Qual Life Res
DOI 10.1007/s11136-012-0292-6

Evaluation of the Patient-Reported Outcomes Information System (PROMIS®) Spanish-language physical functioning items

Sylvia H. Paz · Karen L. Spritzer · Leo S. Morales · Ron D. Hays

Accepted: 4 October 2012

Abstract
Purpose To evaluate the equivalence of the PROMIS® physical functioning item bank by language of administration (English versus Spanish).
Methods The PROMIS® wave 1 English-language physical functioning bank consists of 124 items, and 114 of these were translated into Spanish.
Analysis Item frequencies, means and standard deviations, item-scale correlations, and internal consistency reliability were calculated. The IRT assumption of unidimensionality was evaluated by fitting a single-factor confirmatory factor analytic model. IRT threshold and discrimination parameters were estimated using Samejima's Graded Response Model. DIF by language of administration was evaluated.
Results Item means ranged from 2.53 (SD = 1.36) to 4.62 (SD = 0.82). Coefficient alpha was 0.99, and item-rest correlations ranged from 0.41 to 0.89. A one-factor model fits the data well (CFI = 0.971, TLI = 0.970, and RMSEA = 0.052). The slope parameters ranged from 0.45 ("Are you able to run 10 miles?") to 4.50 ("Are you able to put on a shirt or blouse?"). The threshold parameters ranged from −1.92 ("How much do physical health problems now limit your usual physical activities (such as walking or climbing stairs)?") to 6.06 ("Are you able to run 10 miles?"). Fifty of the 114 items were flagged for DIF based on an R^2 of 0.02 or above criterion. The expected total score was higher for Spanish- than English-language respondents.
Conclusions English- and Spanish-speaking subjects with the same level of underlying physical function responded differently to 50 of 114 items. This study has important implications in the study of physical functioning among diverse populations.

Keywords PROMIS® item banks · IRT · Physical function Spanish items

Electronic supplementary material The online version of this article (doi:10.1007/s11136-012-0292-6) contains supplementary material, which is available to authorized users.

S. H. Paz (✉) · K. L. Spritzer · R. D. Hays
Division of General Internal Medicine and Health Services Research, Department of Medicine, UCLA School of Medicine, 911 Broxton Avenue, Los Angeles, CA 90095-1736, USA
e-mail: shpaz@ucla.edu

L. S. Morales
Group Health Research Institute, Group Health Cooperative, Seattle, WA, USA

L. S. Morales
Department of Health Services, University of Washington, Seattle, WA, USA

R. D. Hays
RAND, 1776 Main Street, Santa Monica, CA 90407, USA

Purpose

Latinos are individuals born into or descended from a Spanish-speaking community [1, 2]. Latinos as a racial/ethnic group stand apart from other racial/ethnic groups in terms of their population growth, socioeconomic characteristics, and health-related issues [3–5]. The US Bureau of Census projects that the proportion of Latinos in the general United States population will increase dramatically with immigration and greater longevity, higher birth rates, and lower infant mortality rates. In fact, the Latino population in the United States was 5 % of the total population in 1970, 12 % in 2000, and 16 % in 2010 [6–8]. The census office

Published online: 03 November 2012

estimates that by 2050, 30 % of the United States population will be Latino [8–10]. A larger percentage of the elder population will consist of minorities, especially Latinos, the fastest-growing minority group among older adults [11].

Physical functioning is an especially important indicator of health for older individuals and one of the strongest predictors of health care utilization and mortality [12–14]. A physical functioning item bank was created for the Patient-Reported Outcomes Measurement Information System (PROMIS®) project [14]. Because interpretation of results requires equivalent responses for different subgroups, differential item functioning (DIF) was previously evaluated for age, gender, and educational attainment for English-language respondents [15]. This paper extends that work by evaluating DIF for the PROMIS® Spanish- and English-language physical functioning items.

Methods

Data sources and measures

English language

The PROMIS® wave 1 English-language physical functioning bank consists of 124 items that assess mobility (lower extremity), dexterity (upper extremity), axial or central (neck and back function), and complex activities that overlap more than one domain (daily living activities) [12–14]. Because of the large number of physical functioning items, two sets of 56 items (112 items) were administered to a randomly selected subgroup of English-language individuals and one set of 56 items to another subgroup (another PROMIS® item bank of 56 items was also administered to this subgroup). The analysis sample consisted of the 728 and 776 individuals in the first and second subgroups that completed at least half of the physical functioning items they were administered ($n = 1{,}504$ overall).

Respondents were recruited by YouGovPolimetrix, a polling firm based in Palo Alto, CA [16]. This firm uses a sample-matching procedure to select a representative sample of the population [17]. The PROMIS® online panel was found to have similar demographic characteristics as the United States census, except that the online panel tended to have more educated individuals [18].

Spanish language

The PROMIS® physical functioning item bank was translated into Spanish using a universal approach for translations and cultural adaptation of instruments [19–22]. The process involved 2 initial forward translations, 1 reconciled version, 1 back translation by a native English speaker, comparison of original with back translation, and reviews by 3 bilingual experts from different Spanish-speaking countries. Fifteen cognitive interviews with native Spanish speakers followed to evaluate the comprehension of the items. The items were divided into three groups, and each group of items was evaluated by five subjects.

Only 114 of the 124 items were administered to the Spanish-language sample. The 10 items that were excluded were those for which the spread of responses was restricted in the English sample, requiring collapsing response categories for analyses. The Spanish-translated items were administered to 640 adult Spanish-speaking Latinos in the Toluna online panel, an independent survey technology provider [23]. All 640 respondents answered all of the 114 physical functioning items administered to the Spanish sample.

Statistical analysis

We estimated item frequencies, means and standard deviations for the Spanish-language physical functioning items. Unidimensionality of the items was examined by fitting a one-factor categorical confirmatory factor analysis model using Mplus [24]. Model fit was assessed by the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI), as well as the root mean square error of approximation (RMSEA). Good model fit is defined by the following cutoffs: CFI >0.95, TLI >0.95, and RMSEA <0.06 [25, 26]. Next, IRT assumptions of local independence and monotonicity were evaluated. Local independence was assessed based on residual correlations among items.

Multilog was used to estimate IRT item parameters (slope and threshold) for Samejima's Graded Response Model [27]. This model yields one slope or discrimination parameter and $(n - 1)$ threshold parameters for polytomous items with n response options. The slope parameter gives information regarding the discrimination of the item between adjoining trait levels. Higher values indicate that items are better able to discriminate between adjacent categories of trait level. The threshold parameter represents the point along the latent trait at which a subject has a 50 % chance of responding in that category or higher.

Differential item functioning (DIF) was assessed by comparing the Spanish language ($n = 640$) with the PROMIS® wave 1 English-language data ($n = 1{,}504$). DIF is present if the probability of selecting a particular response varies by language group when controlling for the underlying level of physical functioning. We evaluated DIF using ordinal logistic regression with IRT-based trait scores estimated from DIF-free "anchor" items (iterative purification) as the conditioning variable. A pseudo R^2 difference of <0.02 between nested models was used to identify potential anchor items. We examined the magnitude of DIF

for English versus Spanish language using test characteristic curves separately for all physical functioning items and for the items identified as having DIF. For those items having DIF, we evaluated whether they had uniform DIF, in which DIF is in the same direction across the entire ability continuum of physical function (response curves for both groups do not cross); or non-uniform DIF, in which the probability of endorsing an item is higher for one group at lower physical functioning, and higher for the other group at higher physical functioning (response curves for both groups cross at a certain point along the continuum). We assessed DIF at the individual level by plotting theta estimates ignoring DIF versus theta estimates accounting for DIF. These analyses were run using Lordif software [28].

To evaluate the impact of using English-language calibrations versus Spanish-language calibrations on theta estimates, we put the Spanish-language item parameters (slopes and thresholds) on the same metric as the English-language parameters using STUIRT [29]. This program produces the Stocking-Lord linking constants used to linearly transform the Spanish item parameter estimates to the English metric. Then, we used Firestar to simulate CAT-based theta estimates from the two sets of item parameters and compared the estimates [30]. Options used for these simulations included imposing a minimum number of items of 5, a maximum number of items of 20, and stopping rule based on a standard error of 0.30.

Results

Descriptive statistics

The average age of the English-speaking sample that completed the physical functioning item bank was 51 with a range from 18 to 93. The average age of the Spanish-speaking sample was 38 with a range from 18 to 77. The English sample was relatively older with fifty-three percent ($n = 802$) being over 50-year old while only fourteen percent of the Spanish sample ($n = 87$) was 50 years or older. Fifty-two percent of the English-speaking sample was female while fifty-eight percent of the Spanish-speaking sample was female (see Table 1).

Thirteen percent of the Spanish-speaking sample reported reading and speaking only Spanish, 48 % speaking Spanish better than English, 39 % reading and speaking both languages equally, and only 1 person reported reading and speaking English better than Spanish. Thirty-three percent reported speaking only Spanish at home, 51 % speaking more Spanish than English at home, and 15 % speaking both equally at home.

The Spanish-speaking sample was less educated with thirty-six percent ($n = 233$) reporting High school education/GED or less, while the English-speaking sample reported twenty percent ($n = 293$) with High School education/GED or less. Another noteworthy difference was the prevalence of some comorbidities being significantly higher in the English-speaking sample: high blood pressure (36 vs. 16 %) and depression (21 vs. 14 %). However, when looking only at those subjects with those comorbidities and who reported being limited in activities because of them, the prevalence was significantly higher in the Spanish sample: high blood pressure (8 % in the English sample versus 37 % in the Spanish sample) and depression (29 % in the English sample versus 55 % in the Spanish sample) (see Table 1).

The 114 physical function items (see Online resource 1) administered to the Spanish sample had 5 response options each: 1 = worst physical function (*cannot do* or *unable to do activity*) and 5 = best physical function (*health does not limit at all in doing activity*). The Spanish sample had no missing data and the range for item means (standard deviation) was 2.53 (1.36) to 4.62 (0.82). The frequencies for the items are provided in Online resource 2. Twenty-nine items had categories with less than 5 responses (*unable to do or cannot do*), see Online resource 3. These categories in these items were collapsed with neighbor categories for analysis. For the 114 items, coefficient alpha was 0.99 and item-rest correlations ranged from 0.41 to 0.89.

Dimensionality (Spanish sample)

A one-factor categorical model was found to fit the data well (CFI = 0.971, TLI = 0.970, and RMSEA = 0.052). Factor loadings (see Table 2) ranged from 0.329 for item PFC33 (*Are you able to run 10 miles (16 km)*?) to 0.939 for item PFB12 (*Are you able to make a bed, including spreading and tucking in bed sheets?*).

Item parameters

Marginal maximum likelihood estimates and standard errors of item parameters from the two-parameter graded response model for all the physical functioning items are provided in Table 3. The slope parameters for this model ranged from 0.45 to 4.50, and the threshold parameters ranged from −1.92 to 6.06. PFC33 (*"Are you able to run 10 miles (16 km)?"*) had the lowest slope while PFA44 (*"Are you able to put on a shirt or blouse?"*) had the highest slope. PFA7 (*"How much do physical health problems now limit your usual physical activities (such as walking or climbing stairs)?"*) had the smallest category threshold (between *cannot do* and *quite a lot*). PFC33

Table 1 Sociodemographic and clinical characteristics of Spanish ($n = 640$) and English ($n = 1,504$) physical function sample

	Spanish	English	Comparison
Age: (mean/SD/range)	37.6 (11.3) 18–77	51.1 (18.3) 18–93	$t(1,869) = 20.8$; $p < 0.0001$
SASH score: (mean/SD/range)[a]	2.02 (0.53) 1–2.75		
Age categories: (*n*/%)			
18–24	100 (16)	137 (9)	Chi(4) = 309.9; $p < 0.0001$
25–34	145 (23)	196 (13)	
35–44	226 (35)	222 (15)	
45–49	82 (13)	146 (10)	
50+	87 (14)	802 (53)	
Gender: (*n*/%)			
Male	271 (42)	716 (48)	2-sided $p < 0.0261$
Female	369 (58)	788 (52)	
Race/ethnicity: (*n*/%)			
Hispanic	640 (100)	158 (11)	
Non-Hispanic White	–	1,123 (75)	
Non-Hispanic Black or African–American	–	156 (10)	
Non-Hispanic other race	–	63 (4)	
Education: (*n*/%)			
Less than High School Grad/GED	91 (14)	23 (2)	Chi(4) = 169.4; $p < 0.0001$
HS graduate/GED	142 (22)	270 (18)	
Some college	199 (31)	581 (39)	
College degree	156 (24)	381 (25)	
Advanced degree	52 (8)	247 (16)	
Comorbidities—ever told you have: (*n*/%)			
High blood pressure	101 (16)	547 (36)	$p < 0.0001$
Chest pain (angina)	27 (4)	79 (5)	n.s.
Hardening of the arteries	4 (1)	59 (4)	$p < 0.0001$
Heart failure or congestive heart failure	4 (1)	41 (3)	$p = 0.0019$
Heart attack (myocardial infarction)	8 (1)	56 (4)	$p = 0.0020$
Stroke or transient ischemic attack (TIA)	5 (1)	42 (3)	$p = 0.0036$
Liver disease, hepatitis, or cirrhosis	11 (2)	37 (2)	n.s.
Kidney disease	27 (4)	34 (2)	$p = 0.0128$
Arthritis or rheumatism	28 (4)	335 (22)	$p < 0.0001$
Asthma	67 (10)	208 (14)	$p = 0.0323$
Chronic lung disease (COPD), chronic bronchitis or emphysema	8 (1)	80 (5)	$p < 0.0001$
Migraines or severe headaches	92 (14)	203 (14)	n.s.
Diabetes or high blood sugar or sugar in urine	60 (9)	142 (9)	n.s.
Cancer other than non-melanoma skin cancer	12 (2)	123 (8)	$p < 0.0001$
Depression	88 (14)	318 (21)	$p < 0.0001$
Anxiety	61 (10)	204 (14)	$p = 0.0092$
Alcohol or drug problem	12 (2)	48 (3)	n.s.
Sleep disorder	48 (8)	156 (10)	$p = 0.0373$
HIV or AIDS	2 (0.3)	5 (0.3)	n.s.
Spinal cord injury	5 (1)	28 (2)	n.s.
Multiple sclerosis	3 (0.47)	6 (0.4)	n.s.
None of the above*	251 (39)	427 (28)	$p < 0.0001$
Comorbidities (activities limited by): (*n*/% of ever told)			
High blood pressure	37 (37)	46 (8)	$p < 0.0001$
Chest pain (angina)	16 (59)	14 (18)	$p < 0.0001$

Table 1 continued

	Spanish	English	Comparison
Hardening of the arteries	2 (50)	11 (19)	n.s.
Heart failure or congestive heart failure	1 (25)	13 (32)	n.s.
Heart attack (myocardial infarction)	7 (88)	8 (14)	$p < 0.0001$
Stroke or transient ischemic attack (TIA)	4 (80)	7 (17)	$p = 0.0018$
Liver disease, hepatitis, or cirrhosis	4 (36)	3 (8)	$p = 0.0197$
Kidney disease	13 (48)	4 (12)	$p = 0.0016$
Arthritis or rheumatism	22 (79)	165 (49)	$p = 0.0029$
Asthma	37 (55)	68 (33)	$p = 0.0010$
Chronic lung disease (COPD), chronic bronchitis or emphysema	4 (50)	39 (49)	n.s.
Migraines or severe headaches	57 (62)	72 (36)	$p < 0.0001$
Diabetes or high blood sugar or sugar in urine	25 (42)	28 (20)	$p = 0.0012$
Cancer other than non-melanoma skin cancer	11 (92)	11 (9)	$p < 0.0001$
Depression	48 (55)	93 (29)	$p < 0.0001$
Anxiety	33 (54)	68 (33)	$p = 0.0034$
Alcohol or drug problem	9 (75)	7 (15)	$p < 0.0001$
Sleep disorder	34 (71)	63 (40)	$p = 0.0002$
HIV or AIDS	1 (50)	2 (40)	n.s.
Spinal cord injury	5 (100)	19 (68)	n.s.
Multiple sclerosis	2 (67)	5 (83)	n.s.

[a] SASH Score: Short Acculturation Scale for Hispanics (SASH); the rating scale ranges from 1 ("Only Spanish") to 5 ("Only English") and an average score <3.0 reflects low acculturation

(*"Are you able to run 10 miles (16 km)?"*) had the largest category threshold (between *without any difficulty* and *with a little difficulty*).

Effect of item local dependency

Lagrange multiplier tests indicated that there was a very large residual correlation of 0.672 between PFC33 (*Are you able to run 10 miles (16 km)?*) and PFC7 (*Are you able to run 5 miles (8 km)?*). In fact, 11 items had a residual correlation of 0.20 or more: PFA13 (*Are you able to exercise for an hour?*), PFA19 (*Are you able to run or jog for 2 miles (3 km)?*), PFA33 (*Are you able to exercise hard for half an hour?*), PFA39 (*Are you able to run at a fast pace for 2 miles (3 km)?*), PFB5 (*Does your health now limit you in hiking a couple of miles (3 km) on uneven surfaces, including hills?*), PFB7 (*Does your health now limit you in doing strenuous activities such as backpacking, skiing, playing tennis, bicycling or jogging?*), PFB51 (*Does your health now limit you in participating in active sports such as swimming, tennis, or basketball?*), PFC7 (*Are you able to run 5 miles (8 km)?*), PFC13 (*Are you able to run 100 yards (100 m)?*), PFC33 (*Are you able to run 10 miles (16 km)?*), and PFC35 (*Does your health now limit you in doing 8 h of physical labor?*).

An iterative process was followed evaluating model fit and parameter changes after dropping one locally dependent item at a time. After dropping all 11 locally dependent items, model fit indices improved to CFI = 0.985, TLI = 0.984, and RMSEA = 0.041. Marginal maximum likelihood estimates of item parameters from a two-parameter graded response model after dropping the 11 local-dependent items revealed that the slope parameters ranged from 1.83 to 4.57 while the threshold parameters ranged from −1.86 to 1.70. Because of some difference in model fit, we evaluated DIF for the full set of items and then after dropping the locally dependent items (see below).

Identification of DIF

Fifty of the 114 items were flagged for DIF based on the R^2 of 2 % (0.0200) or above criterion; 20 uniform and 30 non-uniform (see Table 3). Examination of DIF restricted to 103 items (dropping the 11 items with high residual correlations) showed 44 items with DIF based on the R^2 of 2 % (0.0200) or above criterion. Thirty-three of these items were among the 50 showing DIF in the full set of 114 items, and 30 of them showed the same direction of DIF. We examine DIF impact for the full set of 114 items below.

DIF impact

The impact of language DIF items on test characteristic curves (TCCs) is shown in Fig. 1. The graph on the left of

Table 2 Loadings for 114 physical function items in Spanish sample from categorical one-factor model

Item[a]	Estimate	Standard error	*t* statistic	*p* value
PFA1	0.803	0.017	46.202	0.000
PFA3	0.846	0.014	59.399	0.000
PFA4	0.828	0.016	52.457	0.000
PFA5	0.878	0.013	69.042	0.000
PFA6	0.894	0.012	74.919	0.000
PFA7	0.838	0.015	57.382	0.000
PFA8	0.919	0.010	88.714	0.000
PFA9	0.913	0.009	98.097	0.000
PFA10	0.829	0.015	53.683	0.000
PFA11	0.905	0.011	84.433	0.000
PFA12	0.884	0.012	73.395	0.000
PFA13	0.761	0.019	40.170	0.000
PFA14	0.861	0.014	63.747	0.000
PFA15	0.899	0.012	73.286	0.000
PFA16	0.927	0.008	113.149	0.000
PFA17	0.846	0.015	56.140	0.000
PFA18	0.904	0.011	82.044	0.000
PFA19	0.652	0.026	25.566	0.000
PFA20	0.921	0.009	108.179	0.000
PFA21	0.917	0.009	107.620	0.000
PFA23	0.904	0.011	81.907	0.000
PFA25	0.867	0.013	67.262	0.000
PFA28	0.901	0.013	69.984	0.000
PFA29	0.871	0.013	68.911	0.000
PFA30	0.928	0.007	125.674	0.000
PFA31	0.859	0.013	65.506	0.000
PFA32	0.864	0.013	64.142	0.000
PFA33	0.756	0.019	39.566	0.000
PFA34	0.862	0.014	62.038	0.000
PFA35	0.926	0.008	111.353	0.000
PFA36	0.933	0.008	122.919	0.000
PFA37	0.905	0.011	84.154	0.000
PFA38	0.928	0.007	125.126	0.000
PFA39	0.565	0.030	18.585	0.000
PFA40	0.918	0.009	103.777	0.000
PFA41	0.852	0.014	61.671	0.000
PFA42	0.855	0.013	64.414	0.000
PFA44	0.938	0.008	121.809	0.000
PFA45	0.911	0.013	69.840	0.000
PFA47	0.926	0.009	100.947	0.000
PFA48	0.915	0.011	83.133	0.000
PFA49	0.843	0.014	59.745	0.000
PFA50	0.925	0.009	105.652	0.000
PFA51	0.938	0.007	131.139	0.000
PFA52	0.907	0.011	81.006	0.000
PFA53	0.915	0.011	85.443	0.000
PFA54	0.924	0.009	102.943	0.000
PFA55	0.931	0.007	127.857	0.000
PFA56	0.923	0.009	104.717	0.000
PFB1	0.899	0.011	83.985	0.000
PFB3	0.917	0.009	96.924	0.000
PFB5	0.824	0.015	54.577	0.000
PFB7	0.791	0.018	44.686	0.000
PFB8	0.873	0.012	72.289	0.000
PFB9	0.876	0.013	69.435	0.000
PFB10	0.913	0.010	87.545	0.000
PFB11	0.914	0.011	84.425	0.000
PFB12	0.939	0.007	128.714	0.000
PFB13	0.921	0.010	90.825	0.000
PFB14	0.919	0.011	82.130	0.000
PFB17	0.933	0.007	124.365	0.000
PFB18	0.903	0.013	68.093	0.000
PFB21	0.932	0.007	130.659	0.000
PFB22	0.933	0.008	122.750	0.000
PFB24	0.857	0.014	61.700	0.000
PFB25	0.919	0.011	80.876	0.000
PFB26	0.919	0.011	80.778	0.000
PFB27	0.924	0.009	105.688	0.000
PFB28	0.855	0.014	62.496	0.000
PFB30	0.938	0.008	120.072	0.000
PFB32	0.916	0.011	82.375	0.000
PFB33	0.926	0.008	113.406	0.000
PFB34	0.919	0.010	90.787	0.000
PFB36	0.917	0.012	79.725	0.000
PFB39	0.887	0.012	72.834	0.000
PFB40	0.865	0.014	61.009	0.000
PFB41	0.929	0.008	118.559	0.000
PFB42	0.861	0.013	63.989	0.000
PFB43	0.937	0.008	119.666	0.000
PFB44	0.903	0.010	87.348	0.000
PFB48	0.935	0.008	119.741	0.000
PFB49	0.916	0.010	94.964	0.000
PFB50	0.909	0.011	86.176	0.000
PFB51	0.795	0.017	45.783	0.000
PFB54	0.916	0.011	85.531	0.000
PFB56	0.860	0.014	60.639	0.000
PFC6	0.895	0.013	68.785	0.000
PFC7	0.482	0.033	14.409	0.000
PFC10	0.867	0.013	68.757	0.000
PFC12	0.828	0.016	53.329	0.000
PFC13	0.764	0.019	39.187	0.000
PFC20	0.851	0.014	59.974	0.000
PFC29	0.898	0.011	79.548	0.000
PFC31	0.912	0.010	87.302	0.000

Table 2 continued

Item[a]	Estimate	Standard error	*t* statistic	*p* value
PFC32	0.839	0.014	58.144	0.000
PFC33	0.329	0.038	8.549	0.000
PFC34	0.817	0.016	51.904	0.000
PFC35	0.787	0.019	41.510	0.000
PFC36	0.863	0.014	63.563	0.000
PFC37	0.902	0.011	82.211	0.000
PFC38	0.928	0.009	104.000	0.000
PFC39	0.883	0.012	72.482	0.000
PFC40	0.878	0.012	71.532	0.000
PFC41	0.915	0.009	98.700	0.000
PFC43	0.928	0.008	122.586	0.000
PFC45	0.932	0.009	101.881	0.000
PFC46	0.935	0.009	102.276	0.000
PFC47	0.909	0.010	90.036	0.000
PFC49	0.918	0.010	92.211	0.000
PFC51	0.917	0.009	106.051	0.000
PFC52	0.919	0.011	86.173	0.000
PFC53	0.928	0.010	92.487	0.000
PFC54	0.923	0.010	94.918	0.000
PFC56	0.935	0.008	120.357	0.000

[a] The wording of the items can be seen in Online resource 1

Fig. 1 shows the TCC for all 114 items while the graph on the right shows the TCC for just the 50 items with DIF. These curves indicate that the expected physical functioning total score is higher for Spanish language than English-language respondents. Figure 2 shows the difference between scores ignoring DIF (initial theta) and those that account for DIF (purified). The mean difference is indicated by the dotted line in the right panel (about −1.2), and the median is shown in the box plot in the left panel (about −1.4). These differences are much larger than the standard errors and are substantial in magnitude (greater than a standard deviation). Accounting for DIF tended to result in higher scores, especially for English-language respondents.

Stocking-Lord linking constants to transform linearly the 114 Spanish item parameter estimates to the English metric were as follows:

- Spanish slopes = Spanish calibrated slope/1.21644
- Spanish thresholds = (Spanish calibrated threshold * 1.21644) − 1.749875

The transformation equations for the 64 non-DIF items were as follows:

- Spanish slopes = Spanish calibrated slope/0.995125
- Spanish thresholds = (Spanish calibrated threshold * 0.995125) − 1.117342

The transformed Spanish parameters appear in Online resource 4.

Figure 3 shows the associations between CAT-based theta estimates in the Spanish sample ($n = 640$) based on English parameters (x-axis) and Spanish parameters (y-axis) for the 114 physical functioning items. Figure 4 provides the same but for the 64 non-DIF items. Product-moment correlations between the two CAT-based theta estimates were 0.90 (81 % common variance) and 0.96 (91 % common variance), respectively. Intraclass correlations for the 114 items and for the 64 non-DIF items were 0.89 (CI = 0.87–0.91) and 0. 96 (CI = 0.95–0.97), respectively. Hence, the DIF items have a noteworthy effect on the estimated thetas.

We recommend using English calibrations for the 64 non-DIF items and Spanish calibrations (transformed linearly to English metric) for the other physical functioning items.

Discussion

One of the goals of PROMIS® is to improve precision and enhance the comparability of health outcomes measures [12]. Comparison between different language groups assumes items mean the same to people from the different groups. If subjects respond differently depending on an external variable, group comparisons are problematic. Comparisons of different groups require equivalence in the groups or statistical adjustment for lack of equivalence. Fifty of the 114 items showed differential item functioning in this study of subjects responding to a physical function item bank in English versus Spanish. This indicates that English-language and Spanish-language respondents with the same level of physical functioning respond differently to a substantial number of these items.

One possible explanation for such a high number of items presenting DIF is the validity of the translation. However, the FACIT translation methodology used by PROMIS® to translate the item bank into Spanish [31] is a rigorous and commonly used method. It included representation of people from different countries, which is especially important in Spanish translations to be used in the United States given the diversity of the Spanish-speaking population. A readability analysis of the Spanish item bank was not performed.

If the results of this study are to be generalized to the general United States Latino population, the extent to which the sample of Spanish-language study participants is representative of the United States Latino population is also relevant. According to 2010 Census Data, the United States Latino population has 46 % having a High School diploma or some college and 36 % having less than a

Table 3 Spanish sample slope and threshold parameter estimates (SE) and items presenting DIF (20 with uniform DIF* and 30 with non-uniform DIF**)

Item[a]	Slope	Category threshold			
PFA1**	1.79 (0.14)	−1.39 (0.20)	−0.40 (0.10)	0.69 (0.08)	1.76 (0.09)
PFA3**	2.06 (0.16)	−1.86 (0.27)	−0.52 (0.11)	0.42 (0.07)	1.37 (0.08)
PFA4**	1.99 (0.17)	−1.57 (0.22)	−0.41 (0.11)	0.49 (0.08)	1.30 (0.08)
PFA5*	2.55 (0.20)	−1.52 (0.21)	−0.62 (0.10)	0.29 (0.06)	0.96 (0.07)
PFA6	2.96 (0.24)	−0.76 (0.10)	−0.06 (0.06)	0.63 (0.08)	
PFA7**	2.17 (0.18)	−1.92 (0.29)	−0.60 (0.10)	0.48 (0.07)	1.34 (0.07)
PFA8*	3.71 (0.36)	−1.34 (0.16)	−0.63 (0.07)	−0.17 (0.06)	0.53 (0.06)
PFA9	3.22 (0.25)	−1.46 (0.20)	−0.66 (0.08)	0.09 (0.06)	0.82 (0.06)
PFA10**	2.06 (0.17)	−1.40 (0.19)	−0.64 (0.11)	0.26 (0.08)	1.21 (0.07)
PFA11*	3.21 (0.29)	−1.23 (0.15)	−0.55 (0.08)	0.07 (0.06)	0.95 (0.06)
PFA12	2.70 (0.22)	−1.64 (0.23)	−0.83 (0.10)	−0.02 (0.07)	0.96 (0.06)
PFA13**	1.53 (0.13)	−1.38 (0.20)	−0.49 (0.14)	0.60 (0.09)	2.06 (0.10)
PFA14	2.28 (0.18)	−1.32 (0.18)	−0.56 (0.10)	0.27 (0.08)	1.15 (0.07)
PFA15	3.20 (0.27)	−1.38 (0.18)	−0.76 (0.10)	−0.05 (0.06)	0.67 (0.06)
PFA16	3.57 (0.36)	−0.77 (0.09)	−0.18 (0.06)	0.40 (0.06)	
PFA17	2.13 (0.17)	−1.72 (0.25)	−0.79 (0.12)	0.04 (0.08)	1.01 (0.07)
PFA18	3.17 (0.29)	−1.63 (0.23)	−0.82 (0.10)	−0.11 (0.06)	0.54 (0.06)
PFA19**	1.05 (0.10)	−1.23 (0.24)	−0.01 (0.14)	1.27 (0.11)	2.86 (0.17)
PFA20*	3.29 (0.28)	−0.77 (0.09)	−0.25 (0.06)	0.32 (0.07)	
PFA21*	3.18 (0.32)	−0.67 (0.09)	0.03 (0.06)	0.78 (0.06)	
PFA23*	3.17 (0.29)	−1.38 (0.18)	−0.68 (0.09)	−0.04 (0.06)	0.61 (0.06)
PFA25**	2.35 (0.20)	−1.13 (0.16)	−0.46 (0.09)	0.23 (0.07)	1.19 (0.07)
PFA28	3.35 (0.32)	−1.58 (0.23)	−0.80 (0.09)	−0.15 (0.06)	0.42 (0.06)
PFA29	2.48 (0.20)	−1.41 (0.18)	−0.64 (0.10)	0.15 (0.07)	1.04 (0.07)
PFA30	3.40 (0.29)	−0.68 (0.08)	−0.01 (0.06)	0.68 (0.06)	
PFA31*	2.30 (0.18)	−1.60 (0.22)	−0.66 (0.11)	0.18 (0.07)	1.19 (0.07)
PFA32	2.44 (0.21)	−1.72 (0.26)	−0.67 (0.11)	0.08 (0.07)	0.90 (0.07)
PFA33**	1.53 (0.13)	−1.48 (0.22)	−0.52 (0.13)	0.74 (0.09)	2.00 (0.10)
PFA34	2.36 (0.19)	−1.70 (0.25)	−0.85 (0.12)	−0.01 (0.08)	0.88 (0.07)
PFA35	3.79 (0.39)	−0.73 (0.09)	−0.19 (0.05)	0.29 (0.06)	
PFA36	4.28 (0.44)	−0.72 (0.08)	−0.15 (0.05)	0.31 (0.06)	
PFA37	3.18 (0.27)	−1.63 (0.22)	−0.65 (0.09)	−0.06 (0.06)	0.62 (0.06)
PFA38	3.61 (0.34)	−0.69 (0.08)	−0.06 (0.06)	0.51 (0.06)	
PFA39**	0.80 (0.10)	−1.04 (0.27)	0.12 (0.17)	1.84 (0.15)	3.74 (0.28)
PFA40	3.66 (0.42)	−0.91 (0.10)	−0.27 (0.06)	0.19 (0.06)	
PFA41*	2.26 (0.19)	−1.43 (0.19)	−0.67 (0.11)	0.09 (0.08)	1.14 (0.07)
PFA42**	2.25 (0.19)	−1.39 (0.19)	−0.62 (0.11)	0.13 (0.08)	1.24 (0.07)
PFA44	4.50 (0.48)	−0.81 (0.09)	−0.21 (0.05)	0.23 (0.06)	
PFA45	3.77 (0.35)	−1.46 (0.20)	−0.76 (0.08)	−0.08 (0.06)	0.34 (0.06)
PFA47	3.68 (0.36)	−0.78 (0.09)	−0.13 (0.06)	0.30 (0.06)	
PFA48	3.75 (0.36)	−1.46 (0.19)	−0.72 (0.08)	−0.18 (0.05)	0.22 (0.06)
PFA49	2.19 (0.17)	−1.63 (0.22)	−0.78 (0.11)	0.06 (0.08)	1.11 (0.07)
PFA50*	3.99 (0.49)	−0.82 (0.08)	−0.30 (0.05)	0.06 (0.07)	
PFA51*	4.41 (0.45)	−0.70 (0.07)	−0.19 (0.05)	0.32 (0.05)	
PFA52	3.24 (0.30)	−1.53 (0.21)	−0.76 (0.09)	−0.05 (0.06)	0.54 (0.06)
PFA53*	3.63 (0.33)	−1.38 (0.16)	−0.69 (0.08)	−0.08 (0.06)	0.46 (0.06)
PFA54	4.06 (0.45)	−0.78 (0.08)	−0.22 (0.06)	0.13 (0.06)	

Table 3 continued

Item[a]	Slope	Category threshold			
PFA55	4.05 (0.40)	−0.71 (0.08)	−0.15 (0.05)	0.39 (0.06)	
PFA56	3.66 (0.34)	−0.75 (0.09)	−0.14 (0.06)	0.56 (0.06)	
PFB1*	3.19 (0.24)	−1.60 (0.22)	−0.60 (0.08)	0.21 (0.06)	0.95 (0.07)
PFB3	3.51 (0.29)	−1.52 (0.20)	−0.67 (0.08)	0.04 (0.05)	0.72 (0.07)
PFB5**	1.86 (0.15)	−1.09 (0.16)	−0.31 (0.10)	0.63 (0.08)	1.69 (0.09)
PFB7**	1.62 (0.14)	−0.98 (0.17)	−0.17 (0.10)	0.80 (0.08)	1.92 (0.10)
PFB8**	2.37 (0.20)	−1.22 (0.16)	−0.49 (0.09)	0.16 (0.07)	1.22 (0.07)
PFB9**	2.56 (0.22)	−1.15 (0.15)	−0.71 (0.11)	0.11 (0.07)	0.94 (0.07)
PFB10	3.28 (0.29)	−1.35 (0.17)	−0.71 (0.08)	−0.08 (0.06)	0.57 (0.06)
PFB11	3.51 (0.32)	−1.41 (0.17)	−0.69 (0.08)	−0.04 (0.06)	0.55 (0.06)
PFB12	4.14 (0.38)	−0.66 (0.07)	0.01 (0.05)	0.64 (0.06)	
PFB13	3.93 (0.39)	−1.19 (0.14)	−0.60 (0.07)	−0.10 (0.05)	0.62 (0.06)
PFB14*	4.05 (0.41)	−1.19 (0.13)	−0.75 (0.08)	−0.08 (0.05)	0.42 (0.06)
PFB17	4.12 (0.41)	−0.61 (0.07)	−0.06 (0.05)	0.40 (0.06)	
PFB18	3.44 (0.31)	−1.38 (0.19)	−0.83 (0.09)	−0.19 (0.07)	0.30 (0.07)
PFB21	3.71 (0.35)	−0.68 (0.08)	−0.16 (0.06)	0.38 (0.06)	
PFB22	4.17 (0.41)	−0.68 (0.07)	−0.19 (0.05)	0.33 (0.06)	
PFB24**	2.31 (0.19)	−1.34 (0.18)	−0.53 (0.10)	0.23 (0.07)	1.18 (0.07)
PFB25	4.03 (0.37)	−1.25 (0.15)	−0.73 (0.08)	−0.10 (0.05)	0.40 (0.06)
PFB26	4.12 (0.40)	−1.29 (0.16)	−0.74 (0.07)	−0.18 (0.05)	0.26 (0.06)
PFB27	3.53 (0.26)	−0.68 (0.08)	−0.08 (0.07)	0.44 (0.06)	
PFB28	2.22 (0.18)	−1.15 (0.16)	−0.49 (0.10)	0.23 (0.08)	1.23 (0.07)
PFB30	4.03 (0.38)	−0.75 (0.08)	−0.16 (0.06)	0.39 (0.06)	
PFB32	3.81 (0.36)	−1.21 (0.15)	−0.73 (0.08)	−0.09 (0.06)	0.56 (0.06)
PFB33	3.95 (0.42)	−0.78 (0.08)	−0.18 (0.06)	0.40 (0.06)	
PFB34	3.64 (0.33)	−1.08 (0.14)	−0.56 (0.07)	−0.01 (0.06)	0.67 (0.06)
PFB36	3.60 (0.34)	−1.37 (0.17)	−0.77 (0.08)	−0.17 (0.06)	0.37 (0.06)
PFB39	2.77 (0.24)	−1.32 (0.16)	−0.63 (0.10)	0.02 (0.07)	0.86 (0.07)
PFB40	2.54 (0.20)	−1.32 (0.18)	−0.68 (0.10)	0.00 (0.07)	0.81 (0.07)
PFB41	3.93 (0.39)	−0.74 (0.08)	−0.14 (0.06)	0.28 (0.06)	
PFB42	2.49 (0.21)	−1.34 (0.19)	−0.61 (0.10)	0.18 (0.07)	1.03 (0.07)
PFB43	3.92 (0.34)	−0.68 (0.08)	0.01 (0.05)	0.58 (0.06)	
PFB44**	3.18 (0.26)	−1.15 (0.14)	−0.49 (0.08)	0.20 (0.06)	1.11 (0.07)
PFB48	3.95 (0.36)	−0.70 (0.08)	0.00 (0.05)	0.51 (0.06)	
PFB49*	3.55 (0.35)	−1.44 (0.20)	−0.69 (0.09)	−0.03 (0.05)	0.67 (0.06)
PFB50*	3.36 (0.25)	−1.48 (0.20)	−0.62 (0.09)	0.15 (0.05)	1.15 (0.06)
PFB51**	1.67 (0.15)	−1.19 (0.18)	−0.53 (0.12)	0.41 (0.08)	1.60 (0.10)
PFB54	3.85 (0.31)	−1.44 (0.18)	−0.50 (0.07)	0.08 (0.05)	0.63 (0.07)
PFB56	2.46 (0.22)	−1.74 (0.26)	−0.97 (0.13)	−0.15 (0.08)	0.64 (0.08)
PFC6**	3.39 (0.28)	−1.15 (0.16)	−0.54 (0.08)	0.03 (0.06)	0.68 (0.06)
PFC7**	0.66 (0.09)	−1.05 (0.32)	0.29 (0.20)	2.07 (0.19)	4.18 (0.37)
PFC10**	2.32 (0.19)	−1.29 (0.17)	−0.32 (0.10)	0.49 (0.07)	1.45 (0.07)
PFC12**	1.98 (0.16)	−1.17 (0.16)	−0.39 (0.10)	0.49 (0.07)	1.61 (0.09)
PFC13**	1.48 (0.14)	−1.00 (0.18)	−0.27 (0.13)	0.60 (0.10)	1.82 (0.10)
PFC20**	2.15 (0.18)	−1.39 (0.19)	−0.56 (0.11)	0.26 (0.07)	1.06 (0.07)
PFC29	3.01 (0.26)	−1.49 (0.19)	−0.75 (0.10)	−0.04 (0.06)	0.54 (0.06)
PFC31	3.17 (0.23)	−1.41 (0.19)	−0.62 (0.08)	0.04 (0.07)	0.86 (0.06)
PFC32**	2.02 (0.19)	−1.77 (0.26)	−0.74 (0.13)	0.19 (0.08)	1.21 (0.07)

Table 3 continued

Item[a]	Slope	Category threshold			
PFC33**	0.45 (0.09)	−0.60 (0.42)	1.53 (0.23)	3.60 (0.45)	6.06 (0.79)
PFC34**	1.90 (0.17)	−1.19 (0.19)	−0.58 (0.12)	0.39 (0.08)	1.37 (0.08)
PFC35**	1.69 (0.15)	−1.03 (0.17)	−0.09 (0.10)	0.77 (0.08)	1.73 (0.10)
PFC36**	2.38 (0.18)	−1.13 (0.16)	−0.40 (0.09)	0.46 (0.07)	1.31 (0.08)
PFC37**	3.13 (0.26)	−1.35 (0.17)	−0.56 (0.08)	0.16 (0.06)	0.90 (0.06)
PFC38*	4.14 (0.38)	−1.12 (0.13)	−0.62 (0.07)	−0.08 (0.05)	0.59 (0.06)
PFC39	2.75 (0.23)	−1.33 (0.17)	−0.68 (0.10)	−0.07 (0.07)	0.66 (0.07)
PFC40*	2.68 (0.23)	−1.18 (0.15)	−0.56 (0.09)	0.03 (0.07)	1.05 (0.07)
PFC41*	3.20 (0.27)	−1.38 (0.17)	−0.64 (0.09)	0.07 (0.06)	0.84 (0.06)
PFC43	3.67 (0.35)	−0.68 (0.08)	−0.07 (0.06)	0.46 (0.06)	
PFC45	4.13 (0.41)	−1.34 (0.16)	−0.64 (0.07)	−0.07 (0.06)	0.59 (0.05)
PFC46	4.31 (0.40)	−1.31 (0.15)	−0.63 (0.07)	−0.09 (0.05)	0.54 (0.06)
PFC47	3.41 (0.30)	−1.41 (0.19)	−0.70 (0.09)	0.00 (0.06)	0.62 (0.06)
PFC49	3.72 (0.39)	−1.39 (0.17)	−0.80 (0.09)	−0.21 (0.06)	0.38 (0.06)
PFC51	3.47 (0.36)	−0.75 (0.08)	−0.17 (0.06)	0.51 (0.06)	
PFC52	3.80 (0.32)	−1.35 (0.15)	−0.66 (0.07)	−0.02 (0.06)	0.70 (0.06)
PFC53	4.03 (0.39)	−1.42 (0.18)	−0.64 (0.07)	−0.07 (0.05)	0.58 (0.05)
PFC54*	3.98 (0.35)	−1.17 (0.12)	−0.64 (0.07)	0.00 (0.05)	0.61 (0.06)
PFC56*	3.90 (0.39)	−0.59 (0.07)	−0.12 (0.05)	0.47 (0.06)	

[a] The wording of the items can be seen in Online resource 1

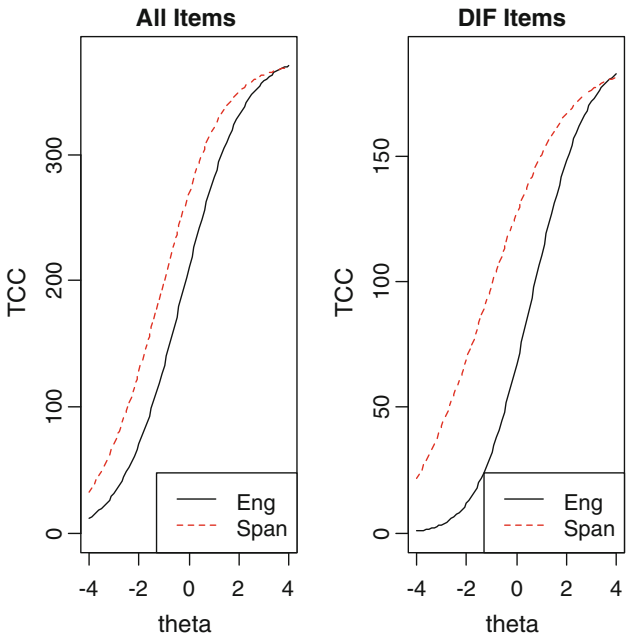

Fig. 1 Impact of DIF on test characteristic curves

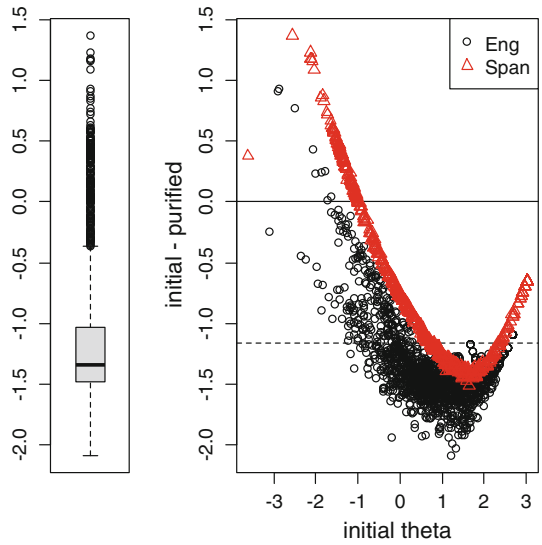

Fig. 2 DIF impact at individual level

completed High School degree, while our sample has 53 % having a High School diploma or some college, and 14 % having less than a completed High School education [32]. We know that Spanish was the language of preference, but we do not know the heritage of the subjects who chose to respond to this online survey. Of the total Latino panel, only 2 % selected Spanish as their language of preference, which makes it harder to understand and characterize the sample. Therefore, the generalizability of the results to the United States Latino population requires future study. Our sample had relatively higher levels of education and low levels of acculturation. In addition, these subjects chose to be part of an online panel, possibly compromising the generalizability to older Latinos or those with lower levels of education. Specifically, online sampling might introduce

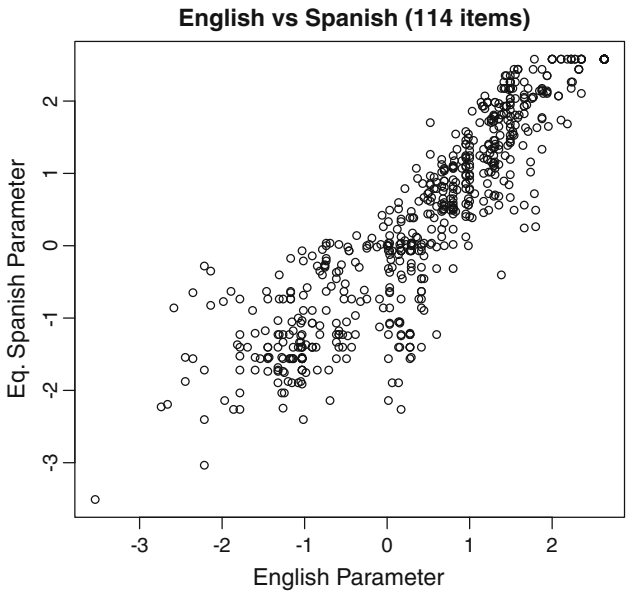

Fig. 3 CAT-based theta estimates using English (*x*-axis) and Spanish (*y*-axis) parameters for 114 items in Spanish sample ($n = 640$, ICC = 0.89)

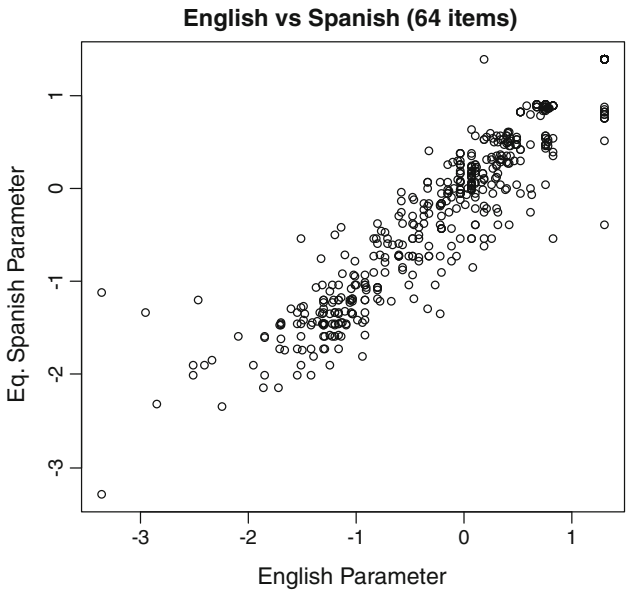

Fig. 4 CAT-based theta estimates using English (*x*-axis) and Spanish (*y*-axis) parameters for 64 non-DIF items in Spanish sample ($n = 640$, ICC = 0.96)

some bias related with computer literacy and its relationship to higher education and socioeconomic status.

Another factor that might need further study is the generalizability of results to older Latinos. In addition, our results cannot be generalized either to a clinical sample with a skewed distribution since our results are based on the sample we studied. It is interesting to note that only in 9 items Spanish-speaking respondents had higher percentages responding to the highest score. When examining these items, all of them correspond to items asking about more strenuous activities; that is, PFC7 *"Are you able to run 5 miles (8 km)?"* PFC32 *"Are you able to climb up 5 flights of stairs?"* and PFC33 *"Are you able to run 10 miles (16 km)?"*

When looking at language differences, it appears that Spanish respondents are much more likely to use response options with the word "some"—that is, *"With some difficulty"* or *"Somewhat."* In fact, for all items, except PFB51 *"Does your health now limit you in participating in active sports such as swimming, tennis, or basketball?"* in which frequencies are the same for both languages, Spanish respondents selected the response with the word "some" more often than English respondents. However, even if not seen in other studies, this could also show that Spanish respondents prefer the middle response more frequently than English respondents. This could show some level of less determination not really committing to an extreme response option. This would be a cultural difference needing further study.

Even though some items had lower discrimination than others, it is still acceptable to include them in the item bank since they are reasonable physical functioning items. However, it is important to note that items with higher discrimination will be selected first in CAT administrations. In order to retain all items in the physical functioning bank, a recalibration was done using a hybrid approach in which English calibration metrics were used for non-DIF items and Spanish calibrations were used for those items presenting DIF. Because some items showed significant language DIF, using English calibrations for all items would lead to inaccurate theta estimates in the Spanish items. Furthermore, there were too many items presenting DIF, so excluding those items was not an option.

As the aging population continues to grow, future studies need to focus in specific physical functioning attributes that decline with aging. Item banks need to ensure that they provide sufficient information regarding the elder population and how the different aspects of physical function decline and are interrelated.

Acknowledgments This paper was supported in part by an NIH cooperative agreement (1U54AR057951). Sylvia H. Paz and Ron D. Hays were supported in part by a grant from the NIA (P30AG021684). Sylvia H. Paz was also supported by NIH/NCRR/NCATS UCLA CTSI Grant Number UL1TR000124. Ron D. Hays was also supported by UCLA/DREW Project EXPORT, NIMHD, (2P20MD000182). The papers' contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf.
2. Shorris, E. (1992). *Latinos: A biography of the people*. New York: W.W. Norton & Co.

3. Morales, L., Kington, R., Valdez, R., et al. (2002). Socioeconomic, cultural, and behavioral factors affecting hispanic health outcomes. *Journal of Health Care Poor Underserved, 13*(4), 477–503.
4. California State Department of Finance. (2002). *Current population survey report: March 2001 data*. Sacramento, November 2002.
5. Los Angeles County Department of Health Services. (2000). *Data collection and analysis division*. Los Angeles: Vital Statistics of Los Angeles County.
6. U.S. Census Bureau: Census 2000 US Demographic profile and population center. Washington, DC 20033 (NP-T4-F) Projections of the total resident population by 5-year age groups, race, and Hispanic origin with special age categories.
7. U.S. Census Bureau: Current population reports (P25-1130) Population projections of the US by age, sex, race, and Hispanic origin.
8. http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf.
9. http://www.census.gov/newsroom/releases/archives/2010_census/cb11-cn146.html.
10. http://www.census.gov/newsroom/releases/archives/population/cb08-123.html.
11. http://seniorjournal.com/NEWS/SeniorStats/5-05-31ProfileOlderAm2004.html.
12. Rose, M., Bjorner, J. B., Becker, J., et al. (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*, 17–33.
13. Bruce, B., Fries, J. F., Ambrosini, D., et al. (2009). Better assessment of physical function: Item improvement is neglected but essential. *Arthritis Research & Therapy, 11*, R191. doi:10.1186/ar2890.
14. Cella, D., Riley, W., Stone, A., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*, 1179–1194.
15. Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., et al. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly, 51*(2), 148–180.
16. http://research.yougov.com/services/scientific_research/ from http://www.polimetrix.com.
17. Rivers, D. (2006). *Sample matching: representative sampling from Internet panels*. Palo Alto, CA: Polimetrix, Inc.
18. Liu, H., Cella, D. F., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the PROMIS internet panel. *Journal of Clinical Epidemiology, 63*(11), 1169–1178.
19. Bonomi, A. E., Cella, D. F., Hahn, E. A., Bjordal, K., Sperner-Unterweger, B., Gangeri, L., et al. (1996). Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. *Quality of Life Research, 5*, 309–320.
20. Cella, D., Hernandez, L., Bonomi, A. E., Corona, M., Vaquero, M., Shiomoto, G., et al. (1998). Spanish language translation and initial validation of the functional assessment of cancer therapy quality-of-life instrument. *Medical Care, 36*, 1407–1418.
21. Lent, L., Hahn, E., Eremenco, S., Webster, K., & Cella, D. (1999). Using cross-cultural input to adapt the Functional Assessment of Chronic Illness Therapy (FACIT) scales. *Acta Oncologica, 38*, 695–702.
22. http://www.nihpromis.org/measures/translations.
23. http://us.toluna.com/.
24. MPlus: Muthen & Muthen. www.statmodel.com/.
25. 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), 22–31.
26. Morales, L. S., Flowers, C., Gutierrez, P., et al. (2006). Item and scale differential functioning of the mini-mental state exam assessed using the Differential Item and Test Functioning (DFIT) framework. *Medical Care, 44*, S143–S151.
27. Du Toit, M. (2003). *IRT from Scientific Software International*. Chicago, IL: SSI, Inc.
28. Choi, S., Gibbons, L., & Crane, P. (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).
29. http://www.education.uiowa.edu/casma/IRTPrograms.htm.
30. Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33*(8), 644–645.
31. Eremenco, S., Cella, D., & Arnold, B. (2005). A comprehensive method for the translation and cross-cultural validation of health status questionnaires. *Evaluation and the Health Professions, 28*(2), 212–232.
32. http://www.census.gov/hhes/socdemo/education/data/cps/2010/tables.html.