Psychological Test and Assessment Modeling, Volume 58, 2016 (1), 119-139

Psychometric evaluation of the PROMIS® Fatigue measure in an ethnically and racially diverse population-based sample of cancer patients

Bryce B. Reeve[1,2,3], Laura C. Pinheiro[2,3], Roxanne E. Jensen[4], Jeanne A. Teresi[5,6,7], Arnold L. Potosky[4], Molly K. McFatrich[3], Mildred Ramirez[6,7] & Wen-Hung Chen[8]

Abstract

Aims: Fatigue is the most prevalent and distressing symptom related to cancer and its treatment affecting functioning and quality of life. In 2010, the National Cancer Institute's Clinical Trials Planning Meeting on cancer-related fatigue adopted the PROMIS® Fatigue measure as the standard to use in clinical trials. This study evaluates the psychometric properties of the PROMIS Fatigue measure in an ethnically/racially diverse population-based sample of adult cancer patients.

Methods: Patients were recruited from four US cancer registries with oversampling of minorities. Participants completed a paper survey 6 - 13 months post-diagnosis. The 14 fatigue items (5-point Likert-type scale; English-, Spanish-, and Chinese-versions) were selected from the PROMIS Fatigue short forms and larger item bank. Item response theory and factor analyses were used to evaluate item- and scale-level performance. Differential item functioning (DIF) was evaluated using the Wald test and ordinal logistic regression (OLR) methods. OLR-identified items with DIF were evaluated further for their effect on the scale scores (threshold $r^2 > .13$).

Results: The sample included 5,507 patients (2,278 non-Hispanic Whites, 1,122 non-Hispanic Blacks, 1,053 Hispanics, and 917 Asians/ Pacific Islanders); 338 Hispanics were given the Spanish-language

[1] *Correspondence concerning this article should be addressed to:* Bryce Reeve, PhD, Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 1101-D McGavran-Greenberg Hall, 135 Dauer Drive, CB 7411, Chapel Hill, NC 27599-7411, USA; email: bbreeve@email.unc.edu

[2] Health Policy and Management, University of North Carolina at Chapel Hill

[3] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill

[4] Lombardi Comprehensive Cancer Center, Georgetown University

[5] Columbia University Stroud Center at New York State Psychiatric Institute

[6] Research Division, Hebrew Home at Riverdale; RiverSpring Health

[7] Division of Geriatrics and Palliative Medicine, Weill Cornell Medical College

[8] RTI Health Solutions

version of the survey and 134 Asians the Chinese version. One PROMIS item had poor discrimination as it was the only positively worded question in the fatigue measure. Among Hispanics, no DIF was found with the Wald test, while the OLR method identified five items with DIF comparing the English and Spanish versions; however, the effect of DIF on scores was negligible (r^2 ranged from .006 - .015). For the English and Chinese translations, no single item was consistently identified by both DIF tests. Minimal or no impact was observed on the overall scale score comparisons among Whites, Blacks, Hispanics, and Asians using the English language scales. However, greater numbers of items with DIF appeared when comparing Asians/ Pacific Islanders with Whites, Blacks, and Hispanics. “How often were you too tired to think clearly” showed consistent DIF.

Conclusions: Twelve of 14 PROMIS fatigue items performed well across the ethnically/racially diverse samples with minimal findings of DIF that would have any effect on comparing or combining scores across cancer populations. Supporting evidence of the validity and reliability of the PROMIS measures will enhance the adoption of the measures in oncology clinical research.

Keywords: differential item functioning, cancer, PROMIS, fatigue, patient-reported outcomes

Introduction

Fatigue is the most prevalent symptom related to cancer and is present in individuals before diagnosis and during and after treatment ends (Barsevick et al., 2010; Reeve et al., 2009). A recent review of the literature found that approximately 60 % of cancer patients experience moderate to severe fatigue during active treatment (Reilly et al., 2013).

Fatigue is also one of the most distressing symptoms as it affects functioning and quality of life (Barsevick et al., 2010). A review of the qualitative literature found that cancer patients commonly describe their fatigue experience as “extreme” and very different from normal fatigue that most people experience; They were “drained” and felt “like a zombie” (Scott, Lasch, Barsevick, & Piault-Louis, 2011). The qualitative study also found fatigue to have deleterious effects on their lives including cognitive impact (“my brain’s gone”), emotional impact (“depression”, “frustration”), physical impact (“trouble keeping up with work”, “too tired to get out of bed”), and social impact (“affects relationship with my kids, …with my [spouse], and ...my friends”; Scott et al., 2011). Observational studies with cancer patients have found fatigue to also co-occur with poor sleep quality, depressed mood, and pain; forming a symptom cluster that severely impacts quality of life (Piper & Cella, 2010). Given its prevalence and quality of life impact, fatigue was recommended by an expert panel of researchers, patient advocates, clinicians, and regulators to be measured in all oncology treatment trials (Reeve et al., 2014) and comparative effectiveness research studies (Basch et al., 2012).

In 2004, the National Institutes of Health (NIH) launched a collaboration with multiple academic centers to create the Patient-Reported Outcomes Measurement Information System® (PROMIS®; Cella et al., 2010). The goal of the PROMIS initiative was to develop standardized, valid, and reliable measures of patient-reported outcomes (PROs) that could be used in disease and non-disease populations and in different arenas of application including clinical research, healthcare delivery settings, and population surveillance. Given its prevalence in multiple disease settings and in the general population, fatigue was one of the first PRO domains created in the network (Cella et al., 2010; Garcia et al., 2007), and the item bank fatigue items were evaluated by Lai et al. (2011). In 2010, the National Cancer Institute's (NCI's) Clinical Trials Planning Meeting on cancer-related fatigue adopted the PROMIS Fatigue measure as the standard to use in oncology clinical trials (Barsevick et al., 2013).

However, despite the widespread adoption of PROMIS measures globally, there lacks an extensive psychometric evaluation of the PROMIS Fatigue measure across ethnically and educationally diverse subgroups in general and among cancer patients in particular. Although cancer patients were included in the original validation and item response theory (IRT)-based calibration of the PROMIS measures, they were pooled with other populations (with and without disease) forming a sample of more than 21,000 individuals to examine the psychometric properties of the PROMIS measures and estimate IRT parameters for the PROMIS item banks (Cella et al., 2010; Lai et al., 2011).

This study evaluates the psychometric properties of the PROMIS Fatigue measure in an ethnically/racially diverse population-based sample of adult cancer patients in the United States. Of particular focus, we tested for differential item functioning (DIF) across key populations for which there may be high likelihood for DIF including language translation (English, Spanish, and Chinese), race/ethnicity (non-Hispanic Whites (NHW), non-Hispanic Blacks (NHB), Hispanics, non-Hispanic Asians/Pacific Islanders (NHAPI)), gender, and age at cancer diagnosis (21 - 49, 50 - 64, and 65 - 84 years). DIF occurs when members from different groups (e.g., race/ethnicity) with the same level of the measured outcome (e.g., fatigue) respond differently to a particular question. Measures with items with DIF will have reduced validity for between group comparisons. A good practice in DIF detection is to perform sensitivity analyses with a second method in order to examine consistency of findings. Thus, we used two different DIF methods including an IRT-based method: the Wald test based on Lord's chi-square (Lord, 1980) as the primary method and the observed score ordinal logistic regression (OLR) model (Zumbo, 1999) as the secondary method. The latter is an extension for polytomous data of the logistic regression approach for binary data (Swaminathan & Rogers, 1990). In addition, the validity of our DIF findings was informed by a group of content experts who reviewed the PROMIS fatigue items without the data results to consider which items are likely to present with DIF.

Methods

Participants

The Measuring Your Health (MY-Health) study recruited cancer patients from four US SEER (Surveillance, Epidemiology, and End Results) Program cancer registries: The Greater Bay Area Cancer Registry covering the San Francisco Bay and surrounding area, the Cancer Registry of Greater California covering the rest of the state except Los Angeles County, the Louisiana Tumor Registry, and the New Jersey State Cancer Registry. The sample was stratified by age and race/ethnicity to ensure a sufficient number of younger and race-ethnic minorities participated. Eligibility criteria required patients to be: 21 - 84 years at diagnosis, a primary diagnosis of one of seven cancers (prostate, colorectal, non-small cell lung, Non-Hodgkin lymphoma, female breast, uterine or cervical), within 6 - 13 months of diagnosis, have no prior cancer diagnosis (except non-melanoma skin cancer), and able to read and respond to questionnaires in English, Spanish or Chinese (Mandarin). Participants were diagnosed between 2010 - 2013. Participants received a $30 gift card or check after completing the survey. The study was approved by IRBs at all participating institutions.

Measures

The PROMIS Fatigue (version 1.0) item bank includes 95 items (i.e., questions and their respective response options) that have been extensively reviewed qualitatively and quantitatively and found to be valid and reliable indicators of fatigue. The items have been IRT-calibrated as a unidimensional measure of fatigue (Lai et al., 2011) that includes both fatigue experiences (e.g., severity) and their impact on patients' lives (e.g., physical, emotional, and social). Several short forms exist. A T-score metric, centered on the US general population which uses IRT-based scoring based on response patterns is available. Also, it is possible to convert raw scores or simple summed score to the T-score equivalent using the PROMIS conversion tables available through the PROMIS websites: www.HealthMeasures.net and www.AssessmentCenter.net.

Fourteen PROMIS Fatigue items were administered in the MY-Health study together with items from seven other PROMIS domains. PROMIS Fatigue items were selected based on the following criteria: 1) representation on commonly-used PROMIS Fatigue short forms (4a, 6a, 7a, 8a); 2) high reliability (i.e., high IRT-information functions); 3) frequent computerized-adaptive testing (CAT) selection at ½ and 1 standard deviations below the U.S. general population mean; and 4) content valid for cancer populations. The item designations from the short forms are shown in Table 1. Each item had five response options and higher scores indicate more fatigue. Spanish and Mandarin translations of the English version of the PROMIS measures followed a strict translation process (Eremenco, Cella, & Arnold, 2005). Details of this process are discussed in the overview paper on the MY-Health survey in this issue.

Patients' demographic information (education, employment status, income, marital status, and insurance coverage), comorbid conditions, and cancer treatment type were collected through self-report via mailed survey. Using the 2010 U.S. Census classification algorithms (Humes, Jones, & Ramirez, 2011), we created the following self-reported race-ethnicity categories: Non-Hispanic White, Non-Hispanic Black, Hispanic, and Asian/Pacific Islander. NHAPI includes Asian, Pacific Islander, and Native American/Alaskan participants. A sensitivity analysis of the three categories indicated similar trends across these race-ethnicity groups. If self-reported race/ethnicity was missing, SEER registry information was used. Clinical information was provided by the SEER Program cancer registry and included age, sex, date of cancer diagnosis, cancer type, and cancer stage.

Table 1:
PROMIS Fatigue Items Descriptive Statistics and Response Frequencies

#	Question Stem	Mean (Std Dev)	*n*	Response Options: *n* (%)					
			5023	Never	Rarely	Sometimes	Often	Always	Missing
1	How often did you feel tired? (SF7a)	3.12 (1.04)	4995	383 (7.62)	840 (16.72)	2009 (40.00)	1304 (25.96)	459 (9.14)	28 (0.56)
2	How often did you experience extreme exhaustion? (SF7a)	2.27 (1.16)	4986	1680 (33.45)	1301 (25.90)	1165 (23.19)	658 (13.10)	182 (3.62)	37 (0.74)
3	How often did you run out of energy? (SF7a)	2.67 (1.13)	4995	930 (18.51)	1262 (25.12)	1590 (31.65)	965 (19.21)	248 (4.94)	28 (0.56)
4	How often did your fatigue limit you at work (include work at home)? (SF7a)	2.57 (1.23)	4959	1267 (25.22)	1115 (22.90)	1378 (27.43)	874 (17.40)	325 (6.47)	64 (1.27)
5	How often were you too tired to think clearly? (SF7a)	2.06 (1.10)	4999	2053 (40.87)	1324 (26.36)	1024 (20.39)	479 (9.54)	119 (2.37)	24 (0.48)
6	How often were you too tired to take a bath or shower? (SF7a)	1.83 (1.06)	4995	2260 (52.96)	1050 (20.90)	863 (17.18)	317 (6.31)	105 (2.09)	28 (0.56)
7	How often did you have enough energy to exercise strenuously? (SF7a)	3.37 (1.30)	4938	461 (9.18)	889 (17.70)	1238 (24.65)	1040 (20.70)	1310 (26.08)	85 (1.69)
8	How often did you have to push yourself to get things done because of your fatigue? (SF8a)	2.65 (1.23)	4984	1147 (22.83)	1142 (22.74)	1401 (27.89)	919 (18.30)	375 (7.47)	39 (0.78)
				Not at all	A little bit	Somewhat	Quite a bit	Very much	Missing
9	How run-down did you feel on average? (SF4a, 6a, 8a)	2.49 (1.17)	4975	1143 (22.76)	1643 (32.71)	1095 (21.80)	826 (16.44)	268 (5.34)	48 (0.96)
10	I feel fatigued. (SF4a, 6a, 8a)	2.52 (1.20)	4985	1136 (22.62)	1623 (32.31)	1025 (20.41)	893 (17.78)	308 (6.13)	38 (0.76)
11	How fatigued were you on average? (SF4a, 6a, 8a)	2.53 (1.16)	4980	1047 (20.84)	1682 (33.49)	1102 (21.94)	872 (17.36)	277 (5.51)	43 (0.86)
12	I have trouble starting things because I am tired. (SF4a, 6a, 8a)	2.27 (1.24)	4991	1774 (35.32)	1343 (26.74)	898 (17.88)	685 (13.64)	291 (5.79)	32 (0.64)
13	How much were you bothered by your fatigue on average? (SF4a, 6a, 8a)	2.45 (1.25)	4945	1380 (27.47)	1471 (29.29)	932 (18.55)	820 (16.32)	342 (6.81)	78 (1.55)
14	To what degree did your fatigue interfere with your physical functioning? (SF4a, 6a, 8a)	2.37 (1.26)	4933	1585 (31.55)	1344 (26.76)	911 (18.14)	777 (15.47)	316 (6.29)	90 (1.79)

Short form (SF) items are in parentheses

Item review by content experts

DIF hypotheses were generated by asking eight clinicians and other content experts to indicate whether or not they expected DIF to be present, and the direction of the DIF with respect to gender, age, race/ethnicity, and language. Three of the members of the panel were clinical or counseling psychologists, two were public health professionals. Of the remaining three, one was a gerontologist, another an epidemiologist, and one did not specify his/her specialty. Experts were provided a detailed definition of DIF to help guide their evaluation of the 14 PROMIS fatigue items.

Analyses

Psychometric analyses of the PROMIS Fatigue measure were consistent with the standards set by the PROMIS network (Reeve et al., 2007). To confirm unidimensionality of the PROMIS Fatigue items used in these analyses, we ran exploratory (EFA) and confirmatory factor analysis (CFA) in MPLUS software (version 7.1, Los Angeles, CA) using the WLSMV estimator (the weighted least squares with adjustments for the mean and variance). The sample was randomly split in half to run exploratory analyses on one half and confirmatory analyses on the other. Model fit statistics and criteria for good fit included the comparative fit index (CFI > 0.95), Tucker-Lewis Index (TLI > 0.95), standardized root mean residuals (SRMR < 0.08), weighted root mean square residual (WRMR < 1.0), and the root mean square error of approximation (RMSEA < 0.06). Not one statistic is universally accepted for all tests of model fit, thus we looked at multiple indicators and multiple analyses to evaluate dimensionality.

IRT-based modeling was carried out using IRTPRO (version 2.1, Skokie, IL). Consistent with the PROMIS psychometric approach, we fit Samejima's Graded Response Model (GRM) to the response data (Samejima, 1969, 1997). IRT model fit was evaluated using the generalization of Orlando and Thissen's $S\text{-}X^2$ indicator for polytomous-response data (Orlando & Thissen, 2000, 2003). Local dependence (LD) was evaluated using Chen and Thissen's G^2 LD index (Chen & Thissen, 1997). For each item, the IRT model provided an estimate of the discrimination parameter and four threshold parameters (5 response options – 1) that describe the likelihood of an individual to respond to one of the five response options conditional on their level of the PRO domain measured by the PROMIS scale. The discrimination parameter (*a*) is an estimate of how well the item discriminates among individuals who have different levels of the measured PRO domain and is also an indicator of how well the item relates to the overall PRO domain being measured by the PROMIS scale. The threshold parameters (*b*) provide estimates of the severity of the item as it relates to measuring different levels of the measured PRO domain. DIF tests described below evaluate DIF in both the discrimination parameters (non-uniform DIF) and the threshold parameters (uniform DIF).

The primary analyses were conducted using the Wald test for DIF, implemented in the software program Item Response Theory for Patient Reported Outcomes (IRTPRO; Cai, Thissen, & du Toit, 2012), which follows the model proposed by Lord (1977,

1980) in which vectors of IRT item parameters are compared. The rationale is that if the vectors of item parameters differ significantly across groups, then the item functions differently for the groups. IRTPRO provides Wald tests to detect DIF for the discrimination parameters and for the threshold parameters. Because the Wald test uses IRT models, DIF tests were only completed if the sample size in each comparison group was 250 or higher.

Sensitivity Analyses: The OLR test for DIF uses the cumulative information of the ordinal responses by comparing the odds of endorsing a response less or equal to *k* versus a response greater than *k.* The odds of the responses, as dependent variable of the logistic regression model, are predicted by the independent variables including group membership, the latent trait of participants, and their interaction. This method will provide a test of DIF of the relationship between the item response and the group membership conditional on the latent trait of the subjects. The OLR method provides tests of uniform DIF with main effects for the OLR model, which represent DIF in the threshold parameters in IRT. OLR methods also test for non-uniform DIF with an interaction term in the OLR model, which represents DIF in the discrimination parameters. The OLR method was applied using a SAS macro designed by one of the authors (W. -H. C.).

For each paired-group examined, we ran multiple analyses. The first attempt was to identify a group of items without DIF to be used as the anchor in the second set of DIF tests of the remaining items. Because we ran multiple DIF tests, we used a Bonferroni-adjusted *p* - value of .004 to identify an item that may have DIF. In addition to examining the significance (*p* - value), magnitude of the DIF was further evaluated by examining the expected item scores and estimating the effect sizes ($R^2 > .13$ indicative of salient DIF; Zumbo & Thomas, 1997). Briefly, the method is to examine nested models, entering the trait variable, followed by the studied group variable, a test of uniform DIF, and then the interaction term to examine non-uniform DIF. The difference in the R-square between the baseline (trait only) model and the last model with group and interaction terms entered is the estimated effect size. Details of this method are described in the overview to this set of articles. The OLR method was used if the group sample sizes were too small; However, we only used the OLR method if the sample size in each group was at least 100 people or more.

Results

Sample Characteristics: The demographic and clinical characteristics for the 5,507 cancer patients included in this study are provided in Table 2. The sample included 2,278 non-Hispanic Whites, 1,122 non-Hispanic Blacks, 1,053 Hispanics (*n* = 338 took the Spanish version), and 917 Asians/Pacific Islanders (*n* = 134 took the Chinese version). Our sample included 1,207 individuals between 21 to 49 years of age at diagnosis, 2,016 individuals between 50 and 64 years of age and 2,248 individuals 65 - 84 years of age at diagnosis. Approximately 59 % of the sample were female and 58 % were married. A majority of patients (95 %) completed the survey within 6 to 12 months from cancer diagnosis. The largest representation by cancer type included breast (30 %), prostate (21

%), colorectal (17 %), and lung (13 %). Cancer treatments included surgery (68 %), chemotherapy (48 %), radiation (57 %), and hormonal therapy (22 %).

Table 2:
Patient Sample Demographic and Clinical Characteristics

Characteristic	**Total**		**English survey language**		**Spanish survey language**		**Chinese survey language**	
	***n* = 5507**	**%**	***n* = 5023**	**%**	***n* = 342**	**%**	***n* = 142**	**%**
Race/Ethnicity								
Non-Hispanic White	2278	41.37	2277	45.33	1	0.29	0	0
Non-Hispanic Black	1122	20.37	1121	22.32	1	0.29	0	0
Hispanic	1053	19.12	710	14.13	8	98.83	5	3.52
NHAPI	917	16.65	782	15.57	1	0.29	134	94.37
Multiple	133	2.42	130	2.59	0	0	3	2.11
Missing	4	0.07	3	0.06	1	0.29	0	0
Age at DX								
Age 21-49	1207	21.92	1077	21.44	104	30.41	26	18.31
Age 50-64	2016	36.61	1828	36.39	135	39.47	53	37.32
Age 65-84	2248	40.82	2082	41.45	103	30.12	63	44.37
Missing	36	0.65	36	0.72	0	0	0	0
Education								
<High School Grad	975	17.70	721	14.35	199	58.19	55	38.73
High School Grad	1055	19.16	975	19.41	55	16.08	25	17.61
Some college	1765	32.05	1686	33.57	57	16.67	22	15.49
College Degree	988	17.94	961	19.13	12	3.51	15	10.56
Graduate Degree	644	11.69	623	12.40	4	1.17	17	11.97
Don’t know	15	0.27	11	0.22	2	0.58	2	1.41
Missing	65	1.18	46	0.92	13	3.80	6	4.23
Employment								
Working	2418	43.91	2208	43.96	164	47.95	46	32.39
Not Working	2982	54.15	2731	54.37	164	47.95	87	61.27
Missing	107	1.94	84	1.67	14	4.09	9	6.34
Married								
No	2246	40.78	2050	40.81	161	47.08	35	24.65
Yes	3200	58.11	2927	58.27	171	50.00	102	71.83
Missing	61	1.11	46	0.92	10	2.92	5	3.52
Survey Mode								
Mailed Survey	5409	98.22	4933	98.21	336	98.25	140	98.59
Phone Survey	98	1.78	90	1.79	6	1.75	2	1.41

Characteristic	**Total**		**English survey language**		**Spanish survey language**		**Chinese survey language**	
	***n* = 5507**	**%**	***n* = 5023**	**%**	***n* = 342**	**%**	***n* = 142**	**%**
Income								
Less than $10,000	584	10.60	484	9.64	78	22.81	22	15.49
$10,000 to $59,999	2176	39.51	1956	38.94	165	48.25	55	38.73
$60,000 to $99,999	912	16.56	894	17.80	11	3.22	7	4.93
$100,000 to $199,999	680	12.35	668	13.30	0	0	12	8.45
$200,000 or more	189	3.43	187	3.72	0	0	2	1.41
Don’t know/Unsure	354	6.43	293	5.83	47	13.74	14	9.86
Refuse to answer	385	6.99	364	7.25	9	2.63	12	8.45
Missing	227	4.12	177	3.52	32	9.36	18	12.68
Insurance								
Private	2273	41.27	2150	42.80	72	21.05	51	35.92
Government	1627	29.54	1390	27.67	171	50.00	66	46.48
Private + Government	1321	23.99	1286	25.60	29	8.48	6	4.23
None	114	2.07	94	1.87	19	5.56	1	0.70
Don’t know, Unsure	90	1.63	47	0.94	30	8.77	13	9.15
Missing	82	1.49	56	1.11	21	6.14	5	3.52
Time from DX to survey completion								
6 - 9 months	2698	48.99	2445	48.68	176	51.46	77	54.23
10 - 12 months	2530	45.94	2321	46.21	148	43.27	61	42.96
13 - 15 months	240	4.36	223	4.44	14	4.09	3	2.11
15+ months	39	0.71	34	0.68	4	1.17	1	0.70
Comorbidity count								
0 conditions	1246	22.63	1091	21.72	100	29.24	55	38.73
1 condition	1323	24.02	1230	24.49	63	18.42	30	21.13
2+ conditions	2938	53.35	2702	53.79	179	52.34	57	40.14
Sex								
Male	2208	40.9	2022	40.25	128	37.43	58	40.85
Female	3263	59.25	2965	59.03	214	62.57	84	59.15
Missing	36	0.65	36	0.72	0	0	0	0
Cancer Type								
Breast	1646	29.89	1481	29.48	112	32.75	53	37.32
Cervix	152	2.76	129	2.57	22	6.43	1	0.70
Colorectal	931	16.91	834	16.60	54	15.79	43	30.28
Lung	723	13.13	706	14.06	13	3.80	4	2.82
NHL	466	8.46	418	8.32	37	10.82	11	7.75
Prostate	1161	21.08	1064	21.18	77	22.51	20	14.08
Uterus	392	7.12	355	7.07	27	7.89	10	7.04
Missing	36	0.65	36	0.72	0	0	0	0

Characteristic	**Total**		**English survey language**		**Spanish survey language**		**Chinese survey language**	
	***n* = 5507**	**%**	***n* = 5023**	**%**	***n* = 342**	**%**	***n* = 142**	**%**
Did you ever have surgery as part of your cancer treatment?								
Yes	3733	67.79	3410	67.89	219	64.04	104	73.24
No	1691	30.71	1551	30.88	107	31.29	33	23.24
Don’t know	25	0.45	15	0.30	7	2.05	3	2.11
Missing	58	1.05	47	0.94	9	2.63	2	1.41
Did you ever receive any chemotherapy as part of your cancer treatment?								
Yes	2632	47.79	2371	47.20	184	53.80	77	54.23
No	2766	50.23	2565	51.07	139	40.64	62	43.66
Don’t know	22	0.40	17	0.34	4	1.17	1	0.70
Missing	87	1.58	70	1.39	15	4.39	2	1.41
Did you ever receive any hormonal therapy as part of your cancer treatment?								
Yes	1203	21.84	1094	21.78	78	22.81	31	21.83
No	4059	73.71	3736	74.38	230	67.25	93	65.49
Don’t know	139	2.52	111	2.21	15	4.39	13	9.15
Missing	106	1.92	82	1.63	19	5.56	5	3.52
Did you ever receive any radiation therapy as part of your cancer treatment?								
Yes	2253	40.91	2072	41.25	130	38.01	51	35.92
No	3164	57.45	2886	57.46	197	57.60	81	57.04
Don’t know	25	0.45	17	0.34	3	0.88	5	3.52
Missing	65	1.18	48	0.96	12	3.51	5	3.52

Descriptive Statistics: The PROMIS fatigue items’ mean, standard deviation, and response frequencies are provided in Table 1. The first eight items listed in the table measure frequency of fatigue experiences using response options of *never*, *rarely*, *sometimes*, *often*, and *always*, and the last six items measure amount (magnitude) using response options of *not at all*, *a little bit*, *somewhat*, *quite a bit*, and *very much*. Item #7 is the only one of the administered items to be framed in a positive manner that was reverse scored for analyses.

Dimensionality: EFA for the 14 items was run on one half of the sample and showed satisfactory fit for most indicators (CFI = 0.993, TLI = 0.992, RMSEA = 0.115, and SRMR = 0.026). However, item #7 had very poor loading (0.102) relative to the other items (range 0.819–0.956). We dropped item #7, reran the EFA and obtained similar fit (CFI = 0.993, TLI = 0.992, RMSEA = 0.119, and SRMR = 0.025), and the loadings ranged from 0.819 – 0.956. CFA for the 13 remaining items on the 2nd half of the sample showed satisfactory fit as well for most of the indices (CFI = 0.993, TLI = 0.992, RMSEA = 0.123, and WRMR = 2.857) with item loadings ranging from 0.813 to 0.965. Although somewhat below the conventional thresholds, the values for the fit statistics are within an acceptable range, given the skewed nature of the data and the sensitivity to inconsequential multidimensionality (Cook, Kallen, & Amtmann, 2009). Consistent with the PROMIS Fatigue item bank, the results support the unidimensionality of the item set, allowing us to apply the IRT model and other statistics to test for DIF in the item set without item #7.

IRT Model Fit: The IRT GRM was fit to the full sample for the 13 items (results not shown). The $S\text{-}X^2$ indicator showed significant *p* - values (indicating lack of fit) for all items except item #2. It is most likely that the $S\text{-}X^2$ statistics are inflated by the large sample size. In this case, the item model fit cannot be appropriately assessed. Two item pairs showed potential LD: items #10 and #11 ($\chi^2 = 49.8$) and items #13 and #14 ($\chi^2 = 44.4$). To examine the degree to which LD resulted in biased parameter estimates, we removed an item from each LD pair and observed how discrimination parameters changed. While the discrimination parameter's magnitude for the LD-paired item decreased when removing one of the LD items, the other items' discrimination parameters remained relatively unchanged. Thus, we kept both pairs of items in the model for DIF testing. However, we repeated DIF testing removing items with possible LD as a sensitivity test, and found that the DIF results were similar. Results provided below include both item pairs. It is important to note that the IRT item discrimination parameters for the first seven items (which use a frequency response scale) ranged from 2.45 to 3.98 and for the last six items (which use an amount response scale) ranged from 4.58 to 6.44.

DIF testing with the English version of the PROMIS Fatigue measure

Review of Fatigue items by content experts: For the 13 items formally reviewed for DIF, the eight content experts did not hypothesize any DIF based on group differences by race/ethnicity or language. Two experts each felt that women would be more likely to report higher levels of fatigue (uniform DIF) associated with items #1 and #4. More items with DIF by age were expected among the content experts. Older individuals were expected to be more likely to report higher levels of fatigue (uniform DIF) for items: #1 (4 experts), #2 (3 experts), #3 (4 experts), #4 (3 experts), #9 (2 experts), #10 (3 experts), #12 (3 experts), and #14 (2 experts).

Results of DIF testing within the English version of the PROMIS Fatigue measure are provided in Table 3. Item #7 was removed prior to DIF testing. Results for both the Wald test and the OLR method are provided next to each other for each race/ethnic group comparisons. In Table 3, the letter "D" stands for DIF detected in the discrimination parameter and the letter "T" stands for DIF detected in the threshold parameters. For all of the items identified as having DIF as described below, the effect of DIF on the PROMIS fatigue scores (for the 13 items) was negligible (R^2 ranged .006-.015).

Table 3:
Tests for Differential Item Functioning in English Language Version of the PROMIS Fatigue Measure*

Item	Race/Ethnicity												Age at diagnosis						Gender	
	White n=2277	Black n=1121	White n=2277	Hisp n=710	White n=2277	Asian n=782	Black n=1121	Hisp n=710	Black n=1121	Asian n=782	Hisp n=710	Asian n=782	21-49 n=1077	50-64 n=1828	21-49 n=1077	65-84 n=2082	50-64 n=1828	65-84 n=2082	Males n=2022	Females n=2965
	Wald	OLR	Wald	OLR	Wald	OLR	Wald	OLR	Wald	OLR	Wald	OLR	Wald	OLR	Wald	OLR	Wald	OLR	Wald	OLR
1			T(18.5)			D(29.4)				D(12.1)		D(29.4)		D(22.2)		D(37.1), T(13.1)		T(14.4)		D(100.1)
2	T(34.3)			D(22.9)	T(17.6)	T(16.1)				D(9.7)		D(23.2)				D(20.6)		D(14.7)		D(17.9)
3	T(24.8)	T(12.1)			T(24.1)	D(51.6)			T(20.6)	D(33.4)		D(48.8)				D(9.5)		T(14.3)		D(57.9)
4	T(20.8)			D(14.3)		D(12.0)			T(24.6)	D(19.8)		D(32.7)				D(17.2)		D(9.6), T(10.6)		D(25.7), T(9.3)
5				D(12.1)	T(40.9)	T(49.6)	D(8.9)		T(31.8)	T(39.6)	T(18.0)	D(18.8), T(16.6)	T(18.2)		T(30.2)	D(12.1), T(10.4)	D(10.0), T(20.6)			T(8.4)
6	T(21.8)	T(18.1)	T(16.1)		T(42.4)	T(38.9)					T(18.6)	T(8.4)								
8				D(10.4)	D(7.6), T(18.2)	D(33.4)			T(19.9)	D(43.2), T(10.9)		D(53.6), T(13.9)				D(22.1)		D(9.8)		D(43.2)
9			T(15.3)		T(18.3)	D(9.1), T(10.6)						D(15.6)		D(8.33)		D(10.7)				D(31.3)
10	D(8.8)																			D(36.9)
11	D(19.8)																			D(29.1)
12												D(12.2)					D(10.6)			D(29.4)
13	D(13.5)		T(16.5)									D(10.5)				D(19.8), T(10.3)				D(33.5)
14	T(19.5)		T(19.3)										T(19.5)							D(24.4)

*Note: Numbers reported in cells are χ^2. Only significant DIF is reported in the table based on threshold of $p < .004$ (.05/13). D = DIF in the Discrimination Parameter. T = DIF in the Threshold parameter. None of the significant DIF findings from the OLR method passed the r^2 threshold (as effect size measure) of 0.13. Item #7 from previous tables was not included in DIF testing due to poor psychometric properties.

Race/ethnicity: The Blacks versus Hispanics comparison resulted in only one item (#5) being identified by the Wald test as having tested positive for DIF. However, the other race/ethnic group comparisons identified several items with DIF, but the number of positive items with DIF varied by groups compared and the DIF method used. Items #3, #5, and #6 were identified with DIF by both methods in three race/ethnic group comparisons; however, the magnitude of DIF never reached the OLR threshold in sensitivity analyses.

As an example of an item with DIF, Figure 1 shows the IRT item characteristic curves (ICCs) for non-Hispanic Whites and Asians/Pacific Islanders that tested item #5 positive for DIF in the threshold parameters (Wald test: $\chi^2 = 40.9$ ($df = 4$), $p = .0001$). The ICCs suggest that, after controlling for underlying differences in fatigue levels, Asians/ Pacific Islanders are more likely to endorse responses reflecting higher frequency for the item "How often were you too tired to think clearly?" than non-Hispanic Whites.

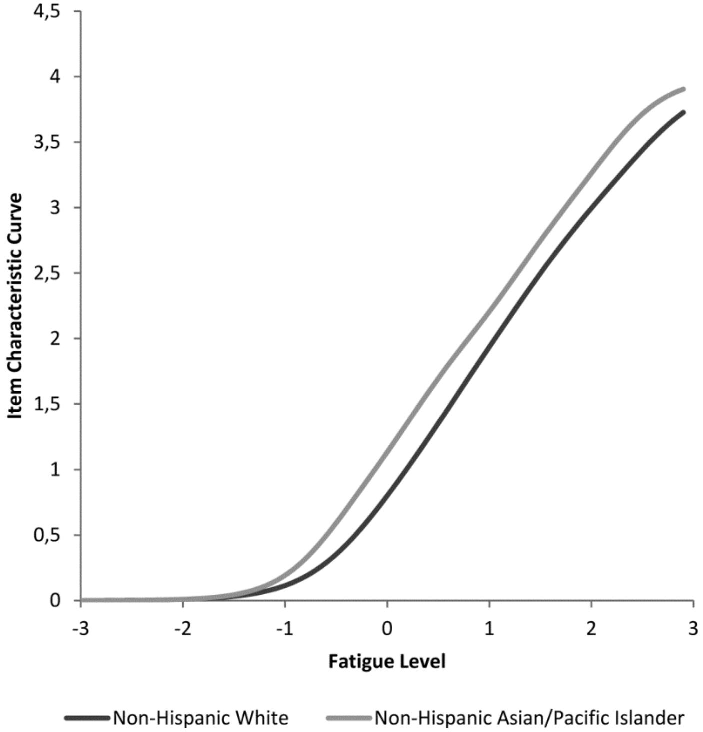

Figure 1:
Item response theory (IRT) - item characteristic curves (ICCs) for non-Hispanic Whites and Asians/Pacific Islanders for PROMIS Fatigue item #5 ("How often were you too tired to think clearly?")

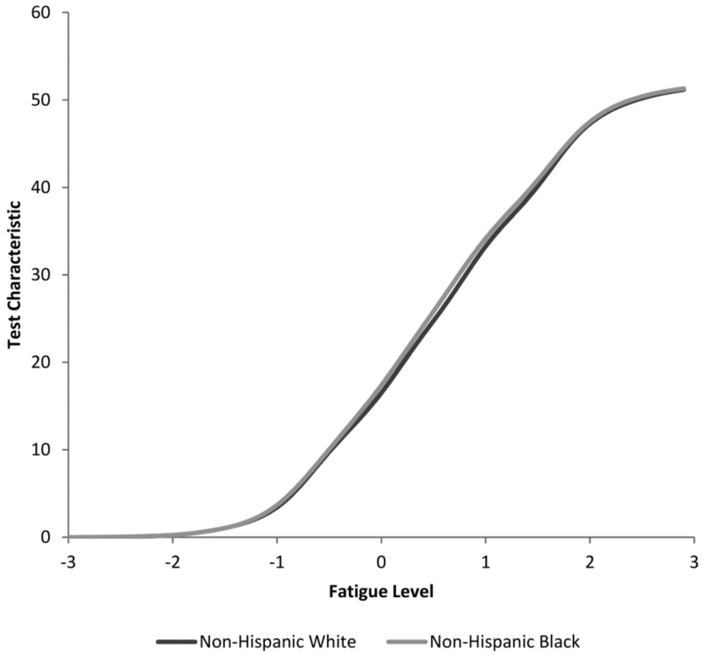

Figure 2:
Item response theory (IRT) Test characteristic curves for non-Hispanic Whites and non-Hispanic Blacks for the 13-item PROMIS Fatigue Measure showing overall impact of DIF from the eight items tested positive by the Wald method

As an example of the (low) impact of DIF, Figure 2 shows the IRT test characteristic curves (TCCs) for non-Hispanic Whites and Blacks for the 13-item PROMIS Fatigue Measure, including the eight items that tested positive by the Wald method. The overlapping TCCs suggest that even with inclusion of the items with DIF, there is a negligible difference in the estimated group for the 13-item set.

Age and Gender: For age group comparisons, item #5 appeared to show DIF consistently for both methods. For gender comparisons, the Wald test did not detect any item with DIF and the OLR method found every item to have DIF except item #6. Neither magnitude of DIF findings for age or gender reached the OLR threshold in sensitivity analyses.

DIF testing with Spanish and Chinese versions of the PROMIS Fatigue measure

Table 4 presents results from DIF tests comparing the different translations of the PROMIS Fatigue measure. We were limited in the number of DIF tests we could conduct within Spanish or Chinese due to small sample sizes. For example, we could not conduct

DIF tests by age groups within either non-English language. For all the items identified with DIF as described below, the effect of DIF on the PROMIS fatigue scores (for the 13 items) was negligible (R^2 ranged from 0.006 - 0.015).

The first set of comparison tests for DIF in the English and the Spanish versions included only Hispanic participants in the study. We did not want to include other races/ethnicities in the English – Spanish version comparisons to minimize possible confounding. Only the OLR method identified items with DIF (#3, #6, #8, #11, and #12).

The second comparison tested for gender DIF within Spanish language, and only the OLR method was used due to small sample sizes in males. Items #1 to #5 were detected with uniform DIF. Within Spanish, item #12 was detected for non-uniform DIF when testing across the age groupings.

The third set of comparison tests for DIF between the English and Chinese versions included only Asian participants in the study. Again, only the OLR method was used due to small sample sizes in the Chinese version. Only items #2 and #4 was identified as exhibiting DIF.

Table 4:
Tests for Differential Item Functioning in Spanish and Chinese Language Version of the PROMIS Fatigue Measure*

Item	English vs Spanish		Spanish only		Spanish only			English vs Chinese	
	English n = 691	Spanish n = 328	Males n = 128	Female n = 214	21-49 yrs n = 104	50-64 yrs n = 135	65-84 yrs n = 103	English n = 782	Chinese n = 134
	Wald	OLR	OLR			OLR		OLR	
1			D(17.1)						
2			D(11.6)					D(9.7)	
3		D(11.8)	D(9.1)						
4			D(12.1)					D(14.2)	
5			D(9.7)						
6		T(9.9)							
8		D(15.4)							
9									
10									
11		D (8.4)							
12		D(10.3)				D(9.0)			
13									
14									

Note: Numbers reported in cells are χ^2 . Only significant DIF is reported in the table based on threshold of $p < .004$ (.05/13). D = DIF in the Discrimination Parameter. T = DIF in the Threshold parameter. None of the significant DIF findings from the OLR method passed the r^2 threshold (as effect size measure) of 0.13. Item #7 from previous tables was not included in DIF testing due to poor psychometric properties. For DIF testing with the Spanish version, only Hispanics were included. For DIF testing with the Chinese version, only Chinese were included.

Discussion

This study addresses the strong need to evaluate the evidence for the validity of the PROMIS Fatigue measure for use in a population-based cohort of cancer patients. Specifically, this article focuses on the assessment of differential item functioning across key demographic groups that are included in cancer outcomes research and in Spanish and Chinese (Mandarin) translations. If present, items with DIF pose a serious threat to the ability of the measure to estimate fatigue levels for individuals from different sociodemographic groups or who respond to different translated forms of the instrument. Items with DIF reduce the validity of across group comparisons or combining items into a scale because their scores may be indicative of a variety of attributes other than those the scale is intended to measure (Thissen, Steinberg, & Wainer, 1993).

The PROMIS Fatigue item bank includes 95 items, and was subjected to extensive psychometric analyses (Lai et al., 2011). Fourteen items were selected for this study (including items from other PROMIS item banks) based on their inclusion in available PROMIS short forms or because they evidenced good psychometric properties in the analyses of the item bank. One PROMIS item, "How often did you have enough energy to exercise strenuously?" evidenced poor discrimination, most likely because it was the only positively worded question in the fatigue scale. Likely, the item wording by itself is clear as it was reviewed qualitatively with cognitive testing methods to make sure it was comprehendible. In addition, our view is that there is nothing different about the cancer experience that would result in this item having poor discrimination, and content experts did not flag it as an item possibly affected by diagnosis. It is also not likely that item responses were affected by the paper administration of the survey in this MY-Health study; although the original PROMIS calibration was via computer, the paper and computer modes of administration have been found to be equivalent (Bjorner et al., 2014). Acquiescent responding is likely responsible for the poor performance of this item as respondents may have not noticed or gotten confused by the switch in meaning regarding the response metric. The item may perform well if better formatting was used for the question or other items with positive-wording were also administered alongside this item.

Method comparisons: Both the Wald test and OLR are powerful for identifying small deviations in the item parameters that may be indicative of DIF. As a result, every one of the 13 fatigue items was identified as possibly having DIF. However, follow-up tests found none of the items had a high magnitude (beyond the OLR threshold) or salient impact on the aggregated fatigue scale score when combined with the other items in the scale. For example, no items were hypothesized by experts to evidence DIF for race/ethnicity, and none with magnitude above threshold were identified with DIF. In other words, differences in scores between the comparison groups (e.g., non-Hispanic Whites and Blacks) when controlling or not for DIF in the item were negligible (e.g., see Figure 2 for an example of an item identified with among the largest DIF). However, a cautionary note is that the effect of DIF from an item could be more of an issue if a computerized adaptive test (CAT) were being used and the item with DIF was selected as one of maybe four or five items administered to the respondent or if the item with DIF was included on a 4-item short form.

Three items were identified with DIF with both the Wald test and the OLR method for multiple comparisons. These included item #3 ("how often did you run out of energy"), item #5 ("how often were you too tired to think clearly"), and item #6 ("how often were you too tired to take a bath or shower?"). Item #3 was also expected by our content experts to present with DIF by age group. It is beyond the scope of this study to determine what may be underlying reasons that account for DIF (see related discussion below), however further qualitative testing would be recommended to identify reasons and further quantitative testing should be done with new data to confirm findings.

It is interesting that items #1 to #6 and item #8 had, on average, more instances of DIF than items #9 to #14 (see Tables 3 and 4). The first seven items (excluding item #7, which was dropped from analyses) use a frequency scale (*never* to *always*) and the last six items use a response metric representing the amount or magnitude of fatigue (*not at all* to *very much*). In addition, the IRT discrimination parameters for the items with the amount response scale were much higher than the items using the frequency response scale. Follow-up qualitative studies would be needed to attempt to uncover possible reasons for these findings; however, we believe fatigue for cancer patients is a chronic, persistent symptom that is an unrelenting experience. Because the PROMIS Fatigue measure uses a seven-day recall period, different groups of patients may interpret the items with frequency response formats differently as the impact of fatigue may remain stable over the week. Thus, the items with the amount response options may be more relevant to capture the differential impact of fatigue on cancer patients' lives. It is important to note that the short form PROMIS scales were comprised primarily of the latter *amount* response category items rather than the *frequency* response category items.

A review of the literature found that no previous study evaluated DIF in the PROMIS Fatigue items in an adult cancer population or among ethnically diverse groups. A study (Lai, Cella, Yanez, & Stone, 2014) using the pediatric version of the PROMIS Fatigue measure found some age-DIF; however, the results are not generalizable to our study because the pediatric study was not conducted in a cancer population and wording of the questions in the adult and pediatric versions are quite different. We did find one study that evaluated DIF in the PROMIS Fatigue items in adults with disabilities including spinal cord injury, muscular dystrophy, post-polio syndrome and multiple sclerosis (Cook, Bamer, Amtmann, Molton, & Jensen, 2012). Using the latent variable OLR method (Choi, Gibbons, & Crane, 2011) this study found no evidence of age DIF for any of the PROMIS Fatigue items (Cook et al., 2012). However, item #6, "How often were you too tired to take a bath or shower?" showed some evidence of DIF by disease diagnosis type, but the impact of the DIF was considered negligible on respondent's overall fatigue scores (Cook et al.).

The Wald test has the advantage of requiring only one model, and therefore is very efficient and relatively less computationally intense than other DIF methods. It also has the advantage of testing DIF for more than two groups. However, the Wald test is IRT model based and thus relies on many IRT model assumptions as well as on robust estimation of item parameters and their error covariance matrix. The OLR method used in this study, on the other hand, relies on fewer assumptions and utilizes observed data. It only requires robust estimation of the latent traits for each subject. However, it is not as efficient as it tests one item at a time. It will also require re-computation of the latent trait for each participant should items be removed from the instrument. The Wald test and OLR methods are different approaches, each with different assumptions; thus convergence of significant findings is not guaranteed. Because the Wald test adjusts internally for multiple comparisons in multigroup testing of DIF, control of type I error (false DIF detection) may be better. Both methods require examination of the magnitude and impact of DIF. The use of a second method in sensitivity analysis is advantageous because only items identified by both methods with significant DIF of high magnitude are flagged, resulting in conservative decisions about items that are most likely to be problematic in practice. Although many studies have been conducted related to the performance of OLR, few studies have compared OLR and the Wald test head-to-head.

Limitations: A limitation of the study is that sample sizes for the Spanish and Chinese versions of the instrument were small, which did not allow us to make demographic comparisons of DIF as we did for the English version.

Conclusions

This is the first study of the performance of the fatigue short form items among ethnically diverse groups and among cancer patients. Evidence supports the structural validity of the PROMIS fatigue short form items in that they measure a single dimension of fatigue experience in ethnically diverse cancer patients. One positively-worded question (#7) did not perform well possibly because it was embedded with 13 other negatively-worded questions selected for inclusion in this study. DIF was detected among all of the items (some more than others); however, the magnitude and impact of the items on the total score was negligible. The Spanish and Chinese (Mandarin) translations of the PROMIS Fatigue measure appear to perform well quantitatively.

Future Directions: In our view, the PROMIS Fatigue short form measure items can be recommended for use in cancer populations. The existing PROMIS short forms include items that focus more on measuring fatigue severity (i.e., intensity or amount) rather than frequency of fatigue experience over the past seven days. Intensity may reflect better the unrelenting fatigue experienced by cancer patients in dealing with the effects of the disease and treatments. Further research is needed to confirm findings from this study and to determine what may be underlying causes of the DIF observed. The findings from this study add to the growing evidence-base related to the validity and reliability of the PROMIS Fatigue short forms, and support their adoption and use in oncology clinical research.

Acknowledgements

Funding for these analyses was provided by the National Institute of Arthritis & Musculoskeletal & Skin Diseases, U01AR057971 (PI: Potosky, Moinpour). The authors thank Stephanie Silver, MPH for editorial review of this paper.

References

Barsevick, A. M., Cleeland, C. S., Manning, D. C., O'Mara, A. M., Reeve, B. B., Scott, J. A., ... Ascpro. (2010). ASCPRO recommendations for the assessment of fatigue as an outcome in clinical trials. *Journal of Pain and Symptom Management, 39*(6), 1086-1099. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20538190 doi:10.1016/j.jpainsymman.2010.02.006

Barsevick, A. M., Irwin, M. R., Hinds, P., Miller, A., Berger, A., Jacobsen, P., ... National Cancer Institute Clinical Trials Planning, M. (2013). Recommendations for high-priority research on cancer-related fatigue in children and adults. *Journal of the National Cancer Institute, 105*(19), 1432-1440. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24047960 doi:10.1093/jnci/djt242

Basch, E., Abernethy, A. P., Mullins, C. D., Reeve, B. B., Smith, M. L., Coons, S. J., ... Tunis, S. (2012). Recommendations for incorporating patient-reported outcomes into clinical comparative effectiveness research in adult oncology. *Journal of Clinical Oncology, 30*(34), 4249-4255. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23071244 doi:10.1200/JCO.2012.42.5967

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24262772 doi:10.1016/j.jclinepi.2013.07.016

Cai, L., Thissen, D., & du Toit, S. H. C. (2012). IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... Group, P. C. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology, 63*(11), 1179-1194. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20685078 doi:10.1016/j.jclinepi.2010.04.011

Chen, W. H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1-30.

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, *18*(4), 447-460.

Cook, K. F., Bamer, A. M., Amtmann, D., Molton, I. R., & Jensen, M. P. (2012). Six patient-reported outcome measurement information system short form measures have negligible age- or diagnosis-related differential item functioning in individuals with disabilities. *Archives of Physical Medicine and Rehabilitation, 93*(7), 1289-1291. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22386213 doi:10.1016/j.apmr.2011.11.022

Eremenco, S. L., Cella, D., & Arnold, B. J. (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Evaluation & the Health Professions, 28*(2), 212-232. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15851774 doi:10.1177/0163278705275342

Garcia, S. F., Cella, D., Clauser, S. B., Flynn, K. E., Lad, T., Lai, J. S., ... Weinfurt, K. (2007). Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. *Journal of Clinical Oncology, 25*(32), 5106-5112. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17991929 doi:10.1200/jco.2007.12.2341

Humes, K. R., Jones, N. A., & Ramirez, R. R. (2011). *Overview of race and Hispanic origin: 2010*. (C2010BR-02). Retrieved from http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf.

Lai, J. S., Cella, D., Yanez, B., & Stone, A. (2014). Linking fatigue measures on a common reporting metric. *Journal of Pain and Symptom Management*. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24698661 doi:10.1016/j.jpainsymman.2013.12.236

Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., 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*(10), S20-S27. doi: http://dx.doi.org/10.1016/j.apmr.2010.08.033

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology : selected papers from the Third International Congress of the International Association for Cross-Cultural Psychology held at Tilburg University, Tilburg, the Netherlands, July 12-16, 1976* (pp. 19-29). Amsterdam: Swets & Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: L. Erlbaum Associates.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64. Retrieved from <Go to ISI>://000167783400003 doi:10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298. Retrieved from doi: 10.1177/0146621603027004004

Piper, B. F., & Cella, D. (2010). Cancer-related fatigue: definitions and clinical subtypes. *Journal of the National Comprehensive Cancer Network, 8*(8), 958-966. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20870639

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., ... Group, P. C. (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*(5 Suppl 1), S22-31. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17443115 doi:10.1097/01.mlr.0000250483.85507.04

Reeve, B. B., Mitchell, S. A., Dueck, A. C., Basch, E., Cella, D., Reilly, C. M., ... Bruner, D. W. (2014). Recommended patient-reported core set of symptoms to measure in adult cancer treatment trials. *Journal of the National Cancer Institute, 106*(7). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25006191 doi:10.1093/jnci/dju129

Reeve, B. B., Potosky, A. L., Smith, A. W., Han, P. K., Hays, R. D., Davis, W. W., ... Clauser, S. B. (2009). Impact of cancer on health-related quality of life of older Americans. *Journal of the National Cancer Institute, 101*(12), 860-868. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19509357 doi:10.1093/jnci/djp123

Reilly, C. M., Bruner, D. W., Mitchell, S. A., Minasian, L. M., Basch, E., Dueck, A. C., ... Reeve, B. B. (2013). A literature synthesis of symptom prevalence and severity in persons receiving active cancer treatment. *Supportive Care in Cancer, 21*(6), 1525-1550. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23314601 doi:10.1007/s00520-012-1688-0

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Richmond, VA: William Byrd Press.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Scott, J. A., Lasch, K. E., Barsevick, A. M., & Piault-Louis, E. (2011). Patients' experiences with cancer-related fatigue: a review and synthesis of qualitative research. *Oncology Nursing Forum, 38*(3), E191-203. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21531669 doi:10.1188/11.onf.e191-e203

Swaminathan H. & Rogers H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size of a model-based approach for studying DIF* The Edgeworth Series in Quantitative Behavioural Science. Working Paper. Edgeworth Laboratory for Quantitative Behavioral Science. University of British Columbia. Prince George, BC.