrapping (random sampling with replacement) was conducted using samples of varying size. Random samples (50,000 replications with replacement) were derived for sample sizes ranging from 25 to 200 (with an increment of 25). Cross-walk-based means were estimated for each random sample. These estimates were compared with the observed sample mean using box plots.

Results

The linking data were responses of 461 participants who participated in the longitudinal survey study at Time 5; the

Table 2 Demographic and disease characteristics of a community sample of individuals with multiple sclerosis at the two data collection time points

Variable	Linking sample (<i>n</i> = 458) Time point 5 <i>n</i> (%) mean ± SD	Cross-validation sample (<i>n</i> = 444) Time point 6 <i>n</i> (%) mean ± SD
Age	52.8 ± 10.9	53.2 ± 10.9
Duration of disease	14.5 ± 9.9	14.9 ± 10.0
Sex		
Women	379 (82.8)	366 (82.4)
Men	79 (17.2)	78 (17.5)
Race		
Caucasian	419 (91.5)	406 (91.4)
Other	39 (8.5)	38 (8.6)
Education completed		
< High school/High school/GED	64 (13.9)	62 (13.9)
Vocational/some college	178 (38.8)	170 (38.3)
College graduate	134 (29.3)	130 (29.3)
University graduate	82 (17.9)	82 (18.5)
Employment status		
Employed	166 (36.2)	163 (36.7)
Married		
Married/live with significant other	318 (69.4)	305 (68.7)
Separated/divorced/single/widowed	140 (31.6)	139 (31.3)
Course of disease		
Relapsing remitting	260 (56.8)	254 (57.2)
Secondary progressive	95 (20.7)	88 (19.8)
Primary progressive	52 (11.4)	52 (11.7)
Progressive relapsing	42 (9.2)	41 (9.2)
Missing	9 (2.0)	9 (2.0)

cross-validation data were responses of the subset of 446 (of 461) who participated at both Time 5 and Time 6. Five participants with missing item responses were removed from the sample (list-wise deletion) during the creation of the cross-walk tables and for the cross-validation. The demographics and clinical variables for the sample at each time point are described in Table 2. As expected, given the overlap between persons participating at Time 5 and Time 6, characteristics of the linking and cross-validation data were very similar.

Phase One: selection of linking method

Similarity of constructs measured

A comparison of the characteristics and content of the MFIS and PROMIS Fatigue SF is reported in Table 3. There are differences between the two instruments. The MFIS has a 4-week time frame; the PROMIS Fatigue SF has a 7-day time frame. Although both instruments cover physical, cognitive, and psychosocial aspects of fatigue,

compared with the MFIS, the PROMIS Fatigue SF has greater coverage of physical aspects of fatigue and less coverage of cognitive aspects.

Strength of the empirical relationship between the scores

The MFIS total score and PROMIS Fatigue SF summative score for the linking sample were 38.9 ± 19.0 (range 0–84) and 20.5 ± 5.9 (range 7–35), respectively. Based on a common 0–100 score range, the two instruments produced very similar medians (48.2 and 47.6 for PROMIS Fatigue SF and MFIS, respectively) and SDs (21.0 and 22.7 for PROMIS Fatigue SF and MFIS). Quantile–Quantile plots for the two instruments suggest a similar score distribution (Fig. 1a, b). The correlation among the MFIS total score and the PROMIS Fatigue SF was $r = 0.849$ ($P < 0.01$). This is close to the value of 0.866 recommended by Dorans [20] when evaluating whether scales are appropriate for the concordance approach to linking.

The CFA of the PROMIS Fatigue SF responses resulted in fit statistics consistent with published guidelines; this

Table 3 Comparison of the Modified Fatigue Impact Scale (MFIS) and Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF)

	MFIS	PROMIS Fatigue SF
Scales, number of items	Physical, 9 items Cognitive, 10 items Psychosocial, 2 items Total, 21 items	Total, 7 items
Type of fatigue assessed based on content analysis	Physical domain ($n = 9$) Impact ($n = 5$) Symptoms ($n = 3$) Cognitive (motivation) ($n = 1$) Cognitive domain ($n = 10$) Impact of symptoms ($n = 10$) Psychosocial domain ($n = 2$) Physical ($n = 1$) Cognitive ($n = 1$)	Physical ($n = 5$) Impact ($n = 2$) Symptoms ($n = 3$) Cognitive ($n = 1$) Impact Psychosocial ($n = 1$) Physical or cognitive
Item stem and time frame	Because of my fatigue during the past 4 weeks...	In the past 7 days...
Response options	5 point scale (never/rarely/sometimes/often/almost always)	5 point scale (never/rarely/sometimes/often/always)
Scoring	MFIS Physical range 0–36 MFIS Cognitive range 0–40 MFIS Psychosocial 0–8 MFIS Total 0–84 (higher score indicates more fatigue)	Summed score 7–35 (higher score indicates more fatigue) T -score metric (mean = 50, SD = 10) (higher score indicates more fatigue)

was not the case for responses to the MFIS. The RMSEA for the PROMIS Fatigue SF and MFIS was 0.10 and 0.32, respectively. CFI values were 0.99 for the PROMIS Fatigue SF and 0.92 for the MFIS responses. TLI values were 0.99 for the PROMIS Fatigue SF and 0.96 for the MFIS. The WRMSR values were 0.77 for the PROMIS Fatigue SF data and 3.93 for the MFIS data. Average absolute residual correlations were 0.12 for the MFIS and 0.03 for the PROMIS Fatigue SF.

Reeve et al. [23] recommended follow-up analysis with EFA when item responses do not meet published fit statistic criteria for unidimensionality. The EFA provided support for a single underlying dimension on both scales. On the MFIS, the proportion of the variance explained by the first factor was 67.0%. The ratio between the first and second eigenvalue was 5.1, and factor loadings for a one-factor solution were >0.4 (0.72–0.95). The scree plot indicated a strong first factor and a weak second factor. For the PROMIS Fatigue SF, the first factor accounted for 66.4% of the variance and the ratio between the first and second factor was 5.3. The scree plot suggested one strong factor. Factor loadings for the 7 items ranged from 0.36 to 0.94 for a one-factor solution.

Invariance of scores across sub-populations

Score invariance was assessed for age, education level, type of MS, and duration of MS using Time 5 data. The effect of age was examined by creating two groups: individuals ≤ 53 years ($n = 232$) and >53 years ($n = 226$). Level of education was classified as less than college ($n = 242$) and a college diploma or university degree ($n = 216$). Those with relapsing remitting MS ($n = 260$) were compared with all other MS types ($n = 189$). Duration of MS was examined by creating two groups: individual with $MS \leq 12$ years ($n = 239$) and $MS > 12$ years ($n = 219$). The difference in SMD values for the MFIS and the PROMIS Fatigue SF for age, education, type of MS, and duration of MS was 0.01, 0.01, 0.03, and 0.08, respectively; these differences in SMD for all these sub-populations met the criterion of <0.11 suggested by Dorans [20].

Phase Two: linking the MFIS and PROMIS Fatigue SF

Cross-walk tables were developed for the MFIS total score and the PROMIS Fatigue SF (and vice versa) using equipercentile methods. The distribution for both the MFIS and PROMIS Fatigue SF were very similar, supporting the

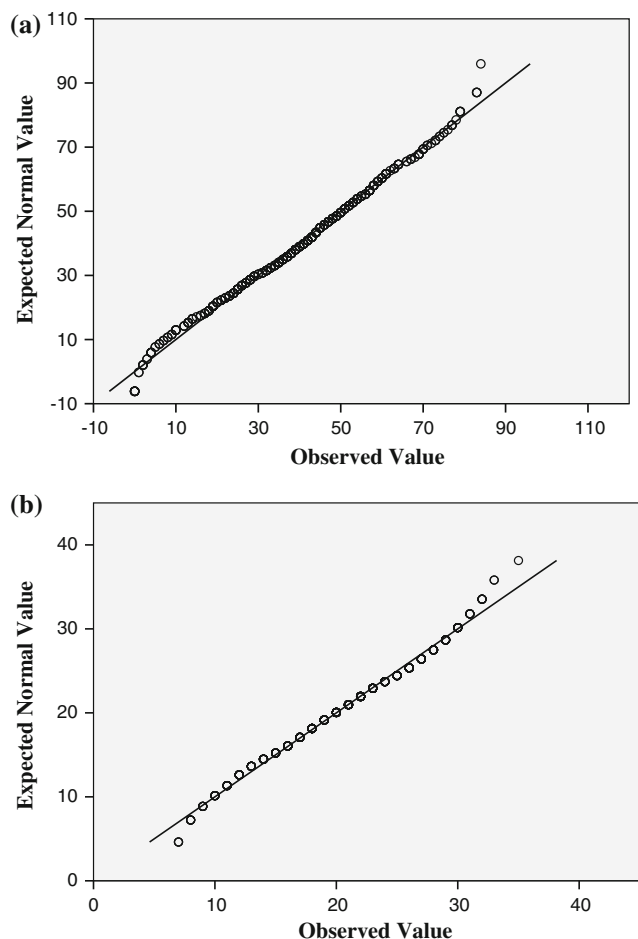


Fig. 1 **a** A normal Q–Q plot of the Modified Fatigue Impact Scale (MFIS) total score. The *plot* suggests the data follow a normal distribution, with a slight deviation at each end. Since the distribution functions of the two instruments are very similar (see Fig. 1b), this suggests it is appropriate to use equipercentile linking. **b**. A normal Q–Q plot of the Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) summative score. The plot suggests the data follow a normal distribution, with a slight deviation at each end. Since the distribution functions of the two instruments are very similar (see Fig. 1a), this suggests it is appropriate to use equipercentile linking

choice of this method (see Fig. 1). Figure 2 shows the function that links PROMIS Fatigue SF scores to the MFIS and the function that links MFIS scores to those of the PROMIS measure. The cross-walk tables are included in Tables 4 and 5. As mentioned, scores on the PROMIS measures are reported on a *T*-score metric that is anchored at the mean of a general U.S. population. In Tables 4 and 5, we have included both the raw summative scores and the scores on the *T*-score metric. An advantage of reporting the linking in the *T*-score metric is that it links MFIS to all PROMIS Fatigue scores whether they were obtained using the version one short form, another form, or based on computer adaptive testing.

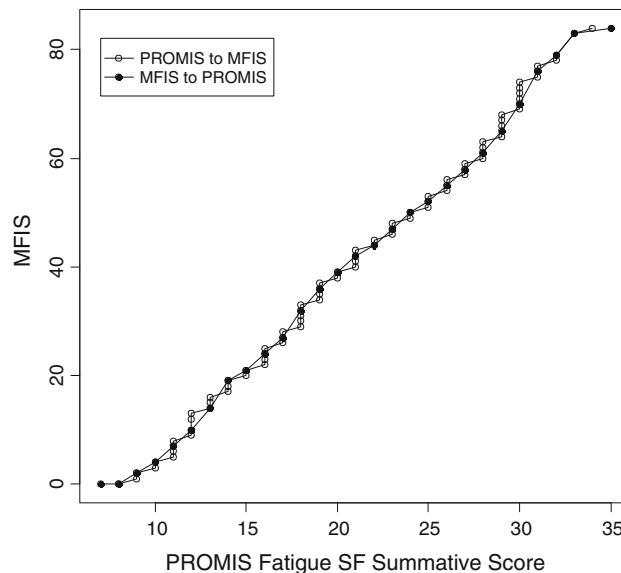


Fig. 2 Equipercetile linking functions from (1) Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) summative score to Modified Fatigue Impact Scale (MFIS) total score and (2) MFIS total score to PROMIS Fatigue SF summative score

Phase Three: validating the cross-walk tables

Scores for the MFIS and PROMIS Fatigue SF in the cross-validation data were very similar to those in the linking data; MFIS cross-validation data mean = 38.0 ± 18.4 (versus MFIS linking sample 38.9 ± 19.0) and PROMIS Fatigue SF cross-validation data mean = 20.4 ± 5.8 (versus PROMIS Fatigue SF linking sample 20.5 ± 5.9).

PROMIS_{DEV} summed item scores ranged from -11.0 to 14.0 with a mean of -0.11 and SD of 3.35 . MFIS_{DEV} total scores ranged from -45.0 to 36.0 with a mean of 0.35 and SD of 10.61 . To evaluate whether the cross-walk tables are more accurate at various levels of fatigue, we correlated deviation scores with the levels of fatigue. Correlations between deviations and fatigue level for the PROMIS Fatigue SF and MFIS were (-0.31) and (-0.30) , respectively, indicating moderately greater deviations with lower fatigue scores. That is, the cross-walks are more accurate at higher than at lower levels of fatigue.

Sample size had a considerable impact on mean deviations. Box plots for sample sizes ranging from 25 to 200 are displayed in Fig. 3 (Fig. 3a, b). Note that in Fig. 3b, PROMIS Fatigue SF scores are reported on the PROMIS *T*-score metric. Although, on average, even small sample sizes yielded mean deviation scores near zero, and the estimates of the majority of samples were within 2 points of the actual sample mean, outlier samples had very different results. For example, some of the 50,000 replications of samples of 25 yielded mean summed MFIS total scores

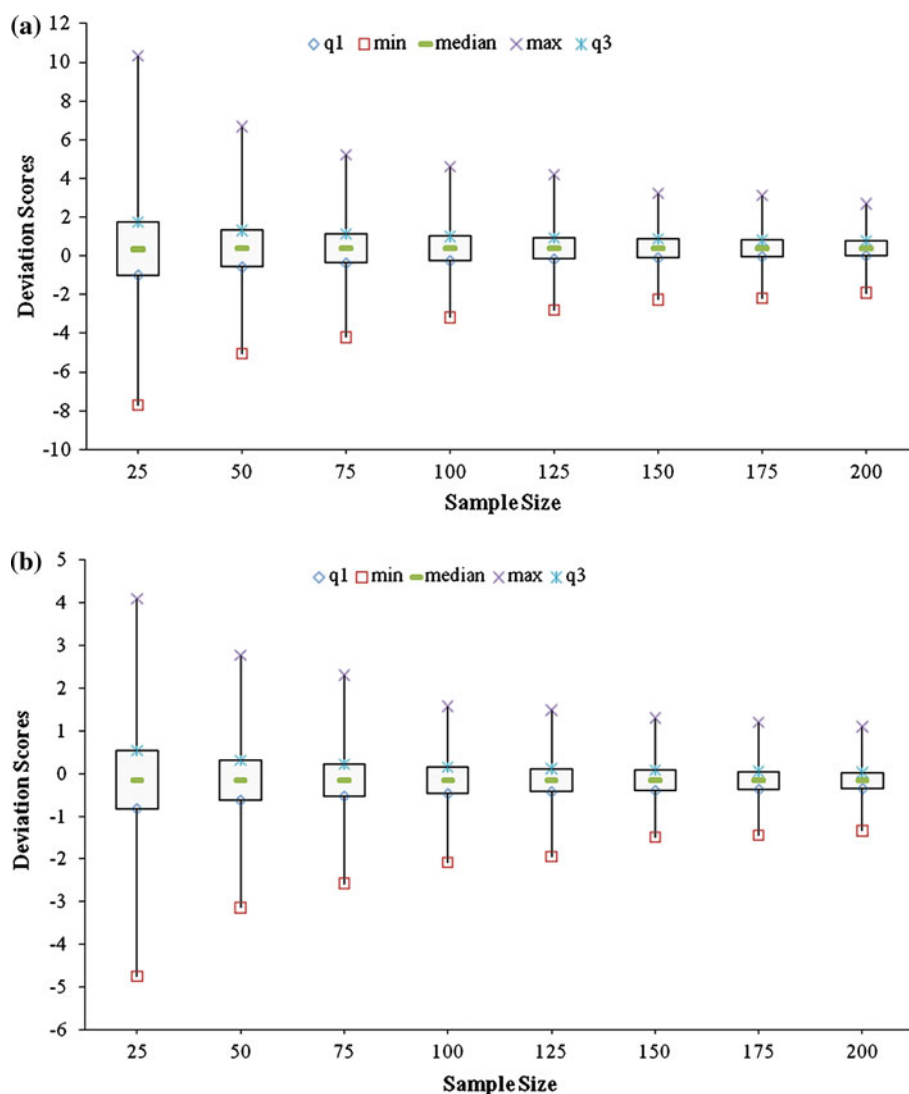


Fig. 3 a. Box plots of the deviation scores (projected scores minus actual scores using the cross-walk tables) for sample sizes ranging from 25 to 200 for the Modified Fatigue Impact Scale (MFIS) total score. Each box plot includes the range (*minimum* and *maximum* values) as well as the interquartile range and median value. The majority of samples were within 2 points of the actual sample mean; however, for some of the replications of samples of 25, the mean MFIS total scores overestimated or underestimated the observed mean of MFIS total scores by 8–10 points. There was less variability in sample sizes greater than 150. **b.** Box plots of the deviation score

(projected score *minus* the actual score using the cross-walk tables) for sample sizes ranging from 25 to 200 for the Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) *T*-score. Each box plot includes the range (*minimum* and *maximum* values) as well as the interquartile range and median value. The majority of samples were within 1 point of the actual sample mean; however, for some of the replications of samples of 25, the mean PROMIS Fatigue SF *T*-score overestimated or underestimated the observed mean PROMIS Fatigue SF *T*-scores by 4–5 points. There was less variability in sample sizes *greater* than 150

overestimated or underestimated the observed mean of MFIS scores by 8–10 points. With larger sample sizes, estimates of sample means were much less variable, especially sample sizes ≥ 150 .

Discussion

A common problem when studies use different instruments to measure patient reported outcomes is establishing the

comparability of scales or units on which the scores are reported. The lack of comparability poses significant challenges in studies focusing on measuring effects or changes, particularly in meta-analytic and longitudinal studies where compilation of findings over several studies can be difficult in the absence of a common measure. By linking different instruments, a common language is created that facilitates score comparability across studies.

The first phase of the current study identified an appropriate linking strategy for creating cross-walk tables

Table 4 Cross-walk table from Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) Scores (summative and *T*-score) to Modified Fatigue Impact Scale (MFIS) total score

PROMIS Fatigue SF (Summative)	PROMIS Fatigue SF (<i>T</i> -score)	MFIS (Total)
7	29.4	0
8	33.4	0
9	36.9	2
10	39.6	4
11	41.9	7
12	43.9	10
13	45.8	14
14	47.6	19
15	49.2	21
16	50.8	24
17	52.2	27
18	53.7	32
19	55.1	36
20	56.4	39
21	57.8	42
22	59.2	44
23	60.6	47
24	62.0	50
25	63.4	52
26	64.8	55
27	66.3	58
28	67.8	61
29	69.4	65
30	71.1	70
31	72.9	76
32	74.8	79
33	77.1	83
34	79.8	84
35	83.2	84

between the MFIS and the PROMIS Fatigue SF. This is a critical step in linking, but one that is seldom described in the health literature [7]. Doran's [20] three criteria for using concordance to link the scores on the MFIS and PROMIS Fatigue SF were supported by the study results.

Findings from the cross-validation of the linking results supported the accuracy of the cross-walks to estimate sample means with sample sizes of 150 or greater. The results do not support clinical use at the individual level or research use in smaller samples. However, with adequate sample sizes, the cross-walks provided good estimates of sample means. The ability to cross-walk the scores will be of use in comparing results from individual studies and, more formally, in meta-analyses.

We found a moderate negative association between deviation scores and fatigue level, i.e., at lower fatigue

levels, deviations tended to be higher than at higher fatigue levels. Some of this association may be explained by the failure of scores on both instruments to discriminate well among people with very low levels of fatigue. The higher deviation scores may be due largely to measurement error rather than bias in the linking procedure.

Although the MFIS has been recommended by the Multiple Sclerosis Council for Clinical Practice Guidelines for use in clinical practice and research [3], the instruments developed based on the recent NIH PROMIS initiative have distinct advantages. These include the availability of multiple assessment formats (e.g., Computer Adaptive Test [CAT] and short forms) and the ability to compare scores to US population norms. We expect these advantages to result in more frequent use of the PROMIS instruments in the future. Given adequate sample sizes, the cross-walk tables developed in this study make possible comparisons of MFIS total scores to the PROMIS general US population norming sample. Furthermore, since all PROMIS instruments are scored on a common metric (*T*-score), these cross-walk tables can be used to associate scores from the MFIS with PROMIS Fatigue scores obtained using other short forms or based on computer adaptive administration of items. Development of MS-specific norms (e.g., type of MS, years with MS) would add substantial additional score interpretability.

The present study has limitations that need to be considered. The cross-walk tables are not suitable for use at the individual level or with small samples. Further, though the linking function successfully associated scores from the two instruments, cross-walked scores are not equivalent and should not be considered interchangeable. The MFIS and the PROMIS Fatigue SF both assess the impact of fatigue, but there are notable differences in how the construct is operationalized in each instrument. The MFIS was developed specifically for persons with MS, whereas the PROMIS Fatigue item bank and SF were developed for the range of chronic conditions. Furthermore, although one of the strengths of this study is that the cross-walk tables were validated in a subsequent time point of a longitudinal study, a stronger design would have been to cross-validate using an independent sample. This should be addressed in future studies.

In conclusion, the cross-walk tables developed in this study enable scores from the MFIS and PROMIS Fatigue SF to be linked and compared. When sample sizes are 150 or greater, scores of the MFIS and PROMIS Fatigue SF can be cross-walked with relatively small estimation error in sample mean estimates. Cross-walking will allow data from studies to be combined to examine effectiveness of MS intervention studies and will support meta-analytic studies.

Table 5 Cross-walk table from Modified Fatigue Impact Scale (MFIS) total score to Patient Reported Outcome Measurement Information System (PROMIS) Fatigue Short Form (SF) Scores (summative and *T*-score)

MFIS (Total)	PROMIS Fatigue SF (Summative)	PROMIS Fatigue SF (<i>T</i> -score)
0	8	33.4
1	9	36.9
2	9	36.9
3	10	39.6
4	10	39.6
5	11	41.9
6	11	41.9
7	11	41.9
8	11	41.9
9	12	43.9
10	12	43.9
11	12	43.9
12	12	43.9
13	12	43.9
14	13	45.8
15	13	45.8
16	13	45.8
17	14	47.6
18	14	47.6
19	14	47.6
20	15	49.2
21	15	49.2
22	16	50.8
23	16	50.8
24	16	50.8
25	16	50.8
26	17	52.2
27	17	52.2
28	17	52.2
29	18	53.7
30	18	53.7
31	18	53.7
32	18	53.7
33	18	53.7
34	19	55.1
35	19	55.1
36	19	55.1
37	19	55.1
38	20	56.4
39	20	56.4
40	21	57.8
41	21	57.8
42	21	57.8
43	21	57.8
44	22	59.2

Table 5 continued

MFIS (Total)	PROMIS Fatigue SF (Summative)	PROMIS Fatigue SF (<i>T</i> -score)
45	22	59.2
46	23	60.6
47	23	60.6
48	23	60.6
49	24	62.0
50	24	62.0
51	25	63.4
52	25	63.4
53	25	63.4
54	26	64.8
55	26	64.8
56	26	64.8
57	27	66.3
59	27	66.3
60	28	67.8
61	28	67.8
62	28	67.8
63	28	67.8
64	29	69.4
65	29	69.4
66	29	69.4
67	29	69.4
68	29	69.4
69	30	71.1
70	30	71.1
71	30	71.1
72	30	71.1
73	30	71.1
74	30	71.1
75	31	72.9
76	31	72.9
77	31	72.9
78	32	74.8
79	32	74.8
80	32	74.8
81	33	77.1
82	33	77.1
83	33	77.1
84	34	79.8

Acknowledgments The contents of this manuscript were developed under grants from the Department of Education, National Institute on Disability and Rehabilitation Research grant numbers H133B031129 and H133B080025, and the National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institute of Health (Grant 5U01AR052171). However, these contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

References

- World Health Organization, & Multiple Sclerosis International Federation. (2008). *Atlas multiple sclerosis resources in the world 2008*. Geneva, Switzerland: World Health Organization.
- Branas, P., Jordan, R., Fry-Smith, A., Burls, A., & Hyde, C. (2000). Treatments for fatigue in multiple sclerosis: A rapid and systematic review. *Health Technology Assessment*, 4, 1–61.
- Multiple Sclerosis Council for Clinical Practice Guidelines. (1998). *Fatigue and multiple sclerosis: Evidence-based management strategies for fatigue in multiple sclerosis*. Washington, DC: Paralyzed Veterans of America.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63, 1179–1194.
- Patient Reported Outcomes Measurement Information System (PROMIS). *PROMIS*. <http://www.nihpromis.org>. Accessed 20 May 2011.
- Holland, P.W. (2007). Chapter 2. A framework and history for score linking. In N.J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *In Linking and Aligning Scores and Scales* (pp. 5–30) Springer, New York.
- Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16, 85–94.
- Bamer, A. M., Cetin, K., Amtmann, D., Bowen, J. D., & Johnson, K. L. (2007). Comparing a self report questionnaire with physician assessment for determining multiple sclerosis clinical disease course: a validation study. *Multiple Sclerosis*, 13, 1033–1037.
- Fisk, J. D., Ritvo, P. G., Ross, L., Haase, D. A., Marrie, T. J., & Schlech, W. F. (1994). Measuring the functional impact of fatigue: Initial validation of the Fatigue Impact Scale. *Clinical Infectious Diseases*, 18, S79–S83.
- Fisk, J. D., Pontefract, A., Ritvo, P. G., Archibald, C. J., & Murray, T. J. (1994). The impact of fatigue on patients with multiple sclerosis. *The Canadian Journal of Neurological Sciences*, 21, 9–14.
- Kos, D., Kerckhofs, E., Carrea, I., Verza, R., Ramos, M., & Jansa, J. (2005). Evaluation of the Modified Fatigue Impact Scale in four different European countries. *Multiple Sclerosis*, 11, 76–80.
- Flachenecker, P., Kumpfel, T., Kallmann, B., Gottschalk, M., Grauer, O., Rieckmann, P., et al. (2002). Fatigue in multiple sclerosis: A comparison of different rating scales and correlation to clinical parameters. *Multiple Sclerosis*, 8, 523–526.
- Tellez, N., Rio, J., Tintore, M., Nos, C., Galan, I., & Montalban, X. (2005). Does the Modified Fatigue Impact Scale offer a more comprehensive assessment of fatigue in MS? *Multiple Sclerosis*, 11, 198–202.
- Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A., et al. (2010). Patient-Reported Outcomes Measurement Information System (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Quality of Life Research*, 19, 1311–1321.
- Christodoulou, C., Junghaenel, D. U., DeWalt, D. A., Rothrock, N., & Stone, A. A. (2008). Cognitive interviewing in the evaluation of fatigue items: Results from the Patient-Reported Outcomes Measurement Information System (PROMIS). *Quality of Life Research*, 17, 1239–1246.
- Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Gershon, R., & Stone, A. (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, S20–S27.
- Fong, T. G., Fearing, M. A., Jones, R. N., Shi, P., Marcantonio, E. R., Rudolph, J. L., et al. (2009). Telephone interview for cognitive status: Creating a crosswalk with the Mini-Mental State Examination. *Alzheimer's & Dementia*, 5, 492–497.
- Leucht, S., Kane, J. M., Etschel, E., Kissling, W., Hamann, J., & Engel, R. R. (2006). Linking the PANSS, BPRS, and CGI: Clinical implications. *Neuropsychopharmacology*, 31, 2318–2325.
- Holzner, B., Bode, R. K., Hahn, E. A., Cella, D., Kopp, M., Sperner-Unterwieser, B., et al. (2006). Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. *European Journal of Cancer*, 42, 3169–3177.
- Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28, 227–246.
- Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, 28, 219–226.
- Muthén, L. K., & Muthén, B. O. (2009). *Mplus: statistical software version 5.21*. Los Angeles, CA: Muthén & Muthén.
- Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45, S22–S31.
- Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Doctor of Philosophy in Education. Los Angeles: University of California.