Establishing a Common Metric for Physical Function: Linking the HAQ-DI and SF-36 PF Subscale to PROMIS® Physical Function

Benjamin D. Schalet, PhD[1], Dennis A. Revicki, PhD[2], Karon F. Cook, PhD[1], Eswar Krishnan, MD[3], Jim F. Fries, MD[3], and David Cella, PhD[1]

[1]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; [2]Outcomes Research, Evidera, Bethesda, MD, USA; [3]Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, USA.

BACKGROUND: Physical function (PF) is a common health concept measured in clinical trials and clinical care. It is measured with different instruments that are not directly comparable, making comparative effectiveness research (CER) challenging when PF is the outcome of interest.

OBJECTIVE: Our goal was to establish a common reporting metric, so that scores on commonly used physical function measures can be converted into PROMIS scores.

DESIGN: Following a single-sample linking design, all participants completed items from the NIH Patient Reported Outcomes Measurement Information System (PROMIS®) Physical Function (PROMIS PF) item bank and at least one other commonly used "legacy" measure: the Health Assessment Questionnaire (HAQ) or the Short Form–36 physical function ten-item PF scale (SF-36 PF). A common metric was created using analyses based on item response theory (IRT), producing score cross-walk tables.

PARTICIPANTS: Participants (N=733) were part of an internet panel, many of whom reported one or more chronic health conditions.

MAIN MEASURES: PROMIS PF, SF-36 PF, and the HAQ–Disability Index (HAQ-DI).

RESULTS: Our results supported the hypothesis that all three scales measure essentially the same concept. Crosswalk tables for use in CER are therefore justified.

CONCLUSIONS: HAQ-DI and SF-36 PF results can be expressed on the PROMIS PF metric for the purposes of CER and other efforts to compare PF results across studies that utilize any one of these three measures. Clinicians seeking to incorporate PROs into their clinics can collect patient data on any one of these three instruments and estimate the equivalent on the other two.

KEY WORDS: patient-reported outcome; PRO; score linking; SF-36; HAQ-DI; PROMIS; physical function.
J Gen Intern Med 30(10):1517–23
DOI: 10.1007/s11606-015-3360-0
© Society of General Internal Medicine 2015

Electronic supplementary material *The online version of this article (doi:10.1007/s11606-015-3360-0) contains supplementary material, which is available to authorized users.*

Received September 2, 2014
Revised March 6, 2015
Accepted March 16, 2015
Published online May 20, 2015

INTRODUCTION

Patient-reported-outcome (PRO) data quantify patients' perspectives on their symptoms, function, and well-being. PROs are frequently employed in clinical research, including clinical trials, to help evaluate treatment effectiveness from the patient's perspective.[1] Patient-reported physical function—including self-care, instrumental activities of daily living, mobility and dexterity—is a frequently assessed endpoint. Physical function can range from low-level activities, such as brushing one's teeth or walking across a room, to strenuous exercise. Measures of physical function can quantify the impact of chronic health conditions, and in so doing, they can help evaluate whether and how well patients are recovering from disease, trauma or restorative surgery.[2–4]

Across these applications, people use different measures of PF. Three of the more common choices include the Health Assessment Questionnaire (HAQ), the SF-36® ten-item PF scale derived from the Medical Outcomes Study, and the Patient Reported Outcomes Measurement System (PROMIS®) PF item bank, including its various short form and computerized adaptive testing (CAT) options. As a result of having these and other instruments to choose from, there is no current way to standardize PF measures around a common language or metric. Yet, item response theory (IRT) measurement and instrument linking methods make this possible.

Adapting the World Health Organization's (2007) tripartite framework of physical, mental, and social health, PROMIS researchers developed multiple item banks[5], including one for physical function.[6–8] Physical Function is one of several PROMIS domains that overlap with concepts in the Body Functions (B) and Activity and Participation (S) components of the International Classification of Functioning.[9] The PROMIS Physical Function bank (PROMIS PF) comprises items that assess a large range of physical ability and target the subdomains of mobility, upper extremity, and central body function. Because the PROMIS PF, like other PROMIS measures, is supported by an item bank, a collection of items that measure the full range of the domain are calibrated to a mathematical model. This allows users to administer the PROMIS PF in a number of ways. Items can be tailored to individual levels of function with a brief computer adaptive test (CAT). Short-forms of different lengths can be administered (e.g., PROMIS PF short forms of 4, 6, and 8 items are available for download). The instruments have generally shown improved measurement precision over existing measures such as the HAQ-DI and the SF-36® PF, particularly for PROMIS PF CAT and in the moderate range of function.[8,10] In addition, the PROMIS metric uses the T-score (mean=50; standard deviation=10), which is centered on the US general population.[11] Thus, a PROMIS Physical Function T-score of 60 can be interpreted as being one standard deviation higher (better function) than the "average person" in the US.

Although these features have made PROMIS an emerging and appealing option for PF assessment, there will likely continue to be researchers and clinicians who prefer to use existing "legacy" PF assessments such as HAQ-DI and the SF-36 PF. For example, pharmaceutical clinical trials in rheumatology often deploy the HAQ-DI, and may continue to do so, because the US Food and Drug Administration (FDA) has recommended a response measure that relies upon HAQ-DI scores.[12] If the HAQ-DI could be co-located on the same underlying PF continuum as PROMIS, then it would give the FDA the opportunity to extend that same response measure to PROMIS, with its improved measurement precision, or it could enable investigators to express their HAQ-DI scores on the PROMIS metric (mean=50; SD=10).

To create a common PF metric for common outcome reporting and comparative effectiveness research (CER), we set out to "link" the scores from legacy measures to PROMIS by establishing the mathematical relationships between legacy and PROMIS scores. If scores from different instruments can be linked to a common metric, a cross-walk table can be constructed that associates scores from one measure to corresponding scores on another measure.

METHOD

Measures

PROMIS Physical Function. The PROMIS PF item bank consists of 124 items that assess mobility (lower extremity), dexterity (upper extremity), axial or central (neck and back function), and complicated actions that cover multiple domains (e.g., daily living activities).[7,8,13] An example of an item is: "Are you able to carry a laundry basket up a flight of stairs?" The five response options range from "Without any difficulty," to "Unable to do." The item bank can be administered in multiple ways. For example, there is a PROMIS PF 10-item short form with items selected to target the range of physical function with high levels of precision. Scores on this short form correlate very highly (r=0.96) with scores on the full item bank. Other forms of the instrument selected from the 124-item bank include a brief CAT, a 20-item short form, as well as CATs that assess mobility or upper extremity exclusively.[14]

Of the 124 items in the bank, we used 76 as anchor items. By combining these PROMIS PF anchor items with the items of a legacy scale and then concurrently calibrating them, we linked the items of the legacy scale to the PROMIS PF metric. These 76 items were selected because they were included in the final PROMIS PF item bank and each had responses in all five response categories. Because PROMIS items are not scored as sums, but rather on a standardized T-score metric using IRT, scores obtained from different item subsets are readily comparable.

Health Assessment Questionnaire—Disability Index (HAQ-DI). The HAQ-DI[15] consists of 20 questions in eight categories (Dressing and Grooming, Hygiene, Arising, Reach, Eating, Grip, Walking, Outside Activities). Each item has four response options, ranging from "No difficulty" to "Unable to do," corresponding to scores from 0 to 3. Ignoring the use of aids and devices, the items may be scored by identifying the highest score (most disability) on each item in each category, summing these eight items, and then dividing by 8. This yields a score from 0 (no disability) to 3 (most disability).[16] Alternatively, some users of the HAQ-DI sum (or average) each of the 20 items, yielding a summary score ranging from 0 to 60. This later scoring rule has not been as well validated.[17] For the current study, we chose to link scores based on each of these two scoring strategies to the PROMIS PF metric.

Short Form-36 Health Survey Physical Function (SF-36 PF). The SF-36 PF is a subset of the SF-36v2,[18,19]which measures multiple domains of physical and mental health. The PF subscale consists of ten items, using a three point scale in which respondents indicate to what extent their health limits their physical function (e.g., climbing stairs). The items are scored such that higher scores indicated better physical function. In this study, we linked to the raw scores, which ranged from 10 to 30. The SF-36v2 manual provides information on how to convert raw scores to normed-based scores.

Sample

The linking sample was selected from a subset of individuals (N=818) who were part of the original PROMIS PF calibration sample.[8] The data were collected during the PROMIS Wave 1 testing phase by Polimetrix (now YouGov; www.research.yougov.com), a national, web-based polling firm. The sample was drawn from nonclinical participants; however, they included both healthy and unhealthy participants, representing a wide range of physical function. Participants provided background information, ratings of global health, and responses to candidate PROMIS PF items. Most of the sample also completed the HAQ-DI (N=733) and the SF-36 PF (N=719). Note that the SF-36 PF responses were a subset of those who completed the HAQ-DI. Table 1 shows the demographics for the larger group (N=733). The sample's mean score was 0.34 on the HAQ-DI (SD=0.43, range 0 to 3) and 25.8 on the SF-36

Table 1 Demographic Characteristics of Participants for Sample to Link HAQ-DI and SF-36 PF to PROMIS Physical Function (N=733)

	Percentage
Gender	
Female	51
Ethnicity	
Hispanic	11
Race	
White	83
Black / African American	12
Native American	4
Asian	1
Education	
Some high school	2
High school diploma or GED	21
Some college/technical degree/vocational program	45
Further educational attainment	33
Mean age (range)	51 (18–88)

Note. Sample size for linking SF-36 PF was slightly smaller (N=719). Numbers do not necessarily sum to 100 % due to rounding

PF (SD=5.0, range 10 to 30). For sample details, see Appendix A.

Analysis

Multi-Method Approach. Our analytic plan followed the multi-method approach applied in the PROsetta Stone Project and recommended by linking experts.[20] This approach includes methods based on IRT and one commonly used non-IRT method (equipercentile linking). IRT is a family of mathematical models that allow researchers to assign unique values (i.e., parameters) to each item based on how likely people with different levels of the measured construct are to endorse reach response category.[21,22] In the current study, the results of each linking method showed a high degree of similarity, consistent with previous reports for the domains of depression[23], anxiety[24], and fatigue[25]. Here we report only the results of the fixed IRT calibration, consistent with other published reports. We fit the data to the graded response model (GRM)[26], which is the standard IRT model for the calibration of PROMIS instruments.[27] Details on linking methods are in Appendix B; we report on the accuracy of linking in Appendix D.

HAQ-DI Scoring Considerations. IRT-based linking methods use individual item scores as the basis for the link. When legacy measures are scored in a complex way, however, this may pose a problem for IRT linking. In the case of the HAQ-DI, the 0 to 3 summary score is obtained by averaging the *maximum* score in each of eight functional categories. However, IRT-based linking is most accurate when parameters are estimated on all possible items; therefore, we linked using all 20 items (not just the eight maximum items). This scoring strategy yields a summary score ranging from 0 to 60 score for each participant. Because this manner of scoring incorporates all of the items, however, the 0 to 60 scale is not directly comparable to the 0 to 3 scale. That is, dividing the 0 to 60 score by 20 would likely result in a lower score (less disability) than using the maximum eight method described above.

Given these considerations, we conducted two different IRT-based links for the HAQ-DI. In one link, we used each of the 20 HAQ-DI items and estimated parameters for them. This resulted in a PROMIS PF cross-walk table to HAQ-DI scores that range from 0 to 60. In the second link, we treated the maximum scores (within each category, e.g., Hygiene) as a single item score, such that we estimated parameters for only those eight worst HAQ-DI items. This resulted in a PROMIS PF cross-walk table to HAQ-DI scores that ranged from 0 to 3. For short-hand, we distinguish the two resulting linkages as *max-8* and *sum-20*.

Linking Assumptions. The first assumption to be tested is that the linked measures are measuring essentially the same concept. We tested this by inspecting item content, calculating correlations, and estimating the proportion of general factor variance of the combined set of items. In addition to linking assumptions, we tested the unidimensionality assumption of IRT using both confirmatory and exploratory factor analytic methods. Since our planned IRT calibrations required only that the combined item set is sufficiently unidimensional, we conducted these analyses on the combined items (e.g., PROMIS PF and HAQ-DI). For details, see Appendix C.

Score Cross-Walk Table and Figures. We used the item parameter estimates derived from the fixed-parameter calibration to construct a cross-walk table by applying expected a posteriori (EAP) summed scoring. Cross-walk tables can be used to map simple raw summed (or mean) scores from each legacy instrument to T-score values on the PROMIS PF metric. To visualize the relationship and demonstrate the ranges, we plotted linked scores from each legacy measure against their corresponding PROMIS PF scores.

RESULTS

Item Content Overlap

Inspection of item content indicated substantial overlap between the PROMIS and legacy measures. For the HAQ-DI, 16 of 20 items had content that was similar to one or more of the 76 PROMIS PF items. The remaining four HAQ-DI items were similar to items in the full PROMIS PF bank. The contents of each of the ten items of the SF-36 PF were represented by one or more of the 76 items on the PROMIS PF bank. At least 20 % of the PROMIS PF and HAQ-DI items assess upper extremity and mobility function exclusively. The

SF-36 PF, however, included only mobility and mixed activities items; no specific upper extremity items are included in the measure.

Correlations and Classical Item Statistics

Correlations between scores on the PROMIS PF and the legacy instruments were high: 0.91 for PROMIS PF and HAQ-DI (sum-20), 0.93 for PROMIS PF and HAQ-DI (max-8), and 0.91 for PROMIS PF and SF-36 PF. These values are well above suggested thresholds for linking.[28] Classical item statistics calculated on both individual and combined instruments suggested relatively high levels of internal consistency and homogeneity. (See Appendix C for details.).

Cross-Walk Tables and Figures

Once we obtained IRT parameters for legacy items, we scored the data to obtain the PROMIS T-score equivalents of each legacy summed score. Tables 2, 3 and 4 map simple raw summed scores from each legacy instrument to T-score values on the PROMIS PF metric. Each raw summed score and corresponding PROMIS T-score is presented with the standard error associated with the scaled score. Because there were too few people with sufficiently severe disability scores above 53 on the 20-item HAQ-DI, we could not estimate PROMIS values associated with HAQ-DI scores worse than 53. The same holds for the top HAQ-DI score (> 2.88) using the maximum-of-eight-categories rule.

To illustrate the cross-walk results, we also provided two figures that map the PROMIS PF scores (x-axis) to each of the two legacy instruments (y-axis). Figure 1 displays the relationships of scores on both the HAQ-DI (sum-20) and the SF-36 PF to scores on the PROMIS PF. The figure shows that PROMIS scores cover a much wider range of physical function than do either of the legacy measures. The HAQ-DI captures scores in the very low range of physical function, whereas the SF-36 PF covers a higher (and narrower) range.

Table 2 HAQ-DI Scores (20 items summed) Associated with PROMIS Physical Function T-Scores

HAQ-DI Score	PROMIS PF T-score	T-Score SE	HAQ-DI Score	PROMIS PF T-score	T-Score SE
53	12.5	1.7	24	29.9	1.5
52	13.4	2.0	23	30.4	1.5
51	14.2	2.1	22	30.8	1.5
50	15.1	2.2	21	31.3	1.5
49	16.0	2.1	20	31.8	1.5
48	16.9	2.1	19	32.3	1.5
47	17.7	2.0	18	32.8	1.5
46	18.4	1.9	17	33.3	1.5
45	19.1	1.8	16	33.9	1.5
44	19.8	1.8	15	34.4	1.5
43	20.4	1.7	14	35.0	1.6
42	21.0	1.7	13	35.5	1.6
41	21.6	1.6	12	36.1	1.6
40	22.1	1.6	11	36.7	1.6
39	22.7	1.6	10	37.4	1.7
38	23.2	1.6	9	38.1	1.7
37	23.7	1.5	8	38.8	1.8
36	24.2	1.5	7	39.6	1.8
35	24.7	1.5	6	40.4	1.9
34	25.2	1.5	5	41.4	2.0
33	25.7	1.5	4	42.5	2.2
32	26.1	1.5	3	43.9	2.6
31	26.6	1.5	2	45.7	2.9
30	27.1	1.5	1	48.6	3.8
29	27.5	1.5	0	56.8	6.8
28	28.0	1.5			
27	28.5	1.5			
26	28.9	1.5			
25	29.4	1.5			

HAQ-DI=Health Assessment Questionnaire–Disability Index; PROMIS PF=PROMIS Physical Function

Table 3 HAQ-DI Scores (Average of Eight Maximum Scores) Associated with PROMIS Physical Function T-Scores

HAQ-DI Score	PROMIS PF T-score	T-score SE
2.88	17.4	3.4
2.75	20.0	3.1
2.63	21.6	3.0
2.50	23.2	2.7
2.38	24.6	2.5
2.25	25.9	2.4
2.13	27.1	2.3
2.00	28.2	2.2
1.88	29.2	2.2
1.75	30.2	2.1
1.63	31.2	2.1
1.50	32.2	2.1
1.38	33.2	2.1
1.25	34.2	2.1
1.13	35.3	2.2
1.00	36.3	2.2
0.88	37.4	2.3
0.75	38.6	2.3
0.63	39.9	2.3
0.50	41.3	2.4
0.38	43.0	2.7
0.25	45.0	2.9
0.13	48.0	3.6
0.00	56.7	6.8

HAQ-DI=Health Assessment Questionnaire–Disability Index; PROMIS PF=PROMIS Physical Function

Table 4 SF-36 PF Scores Associated with PROMIS Physical Function T-Scores

SF-36 PF Score	PROMIS PF T-score	T-Score SE
10	24.5	4.0
11	28.3	2.8
12	30.3	2.5
13	32.0	2.2
14	33.4	2.1
15	34.8	2.0
16	36.0	2.0
17	37.2	2.0
18	38.4	1.9
19	39.5	1.9
20	40.7	1.9
21	41.8	1.9
22	42.9	1.9
23	44.1	2.0
24	45.3	2.0
25	46.7	2.1
26	48.2	2.3
27	49.9	2.5
28	52.0	2.9
29	55.0	3.5
30	61.7	5.7

PROMIS PF=PROMIS Physical Function; SF-36 PF=Short Form 36 Physical Function

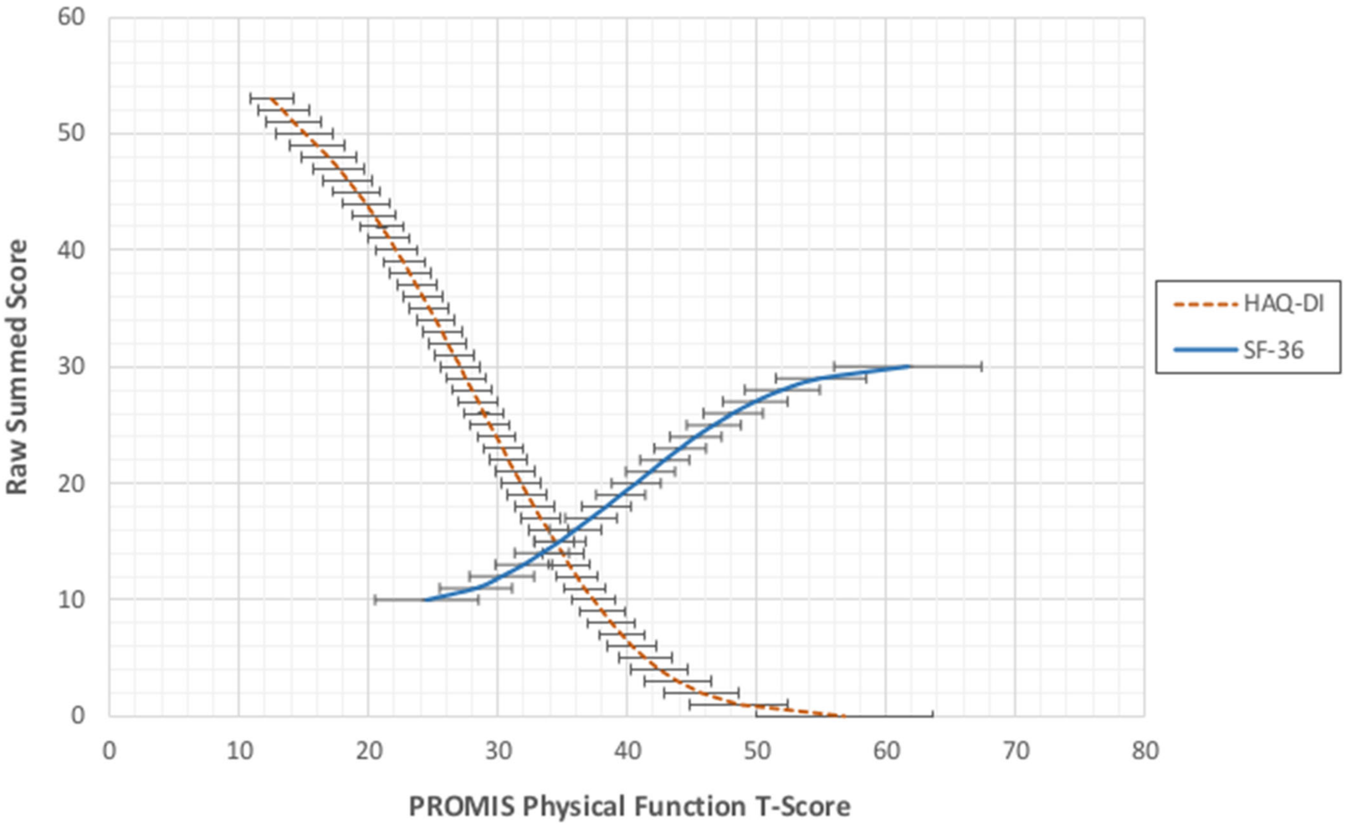

Figure 1 Linking relationships for the HAQ-DI (sum of 20 items) and SF-36 PF to the PROMIS PF metric. The y-axis denotes the raw summed score for both the HAQ-DI and SF-36 PF. The error bars represent±one standard error of measurement derived from the unidimensional IRT model. HAQ-DI=Health Assessment Questionnaire–Disability Index; PROMIS PF=PROMIS Physical Function; SF-36 PF=Short Form 36 Physical Function.

Not shown is the upper range of the PROMIS PF measure, which extends to a T-score of 73. Neither legacy measure extends much beyond the mean of the US population (T-score=50). Figure 2 shows the HAQ-DI on the 0–3 scale.

DISCUSSION

Other researchers have found that measures of physical function are generally amenable to linking and the creation of score cross-walks.[29–31] This study, however, represents the first to link established measures of physical function to the new PROMIS metric. This work has resulted in three cross-walk tables that can be used by researchers and clinicians to convert legacy scores from two popular measures to PROMIS T-scores. In so doing, we have enabled researchers and clinicians to compare scores obtained from one of these instruments with the scores of another.

Our study has a number of strengths. First, the single-group design produces the most robust links.[32] Administering all instruments to all respondents also allowed us to measure directly the accuracy of the linkages by examining differences between actual scores and those predicted by the linking. Second, the correlations of our linked instruments were quite large, exceeding the thresholds recommended by linking experts in the field of high-stakes testing.[28] Finally, our calibrations were not determined by the current sample, but were anchored on PROMIS calibrations derived from the larger standardization sample[8] and centered on the 2000 US census.[11]

The cross-walk tables have several practical uses. For example, current users of legacy measures contemplating a switch to PROMIS PF will be able to “retrofit” their historical patient data by assigning PROMIS PF scores using the cross-walks we provided. This is especially powerful at the aggregate level (e.g., groups of patients), as the error associated with these linkages becomes very small as the sample size exceeds 75 (see Appendix D). Secondly, our results allow clinical investigators to compare results across treatment trials in which different instruments were used. That is, using the IRT-based cross-walk tables, investigators can convert summary mean scores reported in the literature from one metric to another.

These results have particular relevance for investigators conducting clinical trials. Recent draft recommendations from the Federal and Drug Administration (FDA) endorsed the HAQ-DI for use in drug trails of rheumatoid arthritis (RA)[12].

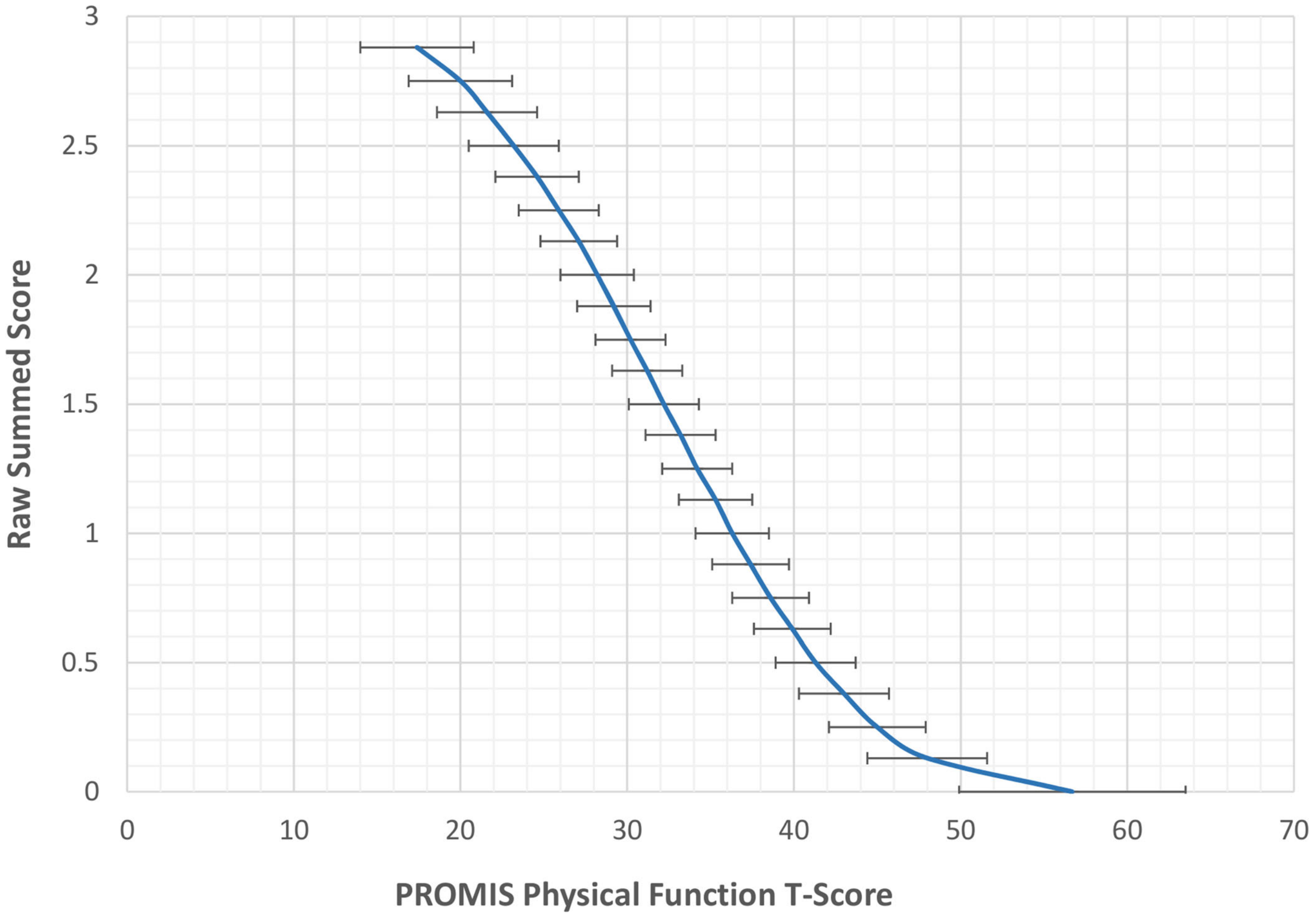

Figure 2 Linking relationship for the HAQ-DI (0 to 3 score) and the PROMIS PF. The HAQ-DI is scored by taking the average of each maximum item score in eight function categories. The error bars represent±one standard error of measurement derived from the unidimensional IRT model. HAQ-DI=Health Assessment Questionnaire–Disability Index; PROMIS PF=PROMIS Physical Function; SF-36 PF=Short Form 36 Physical Function.

Nevertheless, recent studies have demonstrated the advantage of a ten-item CAT and 20-item short-form of the PROMIS PF instrument compared to the HAQ-DI in terms of measurement precision and range of coverage[8,10]. These results suggest that the PROMIS PF-20 can be used in place of the HAQ-DI in clinical trials, and still allow for the estimation of HAQ-DI scores for comparison with earlier clinical trials.

Our results also facilitate PRO use in clinical settings. Given the advantages of PROMIS instruments,[33] clinicians in medical centers who are already administering the HAQ or SF-36 PF may choose to switch to PROMIS PF. By using Tables 2, 3 and 4, they can compare historical patient data to newly obtained scores on PROMIS. Clinicians already using PROMIS may now connect their patient scores to recommended norms and clinical cutoffs established for the HAQ or SF-36[33–35] and can provisionally apply these linked PROMIS cut-offs to inform treatment decisions.

There are some study limitations. First, scores linked to the PROMIS metric based on legacy scores may have more error than scores obtained directly from the PROMIS PF measure and vice versa. Standard errors for cross-walked scores with samples of less than 25 participants may not be adequate for some purposes. Secondly, linking results (regardless of statistical method) may be sensitive to population differences.[32] A recent study, however, found that the linking relationships for PROMIS Pain Interference and the Brief Pain Inventory were very similar when derived from general population and multiple sclerosis groups.[36,37] Nevertheless, it will be necessary to replicate our study with samples drawn from populations with a higher density of scores at either end of the physical function continuum.

In conclusion, the concept of physical function is measured quite comparably by the HAQ-DI, the SF-36 PF, and the PROMIS-PF. We encourage investigators and clinicians to use these cross-walk tables (Tables 2, 3 and 4). We also encourage others to extend this work by linking still other PF measures to PROMIS, making it possible to arrive at a common, unifying language for self-reported physical function.

Acknowledgements: *This research was part of the PROsetta Stone® project, which was funded by the National Institutes of Health/National Cancer Institute grant RC4CA157236 (David Cella, PI). For more information on PROsetta Stone, please see www.prosettastone.org.*

Conflict of Interest: *Karon F. Cook is an unpaid officer of the PROMIS Health Organization and David Cella is an unpaid member of the board of directors and officer of the PROMIS Health Organization. All other authors declare that they do not have a conflict of interest.*

Corresponding Author: *Benjamin D. Schalet, PhD; Department of Medical Social SciencesNorthwestern University Feinberg School of Medicine, 625 N. Michigan Avenue, Suite 2700, Chicago, IL 60611, USA (e-mail: b-schalet@northwestern.edu).*

REFERENCES

1. **Basch E.** New frontiers in patient-reported outcomes: adverse event reporting, comparative effectiveness, and quality assessment. Annu Rev Med. 2014;65(1):307–317. doi:10.1146/annurev-med-010713-141500.
2. **Hung M, Nickisch F, Beals TC, Greene T, Clegg DO, Saltzman CL.** New paradigm for patient-reported outcomes assessment in foot & ankle research: computerized adaptive testing. Foot Ankle Int. 2012;33(8):621–626.
3. **Papuga MO, Beck CA, Kates SL, Schwarz EM, Maloney MD.** Validation of GAITRite and PROMIS as high-throughput physical function outcome measures following ACL reconstruction. J Orthop Res. 2014;32(6):793–801.
4. **Valderas J, Kotzeva A, Espallargues M, Guyatt G, Ferrans C, Halyard M, et al.** The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. Qual Life Res. 2008;17(2):179–193.
5. **Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al.** The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol. 2010;63(11):1179–1194. doi:10.1016/j.jclinepi.2010.04.011.
6. **Fries JF, Cella D, Rose M, Krishnan E, Bruce B.** Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. J Rheumatol. 2009;36(9):2061–2066. doi:10.3899/jrheum.090358.
7. **Rose M, Bjorner JB, Becker J, Fries JF, Ware JE.** Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol. 2008;61(1):17–33.
8. **Rose M, Bjorner JB, Gandek B, Bruce B, Fries JF, Ware JE Jr.** The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. J Clin Epidemiol. 2014;67(5):516–526. doi:10.1016/j.jclinepi.2013.10.024.
9. **Tucker C, Cieza A, Riley A, Stucki G, Lai J, Bedirhan Ustun T, et al.** Concept Analysis of the Patient Reported Outcomes Measurement Information System (PROMIS®) and the International Classification of Functioning, Disability and Health (ICF). Qual Life Res. 2014;6:1677–1686. doi:10.1007/s11136-014-0622-y.
10. **Fries JF, Krishnan E, Rose M, Lingala B, Bruce B.** Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011;13(5):R147. doi:10.1186/ar3461.
11. **Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al.** Representativeness of the PROMIS Internet Panel. J Clin Epidemiol. 2010;63(11):1169–1178.
12. US Food and Drug Administration. Draft Guidance for industry. Qualification process for drug development tools. 2010. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf. Accessed June 30 2014.
13. **Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Ware JE Jr, et al.** Better assessment of physical function: Item improvement is neglected but essential. Arthritis Res Ther. 2009;11(6):R191. doi:10.1186/ar2890.
14. **Hays RD, Spritzer KL, Amtmann D, Lai J-S, DeWitt EM, Rothrock N, et al.** Upper Extremity and Mobility Subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS®) Adult Physical Functioning Item Bank. Arch. Phys. Med. Rehabil. 2013;Epub ahead of print.
15. **Fries JF, Spitz P, Kraines RG, Holman HR.** Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23(2):137–145.
16. **Bruce B, Fries JF.** The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol. 2003;30(1):167–178.
17. **Bruce B, Fries JF.** The Health Assessment Questionnaire (HAQ). Clin. Exp. Rheumatol. 2005;23(5).
18. **Ware JE Jr, Sherbourne CD.** The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual Framework and Item Selection. Med Care. 1992;30(6):473–483.
19. **Ware JE, Kosinski M, Dewey JE.** How to score version 2 of the SF-36 health survey. Lincoln, R.I.: QualityMetric; 2000.
20. **Kolen MJ, Brennan RL.** Test equating, scaling, and linking: methods and practices. New York: Springer; 2004.
21. **Reeve B, Fayers PM.** Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers PM, Hays R, eds. Assessing quality of life in clinical trials. Oxford: New York Oxford University Press; 2005:55–73.
22. **Edelen MO, Reeve BB.** Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res. 2007;16(Suppl 1):5–18. doi:10.1007/s11136-007-9198-0.
23. **Choi SW, Schalet B, Cook KF, Cella D.** Establishing a Common Metric for Depressive Symptoms: Linking the BDI-II, CESD, and PHQ-9 to PROMIS Depression. Psychol Assess. 2014;26(2):513–527. doi:10.1037/a0035768.
24. **Schalet BD, Cook KF, Choi SW, Cella D.** Establishing a common metric for self-reported anxiety: Linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. J Anxiety Disord. 2014;28(1):88–96. doi:10.1016/j.janxdis.2013.11.006.
25. **Lai J-S, Cella D, Yanez B, Stone A.** Linking Fatigue Measures on a Common Reporting Metric. J Pain Symptom Manag. 2014;48(4):639–648. doi:10.1016/j.jpainsymman.2013.12.236.
26. **Samejima F.** Estimation of latent ability using a response pattern of graded scores. . Richmond, VA: Psychometric Society; 1969. Available from: http://www.psychometrika.org/journal/online/MN17.pdf.
27. **Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al.** Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45(5 Suppl 1):S22–S31. doi:10.1097/01.mlr.0000250483.85507.04.
28. **Dorans NJ.** Equating, Concordance, and Expectation. Appl Psychol Meas. 2004;28(4):227–246. doi:10.1177/0146621604265031.
29. **Fisher WP, Eubanks RL, Marier RL.** Equating-the MOS SF36 and the LSU HSI Physical Functioning Scales. J Outcome Meas. 1997;1(4):329–362.
30. **Holzner B, Bode RK, Hahn EA, Cella D, Kopp M, Sperner-Unterweger B, et al.** Equating EORTC QLQ-C30 and FACT-G scores and its use in oncological research. Eur J Cancer. 2006;42:3169–3177.
31. **ten Klooster P, Oude Voshaar M, Gandek B, Rose M, Bjorner J, Taal E, et al.** Development and evaluation of a crosswalk between the SF-36 physical functioning scale and Health Assessment Questionnaire disability index in rheumatoid arthritis. Health Qual Life Outcomes. 2013;11(1):199.
32. **Dorans NJ.** Linking Scores from Multiple Health Outcome Instruments. Qual Life Res. 2007;16(Supplement 1):85–94. doi:10.2307/40212575.
33. **Wagner LI, Schink J, Bass M, Patel S, Diaz MV, Rothrock N, et al.** Bringing PROMIS to practice: Brief and precise symptom screening in ambulatory cancer care. Cancer. 2015;121(6):927–934. doi:10.1002/cncr.29104.
34. **Krishnan E, Sokka T, Häkkinen A, Hubert H, Hannonen P.** Normative values for the Health Assessment Questionnaire disability index: benchmarking disability in the general population. Arthritis Rheum. 2004;50(3):953–960.
35. **Krishnan E, Tugwell P, Fries JF.** Percentile benchmarks in patients with rheumatoid arthritis: Health Assessment Questionnaire as a quality indicator (QI). Arthritis Res Ther. 2004;6(6):505–513.
36. **Chandratre P, Roddy E, Clarson L, Richardson J, Hider SL, Mallen CD.** Health-related quality of life in gout: a systematic review. Rheumatology. 2013;52(11):2031–2040.
37. **Askew R, Kim J, Chung H, Cook K, Johnson K, Amtmann D.** Development of a crosswalk for pain interference measured by the BPI and PROMIS pain interference short form. Qual Life Res. 2013;22(10):2769–2776. doi:10.1007/s11136-013-0398-5.
38. **Cook KF, Kallen MA, Amtmann D.** Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res. 2009;18(4):447–460. doi:10.1007/s11136-009-9464-4.