CrossMark

Assessment of the psychometrics of a PROMIS item bank: self-efficacy for managing daily activities

Ickpyo Hong[1] · Craig A. Velozo[2] · Chih-Ying Li[3] · Sergio Romero[4,5] ·
Ann L. Gruber-Baldini[6] · Lisa M. Shulman[7,8]

Abstract

Purpose The aim of this study is to investigate the psychometrics of the Patient-Reported Outcomes Measurement Information System self-efficacy for managing daily activities item bank.

Methods The item pool was field tested on a sample of 1087 participants via internet ($n = 250$) and in-clinic ($n = 837$) surveys. All participants reported having at least one chronic health condition. The 35 item pool was investigated for dimensionality (confirmatory factor analyses, CFA and exploratory factor analysis, EFA), item-total correlations, local independence, precision, and differential item functioning (DIF) across gender, race, ethnicity, age groups, data collection modes, and neurological chronic conditions (McFadden Pseudo R^2 less than 10 %).

Results The item pool met two of the four CFA fit criteria (CFI = 0.952 and SRMR = 0.07). EFA analysis found a dominant first factor (eigenvalue = 24.34) and the ratio of first to second eigenvalue was 12.4. The item pool demonstrated good item-total correlations (0.59–0.85) and acceptable internal consistency (Cronbach's alpha = 0.97). The item pool maintained its precision (reliability over 0.90) across a wide range of theta (3.70), and there was no significant DIF.

Conclusion The findings indicated the item pool has sound psychometric properties and the test items are eligible for development of computerized adaptive testing and short forms.

Keywords Patient-reported outcome measure · Self-efficacy · Daily activities · Item response theory

Introduction

The Patient-Reported Outcomes Measurement Information System (PROMIS) initiative aims to measure patient-reported health status for physical, mental, and social well-

✉ Craig A. Velozo
velozo@musc.edu

1 Department of Health Sciences and Research, Medical University of South Carolina, Rm B417, 151-B Rutledge Avenue, Charleston, SC 29425-9620, USA

2 Division of Occupational Therapy, Medical University of South Carolina, Rm 330, 151-B Rutledge Avenue MSC 962, Charleston, SC 29425-9620, USA

3 Department of Healthcare Leadership and Management, Medical University of South Carolina, Rm B415, 151-B Rutledge Avenue, Charleston, SC 29425-9620, USA

4 Department of Occupational Therapy, University of Florida, Gainesville, FL, USA

5 Department of Veterans Affairs, Center of Innovation on Disability and Rehabilitation Research, 1601 SW Archer Rd., Gainesville, FL 32608, USA

6 Division of Gerontology, Department of Epidemiology and Public Health, University of Maryland School of Medicine, Rm 213, Howard Hall, 660 West Redwood Street, Baltimore, MD 21201-1559, USA

7 Department of Neurology, University of Maryland Parkinson Disease and Movement Disorders Center, Baltimore, MD, USA

8 Parkinson's Disease and Movement Disorders and The Rosalyn Newman Distinguished Scholar in Parkinson's Disease, 110 Paca Street, Rm 3-S-127, Baltimore, MD 21201-1559, USA

being [1]. Self-efficacy is a subdomain of the mental health category on the PROMIS framework. This study is part of a larger study of the development and validation of self-efficacy for managing chronic disease in which five constructs were developed (managing symptoms, managing emotions, managing daily activities, managing medications and treatments, and managing social interactions). This paper focuses on one of the five constructs, the self-efficacy for managing daily activities. This construct was defined as assessing the subject's self-reported level of confidence performing various basic and instrumental activities of daily living without assistance. Self-efficacy is critical for individuals with chronic conditions in order to successfully perform effective self-care of their conditions [2, 3].

While various self-efficacy scales have been developed to measure disease-specific self-efficacy, such as stroke [4], epilepsy [5], multiple sclerosis [6], cancer [7], arthritis [8], and sickle cell disease [9], only one scale, the Self-Efficacy for Managing Chronic Disease 6-Item Scale (SEMCD6) has been widely used to measure self-efficacy for managing chronic conditions across multiple health conditions [10, 11]. The SEMCD6 includes items related to different domains, such as fatigue, pain, emotional stress, symptoms, activity, and medication; however, it does not provide subscale scores. To address the concerns about the lack of different subscales relevant to managing chronic conditions, we developed five self-efficacy item pools: self-efficacy for managing symptoms, managing emotions, managing daily activities, managing medications and treatments, and managing social interactions. These five constructs target self-efficacy for self-management of various chronic medical conditions and include understudied neurological conditions such as epilepsy, multiple sclerosis, Parkinson's disease, peripheral neuropathy and stroke.

In this paper, we report the psychometrics of the self-efficacy for managing daily activities instrument. The psychometric properties of the self-efficacy for managing daily activities construct were investigated with both classical test theory and item response theory (IRT) methods [12, 13]. Based on the two measurement theory approaches, the test items in the managing daily activities construct were evaluated at the test and item level to provide comprehensive information of item responses and their relationship to the construct. The summary statistics and scoring manuals for the five self-efficacy item banks are available at the PROMIS website (https://www.assessmentcenter.net/Manuals.aspx).

The overall goal of the study was to develop a patient-reported self-efficacy for managing chronic conditions scale that (1) could demonstrate sound psychometric properties of the IRT assumptions (unidimensionality, precision, and local independence), (2) could be used across different demographic groups, and (3) could demonstrate good reliability and validity.

Methods

Study participants

The field test was conducted in patients with diverse chronic conditions. A total of 1087 subjects completed the self-efficacy for managing daily activities item pool. Participants were recruited via two data collection modes between April 2013 and April 2014: a total of 837 participants with chronic neurologic conditions from the clinical setting (University of Maryland Neurology Ambulatory Center) and a total of 250 participants with general medical conditions from the internet survey completed the 35 test items. The neurologic conditions included epilepsy, multiple sclerosis, Parkinson's disease, peripheral neuropathy, and stroke, recruited from an ambulatory care center at the University of Maryland in Baltimore. The inclusion criteria for the neurologic sample were as follows: 18 years of age or older, reside in the community, and diagnosed with any of chronic conditions under study by a treating neurologist. We applied the following exclusion criteria for the clinical sample: (1) cognitive impairment demonstrated by a score ≤20 on the Montreal Cognitive Assessment, (2) inability to give informed consent due to language (e.g., aphasia that interferes with ability to complete questionnaires, or insufficient understanding of the English language, etc.), (3) severe and unstable medical or psychiatric co-morbidities, and (4) pregnant women, prisoners and institutionalized patients.

The general medical conditions internet sample was recruited through a national online panel testing company, Op4G (see op4g.com for more detail). Participants were randomly selected from approximately 250,000 members of Op4G, and the selected sample completed Internet-based surveys using their home computers. Participants had to have at least one of the following chronic medical conditions: chest pain (angina), hardening of the arteries (coronary artery disease), heart failure or congestive heart failure, heart attack (myocardial infarction), stroke or transient ischemic attack (TIA), liver disease, hepatitis, or cirrhosis, kidney disease, arthritis or rheumatism, asthma, chronic lung disease (COPD), chronic bronchitis or emphysema, migraines or severe headaches, diabetes or high blood sugar or sugar in your urine, cancer (other than non-melanoma skin cancer), HIV or AIDS, spinal cord injury, multiple sclerosis, Parkinson's disease, neuropathy, or epilepsy. They could have multiple conditions or others not listed here as long as they had one of these conditions.

They also had to be 18 years of age or older and reside in the community.

This study was approved by the Institutional review boards (IRB) of the Medical University of South Carolina (#Pro00033397), the University of Maryland (#HP-000432550), and the University of Florida (#261-2010).

Self-efficacy for managing daily activities item pool

The initial item pool was developed using the qualitative research methodology approved by the PROMIS group, including literature review of self-efficacy scales, developing an item library, binning and winnowing items, expert researcher ratings (which were done using a Delphi technique), focus groups with patients, cognitive interviews with people with chronic conditions, and expert item revision [14, 15]. Finally, the 35 test items of the self-efficacy for managing daily activities domain were field tested.

The self-efficacy for managing daily activities item pool assesses patient-reported, current level of confidence performing various basic and instrumental activities of daily living without assistance. The item bank is generic rather than disease-specific, instructing the subject to consider all health conditions and all symptoms in their responses. The item pool consisted of 35 test items and a 5-point rating scale: 1 (I am not at all confident); 2 (I am a little confident); 3 (I am somewhat confident); 4 (I am quite confident); and 5 (I am very confident).

Statistical analysis

The psychometric methods employed were introduced by PROMIS [12, 16]. Descriptive statistics were used to describe the demographic characteristics of the sample (Chi-square test, independent t test or Wilcoxon rank-sum test), and traditional (i.e., classic) psychometric methods were used to evaluate the central tendency (mean) and spread (standard deviations) of the test items. SPSS version 21 was used for the descriptive statistics [17]. Item-total correlation and Cronbach's coefficient alpha were used to evaluate internal consistency of the test items. Acceptable criteria for item-total correlation used were greater than 0.4 [18]. For Cronbach's coefficient alpha, we used a criterion of between 0.90 and 0.95 as acceptable for individual level of measurement [19]. The ceiling and floor effect were investigated and less than 15 % of the sample scored at the extreme (i.e., minimum and maximum) was considered as a criterion [20].

We used confirmatory factor analyses (CFA) and exploratory factor analysis (EFA) to investigate the underlying structure (dimensionality) of the item bank. Mplus version 7.11 was used to perform the factor analyses

[21]. CFA was selected over EFA as the first step [12] because the initial item pool was selected by experts and refined using qualitative research methodologies in order to represent self-efficacy for managing daily activities construct. The CFA was conducted with the weighted least squares with adjustments for the mean and variance (WLSMV) estimation and one-factor solution. The rating scales were treated as categorical variables [22]. Factor structures were analyzed using model fit indices, including comparative fit index (CFI > 0.95 for good fit), Tucker–Lewis Index (TLI > 0.95 for good fit), root mean square error of approximation (RMSEA < 0.08 for adequate fit and 0.06 for good fit), and standardized root mean square residual (SRMR < 0.08 for good fit) [12].

If CFA indicates poor model fit, we conducted EFA to investigate the magnitude of eigenvalues for the larger factors and factor loadings to detect the underlying structural patterns [12]. In the EFA, a polychoric correlations matrix was analyzed using the WLSMV estimation. The rating scale on the construct was treated as categorical variables [22]. We reported eigenvalues, the amount of variance explained by the model, the ratio of the first and second eigenvalues, and the factor loadings on the test items. The criteria for unidimensionality were as follows: at least 20 % of the variance explained by the first factor, greater than a value of 4.0 in the ratio of the first and second eigenvalues, and greater than a factor loading of 0.3 on test items [12].

The residual correlation matrix from the single factor CFA was investigated to evaluate local independence of the item bank. The criterion for the violation of the local independence was defined as a residual correlation greater than 0.2 with any of the remaining test items [12].

The IRT analyses were conducted using IRTPRO version 2.1 [23]. With a 2-parameter IRT model (the graded response model), the parameters (item discrimination and step thresholds) were estimated on each test item. We used $S - X^2$, a Pearson X^2 statistic, to investigate item fit to the measurement model [24, 25] and misfit criterion was a p value of less than 0.001 [16]. Standard error of measurement (SEM) was also calculated across theta values to test precision of the item bank. SEM is the reciprocal value of the test information function (TIF) at the estimated ability and is defined as $\mathrm{SEM}(\theta) = \sqrt{(1/Ij(\theta j))}$, where θj is the estimated ability, Ij is information. When the theta values of test items were calibrated with mean 0 and standard deviation 1, cut-off of SEM is 0.3, which is equivalent to reliability of 0.90 [26]. The SEM values were presented graphically over the difficulty level of test items in order to investigate how much the item bank attains measurement precision over the range of self-efficacy scores (T scores).

In this study, Logistic Regression Differential Item Functioning using IRT was used to detect the presence of differential item functioning (DIF) based on trait estimates (thetas) [27]. Five DIF analyses were conducted for gender (male/female), race (white/non-white), ethnicity (Hispanic/non-Hispanic), the age of the sample (under 65/over 65 years), data collection mode(clinical setting/internet survey) variables, and neurological chronic conditions (neuropathy, stroke, multiple sclerosis, Parkinson's disease and epilepsy). The comparison models were categorized as uniform (if the effect is constant), non-uniform (if the effect varies conditional on the trait level), and total DIF (sum of uniform and non-uniform DIF) [27]. The detection index for DIF was McFadden's pseudo R^2. The detection criterion for McFadden's pseudo R^2 was any values greater than 10 % [28].

Results

Demographic characteristics

Table 1 represents the demographic characteristics of the sample. Fifty-one percent were female. The sample median age was 55 (SD = 14.7) years. The majority of the people (37.3 %) are 50–64 years old with the range of 18–89 years. In the sample, 73.2 % were white and 23.7 % non-white. Twenty-two percent had attained an advanced degree. The missing data were less than 3 % for gender (2.9 %), race (3.0 %), ethnicity (2.8 %), and education (2.8 %). The severity of neurological chronic conditions for the clinical sample was rated by clinicians. The sample median disease duration was 7.0 (SD = 10.7) years, and 67.3 % had no-to-mild impairment. There were significant differences in all demographic variables between the clinical and internet sample ($p < 0.05$), except for gender.

Table 2 and Fig. 1 present descriptive statistics about the managing daily activities item pool. A majority of the test items were rated in the highest rating category and demonstrated a ceiling effect (range 33.1–86.0 %). The overall percentage for all the rating categories (1–5) was 6.3, 5.7, 10.2, 14.6, and 62.9 %, respectively. A majority of the items showed skewed response distributions to the ceiling (very confident). The test items demonstrated good item-total correlations (0.59–0.85) and acceptable for individual level of measurement (Cronbach's alpha = 0.97).

Dimensionality

The factor analysis with one-factor solution indicated that the CFI (0.952) and SRMR (0.070) met the model fit criteria, and TLI (0.949) was marginally under the model fit criteria; however, RMSEA (0.09) did not meet the model fit criteria. The EFA was conducted to calculate the eigenvalues on the possible underlying factors. The first factor showed an eigenvalue of 24.34, and the other factors showed an eigenvalue less than 2 (see Table 3). Table 4 represents the EFA factor loadings on the 35 test items. All test items showed a high factor loading (0.72–0.92). The sum squared loadings of the 35 items were 24.97, and the ratio of first to second eigenvalue was 12.4. The test items explained 71 % of variance.

High residual correlations (−0.22 to −0.29) were found between item 13—"Exercise at a vigorous level for 10 min" and item 2—"eat without help," 3—"personal hygiene without help," 26—"use telephone to schedule appointments," and 34—"can find new ways to manage daily activities when the old way doesn't work."

Model fit

Nine items (items 2, 3, 17, 22, 23, 25, 26, 34, and 35) misfit the measurement model (Table 4). In order to investigate the impact of the misfit items, we created a scatter plot of between person measures calibrated from the full item pool (35 items) and person measures calibrated from non-misfit items (26 items). In spite of the narrow error bands, there were no person measures located outside the 95 % confidence interval (i.e., over 5 % of the sample). Based on these findings, we concluded that the misfit items had a negligible or no effect on the person measures.

Item calibrations

Table 4 summarizes the item calibrations for the item bank. The most challenging test items were item 13—"exercise vigorously for 10 min," item 24—"keep doing my usual activities at work," and item 30—"maintain a regular exercise program." The average thresholds of the most challenging items were −0.13, −0.63, and −0.69, respectively. The least challenging test items were item 3—"personal hygiene without help," item 33—"take medications with correct dose and times," and item 2—"eat without help." The average thresholds of the least challenging items were −1.72, −1.76, and −1.86, respectively.

Precision

The theta values were converted into T scores with average 50 and standard deviation 10 for a US clinical population

Table 1 Demographic characteristics

Characteristic	Total sample N (%) N = 1087	Clinic sample N (%) N = 837	Internet sample N (%) N = 250	p value
Gender				0.329
Male	458 (42.1)	343 (40.9)	115 (46.0)	
Female	599 (55.1)	464 (55.4)	135 (54.0)	
Missing	30 (2.8)	30 (3.7)	0 (0)	
Race				<0.001
White	796 (73.2)	597 (71.3)	199 (79.6)	
Non-White	258 (23.7)	207 (24.7)	51 (20.4)	
Missing	33 (3.0)	33 (4.0)	0 (0)	
Ethnicity				<0.001
Hispanic	64 (5.9)	14 (1.7)	50 (20.0)	
Non-Hispanic	992 (91.3)	792 (94.6)	200 (80.0)	
Missing	31 (2.9)	31 (3.7)	0 (0)	
Education				<0.001
Less than high school	21 (1.9)	3 (0.4)	18 (7.2)	
High school diploma	190 (17.5)	167 (20.0)	23 (9.2)	
Some college	271 (24.9)	205 (24.5)	66 (26.4)	
College degree	336 (30.9)	244 (29.2)	92 (36.8)	
Advanced degree	239 (22.0)	188 (22.5)	51 (20.4)	
Missing	30 (2.8)	30 (3.6)	0 (0)	
Chronic conditions[a]				
Neuropathy	238 (22.0)	214 (25.6)	24 (9.6)	<0.001
Stroke	181 (16.8)	172 (20.5)	9 (3.6)	<0.001
Multiple sclerosis	178 (16.5)	166 (19.8)	12 (4.8)	<0.001
Parkinson's disease	171 (15.8)	169 (20.2)	2 (0.8)	<0.001
Epilepsy	169 (15.6)	160 (19.1)	9 (3.6)	<0.001
Severity of chronic conditions[b]				NA
No impairment	184 (16.9)	184 (22.2)	NA	
Mild impairment	374 (34.4)	374 (45.1)	NA	
Moderate/severe impairment	271 (24.9)	271 (32.9)	NA	
Missing	258 (23.7)	8 (1.0)	NA	
Disease duration (median, SD)	7.0 (10.7)	7.0 (10.7)	NA	NA
Age (median, SD)	55 (14.7)	54.4 (14.5)	51.8 (15.3)	0.048

[a] Participants had multiple chronic conditions

[b] Clinicians rated; *NA* Not applicable

of individuals with at least one chronic condition. The raw scores of the item pool were converted into T scores ("Appendix"). The sample demonstrated the average T score of 50.3 (SD = 9.5), range from 16.31 to 65.11, and median of 50.1. Figure 2 represents the precision of the 35 test items. The x-axis is the converted T score, and the y-axis is the standard error of measurement, indicating the level of precision. The item bank demonstrated the highest precision (SEM = 0.095) at the T score of 39. A wide T score distribution ($20 \leq T$ score ≤ 57.0) was identified with a high reliability of 0.90. The T score values over 62 demonstrated a low reliability of less than 0.80, and 14.6 % ($n = 159$) of the sample fell in this range.

Differential item functioning

The item bank showed no DIF for gender (male/female), race (white/non-white), ethnicity (Hispanic/non-Hispanic), age (under 65/over 65 years), data collection mode (clinical setting/internet survey), and neurological chronic conditions. All magnitudes of McFadden Pseudo R^2 for the DIFs were less than 10 % – gender (0.01–3.21 %), race (0.04–2.37 %), ethnicity (0.01–5.32 %), age (0.00–0.90 %), data collection mode (0.01–4.51 %)–and the five chronic conditions: neuropathy (0.00–3.30 %), stroke (0.04–1.41 %), multiple sclerosis (0.01–1.44 %), Parkinson's disease (0.02–1.77 %) and epilepsy (0.00–3.30 %).

Table 2 Descriptive statistics for PROMIS self-efficacy for managing daily activities

ID	Item	N	Mean (SD)	Median	Range	Ceiling (%)	Item-total correlation
1	I can take a bath or shower	1087	4.58 (0.88)	5	1–5	819 (75.3)	0.684
2	I can eat without help from anyone	1087	4.76 (0.68)	5	1–5	935 (86.0)	0.591
3	I can take care of my personal hygiene without help from anyone (for example: brush my teeth, comb my hair, shave, apply makeup)	1087	4.73 (0.74)	5	1–5	915 (84.2)	0.627
4	I can dress myself in the way I want to be dressed (including buttoning clothes and putting on shoes)	1087	4.58 (0.87)	5	1–5	825 (75.9)	0.666
5	I can get in and out of bed without falling	1086	4.54 (0.90)	5	1–5	796 (73.3)	0.719
6	I can get in and out of a chair	1087	4.57 (0.84)	5	1–5	802 (73.8)	0.723
7	I can get to the bathroom in time	1087	4.41 (0.99)	5	1–5	735 (67.6)	0.662
8	I can manage my clothes when I need to use the toilet.	1085	4.64 (0.79)	5	1–5	843 (77.6)	0.690
9	I can stand for 5 min (for example, waiting in a line, waiting for a bus)	1085	4.33 (1.17)	5	1–5	747 (68.7)	0.756
10	I can walk around inside my house	1085	4.55 (0.94)	5	1–5	817 (75.2)	0.750
11	I can walk a block (about 300 feet or 100 meters) on flat ground	1083	4.22 (1.28)	5	1–5	713 (65.6)	0.789
12	I can exercise at a moderate level for 10 min (for example, walking briskly, biking, swimming, aerobics)	1083	3.95 (1.40)	5	1–5	601 (55.3)	0.761
13	I can exercise at a vigorous level for 10 min (for example, running, jogging)	1085	3.13 (1.62)	3	1–5	360 (33.1)	0.636
14	I can get around in an unfamiliar environment	1085	3.96 (1.24)	4	1–5	527 (48.6)	0.795
15	I can travel to a new destination alone	1083	3.92 (1.38)	5	1–5	567 (52.4)	0.776
16	I can go outside in challenging weather for me.	1085	3.96 (1.32)	5	1–5	564 (51.9)	0.796
17	I can climb one flight of stairs (with or without rails)	1085	4.03 (1.36)	5	1–5	629 (58.0)	0.768
18	I can go shopping and run errands	1086	4.21 (1.21)	5	1–5	670 (61.7)	0.839
19	I can perform my daily activities even if someone is rushing me	1086	4.01 (1.24)	5	1–5	555 (51.1)	0.832
20	I can lift and carry groceries	1087	4.02 (1.33)	5	1–5	616 (56.7)	0.821
21	I can perform my household chores	1085	4.10 (1.26)	5	1–5	631 (58.2)	0.847
22	I can drive a car	1085	3.96 (1.51)	5	1–5	652 (60.1)	0.623
23	I can use public transportation	1083	4.00 (1.41)	5	1–5	626 (57.8)	0.738
24	I can keep doing my usual activities at work	1076	3.76 (1.57)	5	1–5	565 (52.5)	0.762
25	I can use a computer (for example, use keyboard, see screen, login)	1084	4.35 (1.15)	5	1–5	744 (68.6)	0.574
26	I can use a telephone to schedule an appointment	1085	4.68 (0.79)	5	1–5	888 (81.8)	0.624
27	I can engage in hobbies or recreational activities	1085	4.19 (1.20)	5	1–5	662 (61.0)	0.785
28	I can take care of others (for example, cook for others, help them dress, watch children)	1082	3.94 (1.37)	5	1–5	572 (52.9)	0.836
29	I can maintain my finances (for example, write checks, pay bills)	1082	4.39 (1.06)	5	1–5	729 (67.4)	0.657
30	I can maintain a regular exercise program	1081	3.73 (1.36)	4	1–5	461 (42.6)	0.693
31	I can concentrate on something difficult	1082	3.98 (1.20)	4	1–5	508 (47.0)	0.703
32	I can prepare my own meals (for example, plan and cook full meals by myself)	1082	4.27 (1.16)	5	1–5	695 (64.2)	0.759
33	I can take my medications in the right doses and at the right times	1081	4.54 (0.91)	5	1–5	793 (73.4)	0.593
34	I can find new ways to manage daily activities when the old way doesn't work	1082	4.29 (0.99)	5	1–5	610 (56.4)	0.723
35	I can recognize risks (for example, handling hot liquids, walking on uneven ground) and take steps to prevent accidents	1082	4.45 (0.96)	5	1–5	741 (68.5)	0.662

Fig. 1 Distribution of the sum of rating scale frequency
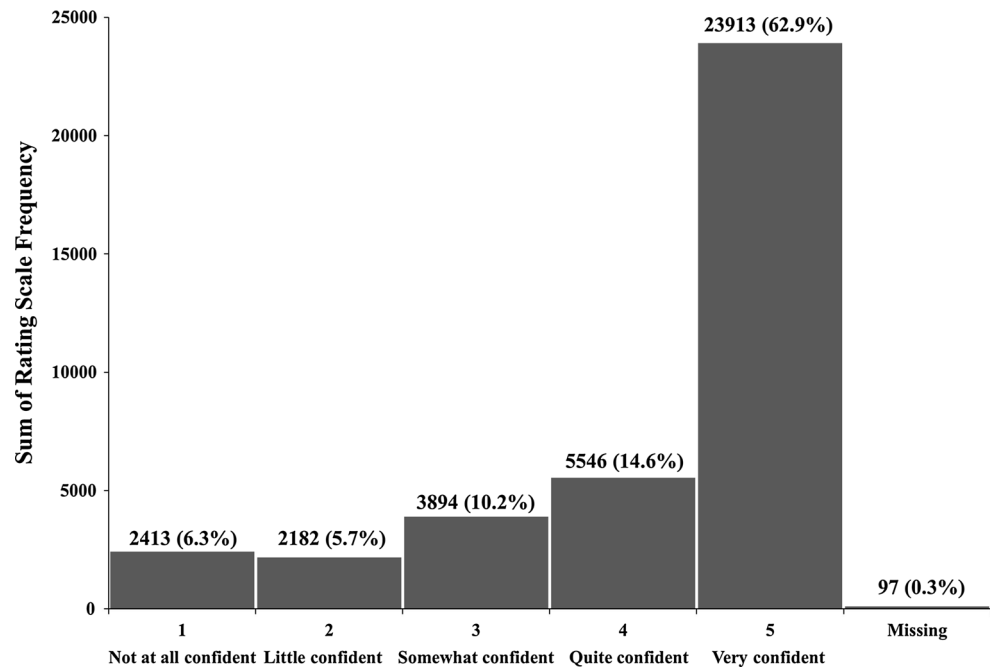
Table 3 EFA eigenvalue of factors on the item bank

Factor	Eigenvalue
1	24.34
2	1.97
3	1.39
4	0.75
5	0.58
6	0.53
7	0.45
8	0.44

Final item bank

Item 13—"I can exercise at a vigorous level for 10 min, i.e., running, jogging" showed a high residual correlation (over 0.2) with item 2, 3, 26, and 34 (−0.22 to −0.29). There was a high residual correlation between item 13 and items 2, 3, 26, and 34; however, item 13 was identified as the most challenging item in the bank. Thus, the item was considered important to measure a wide range of the latent trait. For this reason, item 13 was not removed from the item bank. The findings from the investigation of item calibrations and step thresholds of the 35 test items are reported in Table 4.

Discussion

We developed the PROMIS self-efficacy for managing daily activities item bank and evaluated its psychometrics. The item bank demonstrated a single measurement

construct and precisely measured individuals with a wide range of abilities. Based on its acceptable psychometric properties, the item bank can be used for a computerized adaptive test (CAT) which measures self-efficacy for managing daily activities for individuals among neurological chronic conditions. In addition, the item bank can be used to create various short forms where necessary to reduce administrative burdens.

In keeping with the purpose of the PROMIS initiative [1], this study established a new domain of self-efficacy as part of the existing PROMIS Domain Framework. Self-efficacy is the last domain in the PROMIS mental health category. The item bank for self-efficacy demonstrated sufficient psychometric properties compared to the rest of PROMIS additional domains, such as anxiety, depression, pain, fatigue and anger [16, 29–31]. Once these item banks were created and validated, they were further used for CAT and creating various length of short forms [16, 29–31]. Since the study item bank demonstrated acceptable psychometric properties, various short forms (i.e., 4 and 8 items) of self-efficacy for managing daily activities can be created while maintaining good psychometric properties. In addition, the item bank can be used to develop a CAT to reduce administrative burdens while maintaining measurement precision.

We expected a slightly higher RMSEA value than the common RMSEA criterion because there was a large number of test items in the item pool [30]. However, since TLI marginally met the model fit criteria, we conducted EFA to investigate a potential secondary dimension. The

Table 4 Managing daily activities item bank: item calibrations, step thresholds, factor loadings, and item fit statistics

ID	Item stem	a	$b1$	$b2$	$b3$	$b4$	Average thresholds	Loadings	Item fit statistics		
									$S - X^2$	P	df
1	Take a bath or shower	3.08	−2.18	−1.75	−1.33	−0.79	−1.51	0.83	157.31	0.0823	134
2	Eat without help	3.03	−2.61	−2.03	−1.61	−1.20	−1.86	0.87	146.43	0.0002	90
3	Personal hygiene without help	3.23	−2.27	−2.00	−1.49	−1.10	−1.72	0.88	143.25	0.0005	92
4	Dress myself as desired	2.77	−2.43	−1.84	−1.31	−0.82	−1.60	0.82	174.30	0.0338	142
5	Get in and out of bed without falling	3.26	−2.11	−1.75	−1.21	−0.71	−1.45	0.87	168.74	0.0146	131
6	Get in and out of a chair	3.27	−2.32	−1.81	−1.28	−0.72	−1.53	0.88	180.46	0.0046	134
7	Get to the bathroom in time	2.46	−2.38	−1.73	−1.07	−0.57	−1.44	0.78	204.82	0.0313	169
8	Manage clothes when toileting	3.43	−2.21	−1.91	−1.40	−0.86	−1.60	0.87	153.12	0.0085	114
9	Stand for 5 min	3.57	−1.66	−1.33	−0.96	−0.60	−1.14	0.87	220.57	0.0018	163
10	Walk around inside my house	4.41	−1.86	−1.52	−1.20	−0.76	−1.34	0.89	155.26	0.0122	118
11	Walk a block on flat ground	3.92	−1.44	−1.21	−0.87	−0.50	−1.01	0.91	218.72	0.0028	164
12	Exercise moderately for 10 min	2.94	−1.36	−1.05	−0.61	−0.20	−0.81	0.88	245.94	0.0450	210
13	Exercise vigorously for 10 min	1.95	−0.83	−0.39	0.12	0.59	−0.13	0.78	253.35	0.0218	210
14	Get around in an unfamiliar environment	3.07	−1.69	−1.14	−0.58	0.02	−0.85	0.87	229.13	0.0427	194
15	Travel to a new destination alone	2.99	−1.40	−1.03	−0.60	−0.10	−0.78	0.87	242.93	0.0182	199
16	Go outside in challenging weather	3.39	−1.47	−1.07	−0.59	−0.09	−0.81	0.87	229.22	0.0342	192
17	Climb one flight of stairs	3.19	−1.43	−1.08	−0.67	−0.27	−0.86	0.86	273.99	0.0002	195
18	Go shopping and run errands	4.75	−1.49	−1.21	−0.83	−0.37	−0.98	0.91	167.40	0.1194	147
19	Perform daily activities when rushed	3.95	−1.55	−1.14	−0.61	−0.05	−0.84	0.89	207.71	0.0066	160
20	Lift or carry groceries	4.06	−1.41	−1.06	−0.63	−0.23	−0.83	0.90	212.25	0.0175	171
21	Perform my household chores	4.79	−1.49	−1.07	−0.70	−0.26	−0.88	0.92	208.13	0.0020	153
22	Drive a car	2.10	−1.32	−1.11	−0.87	−0.40	−0.93	0.76	323.72	0.0001	200
23	Use public transportation	3.05	−1.33	−1.04	−0.74	−0.28	−0.85	0.84	303.19	0.0001	197
24	Keep doing my usual activities at work	3.06	−1.04	−0.84	−0.53	−0.11	−0.63	0.86	214.54	0.1486	194
25	Use a computer	1.98	−2.04	−1.69	−1.23	−0.67	−1.41	0.72	277.16	0.0001	185
26	Use telephone to schedule appointments	2.86	−2.37	−1.93	−1.50	−1.04	−1.71	0.82	184.63	0.0001	106
27	Engage in hobbies or recreational activities	3.51	−1.72	−1.24	−0.80	−0.35	−1.03	0.86	227.97	0.0020	170
28	Can take care of others	4.03	−1.33	−0.99	−0.59	−0.11	−0.76	0.90	221.20	0.0103	175
29	Maintain my finances	2.24	−2.21	−1.67	−1.19	−0.57	−1.41	0.77	222.70	0.0271	184
30	Maintain a regular exercise program	2.27	−1.60	−0.99	−0.38	0.23	−0.69	0.78	238.18	0.1231	214
31	Concentrate on something difficult	2.26	−1.98	−1.31	−0.64	0.09	−0.96	0.78	275.47	0.0044	217
32	Prepare my own meals	3.09	−1.73	−1.39	−0.91	−0.45	−1.12	0.84	180.69	0.4510	179
33	Take medications with correct dose and times	1.90	−2.61	−2.10	−1.53	−0.80	−1.76	0.74	216.44	0.0032	163
34	Can find new ways to manage daily activities when the old way doesn't work	2.62	−2.28	−1.71	−1.01	−0.20	−1.30	0.81	243.13	0.0003	173
35	Recognize risks and prevent accidents	2.53	−2.28	−1.76	−1.17	−0.59	−1.45	0.79	245.93	0.0001	165

Item parameter estimates = $a(\theta - b)$; a = discrimination (slope); $b1$ = threshold 1; $b2$ = threshold 2; $b3$ = threshold 3; $b4$ = threshold 4; Loadings = the EFA factor loadings on the test items

EFA indicated that the item pool sufficiently met the unidimensional factor criteria.

Only one of 35 test items, item 13—"exercise vigorously for 10 min"—showed local dependency with items 2, 3, 24, and 34 (residual correlations = −0.22 to −0.29). High residual correlations estimated by the single factor CFA indicate that the relationships between item 13 and these four items are stronger than the relationships between item 13 and the latent construct. However, the relationships between item 13 and the four items are not interpretable. For example, the residual correlation between item 13 and item 2 was −0.29, indicating that patients having high confidence in exercising vigorously have low confidence in eating. Analyses conducted in this study do not speak to

potential explanations for such findings. Item 13 was the most challenging item in the item pool and therefore effective in measuring individuals who respond at the higher ranges of the measure. For these reasons, we did not remove item 13 from the item bank.

The distributions of responses were skewed to endorsing high confidence levels, indicating that the test items in the bank were relatively less challenging than the confidence levels of our sample. In other words, the sample's confidence levels were higher than the difficulty of the test items. This finding may explain the low precision for participants (14.6 % of the sample) with high confidence levels, those with T score greater than 62 (lower than reliability 0.80). Theoretically, the IRT model calibrates parameters based on relationships between person ability and item difficulty. When person ability is equivalent to item difficulty, the item information function carries the most information and low standard errors of measurement [32]. However, when there is a discrepancy between person ability and item difficulty, there is a high probability of errors that result in low reliability. In the item bank, the most challenging test items were item 13—"exercise vigorously for 10 min," item 24—"keep doing my usual activities at work," and item 30—"maintain a regular exercise program." Although these test items were the most challenging items in the item bank, other items may not be challenging enough to measure the self-efficacy for individuals with higher confidence. In the item pool development process, the research team made efforts to develop particularly challenging items, such as items 12, 13, 14, 15, 16, 19, 20, 23, 24, 28, 30, 31, 34, and 35. Those items were identified as challenging items and were modified based on feedback from patients through focus groups and cognitive testing. For instance, item 28—"I can take care of others (for example, cook for others, help them dress, watch children)" which demonstrated lower proportions of ratings at the ceiling (about 50 %). However, most items demonstrating the highest proportions of ratings at the ceiling are related to basic and instrument activities of daily living. A possible explanation of the ceiling effect would be that it occurred due to characteristics among participants who reside in their community, a population in which 67.3 % of have no-to-mild impairment levels. Their medical conditions may be stabilized and good enough for living in the community. Therefore, additional research may be needed to develop more challenging test items in order to measure precisely self-efficacy for managing more challenging daily activities.

Although the conceptual domains of self-efficacy were defined and modified by qualitative methods (Delphi, focus groups, and cognitive interviews), the self-efficacy for managing daily activities item bank have similar test item content compared to the PROMIS physical functioning item bank (v. 1.2). In spite of this similarity, we pursued this self-efficacy measure because from the Delphi phase, being physically active was identified as an important aspect of self-efficacy for managing chronic disease. For instance, item 11, "I can walk a block (about 300 feet or 100 m) on flat ground" with the response options ranging from "I am not at all confident" to "I am very confident" is very similar to a test item of the PROMIS physical functioning item bank (v. 1.2), "Are you able to walk a block (100 m) on flat ground?" with the primary difference being in the response options ranging from "Without any difficulty" to "Unable to do" [33]. We believe this is a generic limitation of self-reported outcome measures in measuring different conceptual domains. To clearly distinguish the differences and similarities of the self-efficacy item bank with other conceptual constructs, future studies will be needed to investigate the convergent and divergent validity between the self-efficacy item bank and other self-efficacy measures and non-self-efficacy measures, respectively.

The item bank demonstrated no evidence of DIF across gender, race, ethnicity, age, data collection site, and the five neurological chronic conditions. However, sample size could affect the significant level of DIF when there is a small sample in a comparison group and a reference group or both. The managing daily activities item bank consisted of a 5-point rating scale, indicating the need for a sample size of at least 200 to detect moderate uniform DIF [34]. Although there was no DIF for demographics and the five neurological chronic conditions, the sample sizes for ethnicity ($n = 64$) and four chronic conditions, including stroke, multiple sclerosis, Parkinson's disease, and epilepsy ($n = 169$–181), were less than 200. These small sample sizes in the comparison group might inflate the type II error rate. Thus, further studies are recommended to replicate with larger samples ($n > 200$) to test DIF for ethnicity and the four chronic conditions.

This study has several limitations. First, our sample had a predominance of neurological conditions. While there was no DIF for the data collection mode (clinical site and internet survey), there were significant differences in demographics and the neurological chronic conditions. The different recruitment methods may cause a selection bias, and the sample differences might affect psychometric properties. Secondly, while clinicians rated the item pool and other information for the clinical sample, the collected data was self-reported from the internet sample. Since data administration methods were different for the two samples, the quality of data may not be consistent.

In conclusion, the PROMIS self-efficacy for managing daily activities item bank was developed and tested by quantitative measurement theory methods. The psychometrics of the item bank indicated that self-efficacy for managing chronic conditions can be reliably measured,

Fig. 2 Standard error of measurement of the managing daily activities item bank
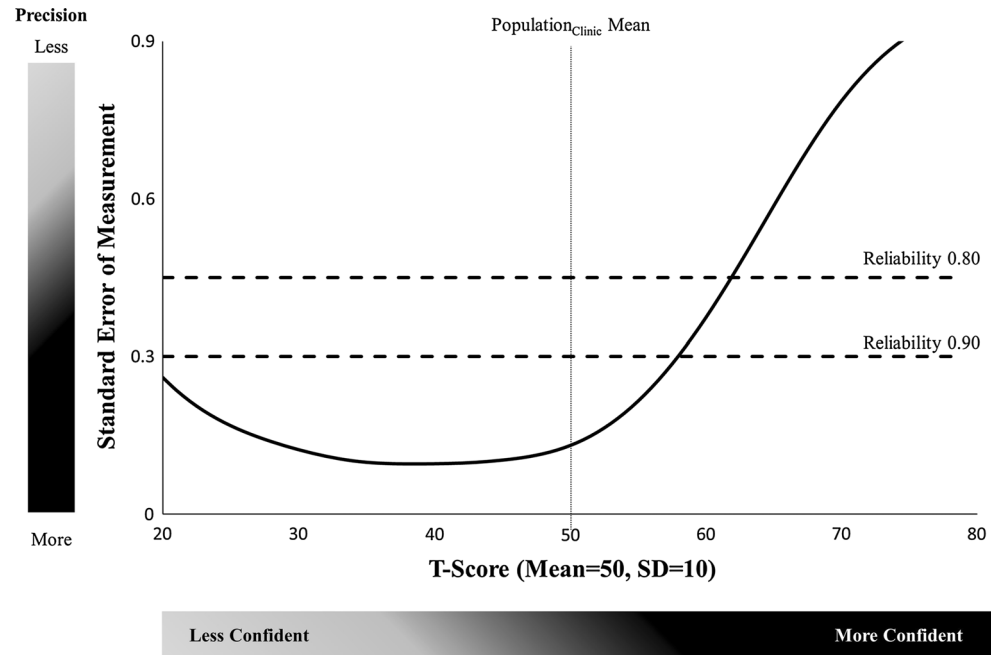
maintaining an acceptable precision across a wide range of confidence levels. Based on the sound psychometrics, a CAT for the item bank and various short forms (i.e., 4 and 8 items) can be developed in order to reduce the burden of completing the 35-item bank. Once a CAT is developed and tested in clinical settings in people with a range of self-efficacy, the item bank may be improved by tracking the trajectory of change in self-efficacy in a longitudinal study and by introducing more challenging test items that result in improved precision when measuring high levels of self-efficacy. Further research is needed to investigate the untested psychometrics of the item bank, such as predictive validity, test–retest reliability, and responsiveness.

Appendix

Raw scores were converted into a *T* score metric

Raw score	*T* score	SE	Raw score	*T* score	SE	Raw score	*T* score	SE
35	16.31	3.73	82	35.05	0.87	129	42.63	0.87
36	18.98	2.91	83	35.17	0.91	130	42.75	0.93
37	20.28	2.70	84	35.30	0.97	131	42.89	1.01
38	21.32	2.52	85	35.45	1.06	132	43.06	1.10
39	22.18	2.37	86	35.63	1.14	133	43.25	1.19
40	22.97	2.21	87	35.84	1.22	134	43.48	1.25
41	23.65	2.08	88	36.07	1.26	135	43.72	1.28
42	24.26	1.99	89	36.31	1.27	136	43.97	1.28
43	24.82	1.90	90	36.55	1.24	137	44.21	1.23
44	25.33	1.84	91	36.76	1.17	138	44.43	1.17
45	25.81	1.77	92	36.95	1.08	139	44.63	1.09
46	26.26	1.71	93	37.11	0.98	140	44.80	1.03
47	26.67	1.65	94	37.24	0.89	141	44.95	0.99
48	27.05	1.60	95	37.35	0.82	142	45.10	0.99
49	27.42	1.57	96	37.44	0.78	143	45.25	1.03
50	27.77	1.55	97	37.53	0.78	144	45.42	1.10
51	28.12	1.53	98	37.62	0.81	145	45.62	1.18
52	28.46	1.51	99	37.72	0.87	146	45.84	1.26

Raw score	T score	SE	Raw score	T score	SE	Raw score	T score	SE
53	28.78	1.48	100	37.85	0.95	147	46.09	1.31
54	29.09	1.43	101	38.00	1.05	148	46.36	1.34
55	29.37	1.37	102	38.17	1.14	149	46.64	1.33
56	29.64	1.32	103	38.38	1.22	150	46.90	1.29
57	29.88	1.30	104	38.61	1.26	151	47.16	1.25
58	30.13	1.30	105	38.85	1.26	152	47.39	1.22
59	30.38	1.32	106	39.08	1.23	153	47.62	1.22
60	30.64	1.34	107	39.29	1.16	154	47.85	1.25
61	30.91	1.36	108	39.47	1.07	155	48.10	1.31
62	31.18	1.36	109	39.63	0.97	156	48.37	1.38
63	31.45	1.33	110	39.76	0.89	157	48.68	1.44
64	31.70	1.27	111	39.86	0.82	158	49.01	1.49
65	31.92	1.20	112	39.96	0.78	159	49.36	1.52
66	32.11	1.13	113	40.04	0.78	160	49.72	1.53
67	32.29	1.07	114	40.14	0.81	161	50.08	1.55
68	32.45	1.05	115	40.24	0.88	162	50.46	1.60
69	32.62	1.06	116	40.37	0.97	163	50.85	1.67
70	32.79	1.10	117	40.52	1.06	164	51.29	1.76
71	32.98	1.17	118	40.70	1.15	165	51.77	1.87
72	33.20	1.23	119	40.91	1.22	166	52.31	1.98
73	33.43	1.28	120	41.14	1.26	167	52.91	2.11
74	33.68	1.29	121	41.38	1.27	168	53.56	2.25
75	33.92	1.28	122	41.61	1.23	169	54.31	2.44
76	34.16	1.22	123	41.83	1.16	170	55.19	2.67
77	34.37	1.15	124	42.01	1.07	171	56.38	3.21
78	34.54	1.06	125	42.17	0.98	172	57.28	3.26
79	34.70	0.97	126	42.30	0.91	173	58.68	3.56
80	34.82	0.90	127	42.42	0.86	174	60.39	3.82
81	34.94	0.87	128	42.53	0.85	175	65.11	5.41

The raw scores were converted into T scores with average 50 and standard deviation 10

SE standard error

References

1. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–S11. doi:10.1097/01.mlr.0000258615.42478.55.

2. Bodenheimer, T., Lorig, K., Holman, H., & Grumbach, K. (2002). Patient self-management of chronic disease in primary care. *Journal of the American Medical Association, 288*(19), 2469–2475.

3. Leventhal, H., Weinman, J., Leventhal, E. A., & Phillips, L. A. (2008). Health psychology: The search for pathways between behavior and health. *Annual Review of Psychology, 59*, 477–505. doi:10.1146/annurev.psych.59.103006.093643.

4. Jones, F., Partridge, C., & Reid, F. (2008). The Stroke Self-Efficacy Questionnaire: Measuring individual confidence in functional performance after stroke. *Journal of Clinical Nursing, 17*(7b), 244–252. doi:10.1111/j.1365-2702.2008.02333.x.

5. Gramstad, A., Iversen, E., & Engelsen, B. A. (2001). The impact of affectivity dispositions, self-efficacy and locus of control on psychosocial adjustment in patients with epilepsy. *Epilepsy Research, 46*(1), 53–61.

6. Schwartz, C. E., Coulthard-Morris, L., Zeng, Q., & Retzlaff, P. (1996). Measuring self-efficacy in people with multiple sclerosis: A validation study. *Archives of Physical Medicine and Rehabilitation, 77*(4), 394–398.

7. Beckham, J. C., Burker, E. J., Lytle, B. L., Feldman, M. E., & Costakis, M. J. (1997). Self-efficacy and adjustment in cancer patients: A preliminary report. *Behavioral Medicine, 23*(3), 138–142. doi:10.1080/08964289709596370.

8. Wilcox, S., Schoffman, D. E., Dowda, M., & Sharpe, P. A. (2014). Psychometric properties of the 8-item english arthritis self-efficacy scale in a diverse sample. *Arthritis, 2014*, 385256. doi:10.1155/2014/385256.

9. Edwards, R., Telfair, J., Cecil, H., & Lenoci, J. (2001). Self-efficacy as a predictor of adult adjustment to sickle cell disease: One-year outcomes. *Psychosomatic Medicine, 63*(5), 850–858.

10. Lorig, K., Chastain, R. L., Ung, E., Shoor, S., & Holman, H. R. (1989). Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. *Arthritis and Rheumatism, 32*(1), 37–44.

11. Lorig, K. R., Sobel, D. S., Ritter, P. L., Laurent, D., & Hobbs, M. (2001). Effect of a self-management program on patients with chronic disease. *Effective Clinical Practice, 4*(6), 256–262.

12. 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*(5 Suppl 1), S22–S31. doi:10.1097/01.mlr.0000250483.85507.04.

13. van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

14. DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(5 Suppl 1), S12–S21. doi:10.1097/01.mlr.0000254567.79743.e2.

15. 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*(11), 1179–1194. doi:10.1016/j.jclinepi.2010.04.011.

16. Revicki, D. A., Chen, W. H., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., et al. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain, 146*(1–2), 158–169. doi:10.1016/j.pain.2009.07.029.

17. IBM SPSS Statistics for Windows. (2013). (22.0 ed.). Armonk, NY: IBM Corp.

18. Hair, J. F., Anderson, R. E., Tatham, R. I., & Black, W. C. (1998). *Multivariate data analysis with readings*. Prentice Hall: NY.

19. Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. doi:10.1007/BF02310555.

20. McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 4*(4), 293–307.

21. Muthén, L. K., & Muthén, B. O. (2012). *Mplus* (7.11th ed.). CA: Los Angeles.

22. Revicki, D. A., Cook, K. F., Amtmann, D., Harnam, N., Chen, W. H., & Keefe, F. J. (2014). Exploratory and confirmatory factor analysis of the PROMIS pain quality item bank. *Quality of Life Research, 23*(1), 245–255. doi:10.1007/s11136-013-0467-9.

23. Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* (2.1st ed.). Lincolnwood, IL: Scientific Software International.

24. Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.

25. Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - \chi^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298. doi:10.1177/0146621603027004004.

26. Dodd, B., Koch, W., & De Ayala, R. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*(2), 129–143.

27. 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.

28. Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241–256. doi:10.1002/sim.1713.

29. Irwin, D. E., Stucky, B., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J.-S., et al. (2010). An item response analysis of the pediatric PROMIS anxiety and depressive symptoms scales. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 19*(4), 595–607. doi:10.1007/s11136-010-9619-3.

30. Lai, J.-S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation, 92*(10), S20–S27.

31. Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment, 18*(3), 263–283. doi:10.1177/1073191111411667.

32. Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9 Suppl), 28–42.

33. Rose, M., Bjorner, J. B., Becker, J., Fries, J. F., & Ware, J. E. (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology, 61*(1), 17–33. doi:10.1016/j.jclinepi.2006.06.025.

34. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62*(3), 288–295. doi:10.1016/j.jclinepi.2008.06.003.