# Progress in Assessing Physical Function in Arthritis: PROMIS Short Forms and Computerized Adaptive Testing

JAMES F. FRIES, DAVID CELLA, MATTHIAS ROSE, ESWAR KRISHNAN, and BONNIE BRUCE

*ABSTRACT*. ***Objective.*** Assessing self-reported physical function/disability with the Health Assessment Questionnaire Disability Index (HAQ) and other instruments has become central in arthritis research. Item response theory (IRT) and computerized adaptive testing (CAT) techniques can increase reliability and statistical power. IRT-based instruments can improve measurement precision substantially over a wider range of disease severity. These modern methods were applied and the magnitude of improvement was estimated.

***Methods.*** A 199-item physical function/disability item bank was developed by distilling 1865 items to 124, including Legacy Health Assessment Questionnaire (HAQ) and Physical Function-10 items, and improving precision through qualitative and quantitative evaluation in over 21,000 subjects, which included about 1500 patients with rheumatoid arthritis and osteoarthritis. Four new instruments, (A) Patient-Reported Outcomes Measurement Information (PROMIS) HAQ, which evolved from the original (Legacy) HAQ; (B) "best" PROMIS 10; (C) 20-item static (short) forms; and (D) simulated PROMIS CAT, which sequentially selected the most informative item, were compared with the HAQ.

***Results.*** Online and mailed administration modes yielded similar item and domain scores. The HAQ and PROMIS HAQ 20-item scales yielded greater information content versus other scales in patients with more severe disease. The "best" PROMIS 20-item scale outperformed the other 20-item static forms over a broad range of 4 standard deviations. The 10-item simulated PROMIS CAT outperformed all other forms.

***Conclusion.*** Improved items and instruments yielded better information. The PROMIS HAQ is currently available and considered validated. The new PROMIS short forms, after validation, are likely to represent further improvement. CAT-based physical function/disability assessment offers superior performance over static forms of equal length. (J Rheumatol 2009;36:2061–6; doi:10.3899/jrheum.090358)

*Key Indexing Terms:*
ITEM RESPONSE THEORY COMPUTERIZED ADAPTIVE TESTING PROMIS
HEALTH ASSESSMENT QUESTIONNAIRE DISABILITY INDEX

With the advent of modern psychometric tools such as item response theory (IRT)[1,2] and computerized adaptive testing (CAT)[3], shorter instruments with more precise item estimates are now possible and require fewer subjects for the same degree of statistical power. These permit a quantum advance in the science of outcome assessment.

*From the Department of Medicine, Stanford University School of Medicine, Palo Alto, California; Evanston Northwestern Healthcare and Northwestern University, Evanston, Illinois; QualityMetric, Inc., Lincoln, Rhode Island, USA; and Department of Medicine, University Clinic Hamburg Eppendorf, Hamburg, Germany.*

*Supported by the US National Institutes of Health Patient-Reported Outcomes Measurement Information System Roadmap Cooperative Agreements U01 AR052158 and U01 AR52177.*

*J.F. Fries, MD, Professor of Medicine, Stanford University School of Medicine; D. Cella, PhD, Professor of Psychiatry and Behavioral Sciences, Northwestern University Medical School; M. Rose, MD, Senior Scientist, QualityMetric; E. Krishnan, MD, Assistant Professor of Psychiatry and Behavioral Sciences; B. Bruce, DrPH, MPH, RD, Senior Research Scientist, Stanford University.*

*Address correspondence to Dr. J.F. Fries, Division of Immunology and Rheumatology, Stanford University School of Medicine, 1000 Welch Rd., Suite 203, Palo Alto, CA 94304, USA. E-mail: jff@stanford.edu*

The concept of a physical function/disability outcome domain in arthritis is nearly 70 years old[4], and the major patient reported outcome (PRO) instruments[5-7] are more than 30 years old. These measures are accepted, essentially universally used, and have become a standard of PRO assessment in observational and clinical studies. The term "disability" is traditional in rheumatology, where the clinical goal is to reverse decrements caused by the disease. The term "physical function," conceptually the reciprocal of disability, encourages assessment of a wider range of functioning, including functioning better than the population average. The term "physical function" is gradually gaining favor, but both are used here.

The Patient Reported Outcomes Measurement Information System (PROMIS) is a US National Institutes of Health (NIH) Roadmap initiative involving 7 primary institutions with the aim of improving outcome assessment science and effectiveness[8]. The purpose of this PROMIS project was to determine the degree of improvement to be expected from development of IRT and CAT-based instruments and whether new comparability problems between Internet and mail-administered instruments might be introduced.

## MATERIALS AND METHODS

*Domain definition.* For the PROMIS, the "physical function/disability" domain is a subdomain of physical health, which is in turn a subdomain of health (http://www.nihpromis.org). In IRT, a "latent trait" representing a concept such as "physical function" is estimated from multiple items addressing different facets of the trait. The PROMIS definition of the physical function latent trait is the ability to perform activities of daily living (ADL) and instrumental ADL (available from http://www.nihpromis.org). This trait is based on the ability to perform, not on whether or not an activity actually has been performed. It has a capability stem, a capability response, avoids time-based response options, uses the present tense, and avoids attribution to disease or other limiting context.

*Development of the physical function item bank.* A PROMIS physical function/disability item bank containing 124 items was developed from 1865 extant physical-function related items that were identified from 160 published English-language instruments. These items underwent extensive qualitative evaluation with patient surveys and focus groups[8-11]. They were empirically tested in more than 21,000 persons from the general population, which included clinical samples of 1473 adults with self-reported arthritis [osteoarthritis, n = 916; rheumatoid arthritis (RA), n = 557], all of whom were recruited from an Internet sample (http://www.nihpromis.org). The Legacy (original) HAQ and Legacy Physical Function-10 (PF-10) items were included in the item bank, bringing the total number of items to 199. During the quantitative evaluation process, each item was answered by more than 2200 subjects. IRT methods[12] were used to calibrate the final PROMIS physical function item bank.

*Development of new instruments.* Three new short forms and a CAT were developed. "Short forms" are fixed (static) questionnaires of limited length (generally 5–20 items) designed to use the best items to estimate the latent trait. CAT instruments select the best items for the particular patient from a larger pool, sequentially selecting items until a given degree of accuracy is achieved, generally after 5—10 items have been administered. Instruments derived from the PROMIS item bank were compared with Legacy instruments and with each other. The Legacy instruments were the 20-item HAQ[5,13] and the 10-item PF-10 of the Medical Outcome Study Short-form 36[7]. The new instruments were: (A) a 20-item PROMIS HAQ, which evolved from the HAQ; (B) a PROMIS 10-item static, or short, form with items selected as the "best" from the 199 physical function items; (C) a PROMIS 20-item static form also selected from the "best" PROMIS items; and (D) a CAT designed for simulation of dynamic testing of the physical function item bank. The simulated CAT sequentially selected items from the 199 physical function items, which had the greatest information content for the individual patient. The simulated CAT was terminated after 10 items for each patient. Simulation of CAT performance is done by administering the entire item bank to the test population, then using the computer to selectively follow test sequences unique to the individual and ignoring the remaining items. Thus, the group simulated a CAT as each patient would have responded to a personalized 10-item subset of items.

*Evolution of the PROMIS HAQ.* The PROMIS HAQ was evolved from the Legacy HAQ and contains the same 20 items but has a few distinct differences. A fifth response option, "with a little bit of difficulty," was added, the context was changed to the present tense, item clarity was improved, and the aids and devices items improved and reduced from 44 to 24 items[14]. The scoring algorithm was changed from the HAQ's 0–3 unit scale to a 0–100 unit scale. Completion time was reduced by over one-third. However, the many validation studies performed for the HAQ over the years[15] should apply to the PROMIS HAQ as the 2 instruments are closely similar.

*Internet versus traditional mailed administration.* CAT administration requires that an outcome instrument be completed electronically (e.g., a PC, tablet-PC, Internet, personal digital assistant, telephone, etc.). This requires, at a minimum, that computer-based data collection be as practical and valid as traditional mailed administration of the same items.

*Mode of administration study.* In an initial study of the effects of mode of administration, we compared Internet with mailed administration of our tools. We placed 378 subjects from RA, osteoarthritis, and normal aging cohorts over the age of 65[16-18] into 2 groups by their reported capabilities in Internet use (experienced, not experienced). The Internet-experienced group was then randomly divided to complete instruments either over the Internet or by mail. The 3 groups were further randomized to receive either the Legacy or the improved PROMIS items. Figure 1 presents the procedure and response rates of group formation. Both the Internet and mailed administration groups received the same 3 rounds of followup: E-mails or telephone calls, postcards or E-mails, and repeat mailings of questionnaires or E-mails with online links. Every attempt was made to ensure that followup protocols were identical among groups.

## RESULTS

*Mode of administration study.* The majority of subjects were white (~95%) and female (~65%). They averaged 16 years of education. Those with computer experience were on average 6 years younger than those not computer-experienced. Completion rates across the 6 groups were similarly high and comparable, ranging from 92% to 98% (chi-square = 3.59, p = 0.61; Figure 1). Thus, under these circumstances response rates in those solicited over the Internet were nearly as high as those queried by mail. Completion rates were also similar for Legacy and PROMIS instruments.

On examination of the issues of missing data and subjects who reported inability to perform a task, we discovered a striking anomaly, which had not been previously noted. More than half of all missing data over all 20 items from both the Legacy HAQ and the PROMIS HAQ, and more than half of all ceiling effects occurred with the single item "Are you able to get in and out of a bathtub?" Upon debriefing, about one-fourth of subjects who did not have or did not use a bathtub had either left the item blank or had reported erroneously that they were unable to do the activity, suggesting that this item is no longer a good "activity of daily living" indicator.

*Information content of items and instruments.* Figures 2, 3, and 4 show distinctions in the information content between new and Legacy items and instruments[19,20]. The X axis represents the disability level with a population mean set to zero and with each one unit above or below zero representing one standard deviation (SD). The Y axis represents the measurement precision curves [standard error (SE)], which demonstrate precision at different levels of physical function/disability. An optimal instrument would have the greatest measurement precision over the broadest range of disease severity. An SE of 2.3 is equivalent to a reliability (internal consistency) of 0.95. A useful metric for comparison is the number of SD covered at an SE of 2.3 or less, which approximates an area under the curve approach.

*PROMIS HAQ validation.* The PROMIS HAQ, which was derived from the 20-item HAQ, closely parallels its Legacy antecedent. Spearman correlations (Table 1) between individual PROMIS HAQ items ranged from 0.37 to 0.75 (all $p < 0.0001$), suggesting that these items assess a similar latent trait. The same result was obtained with Legacy HAQ items

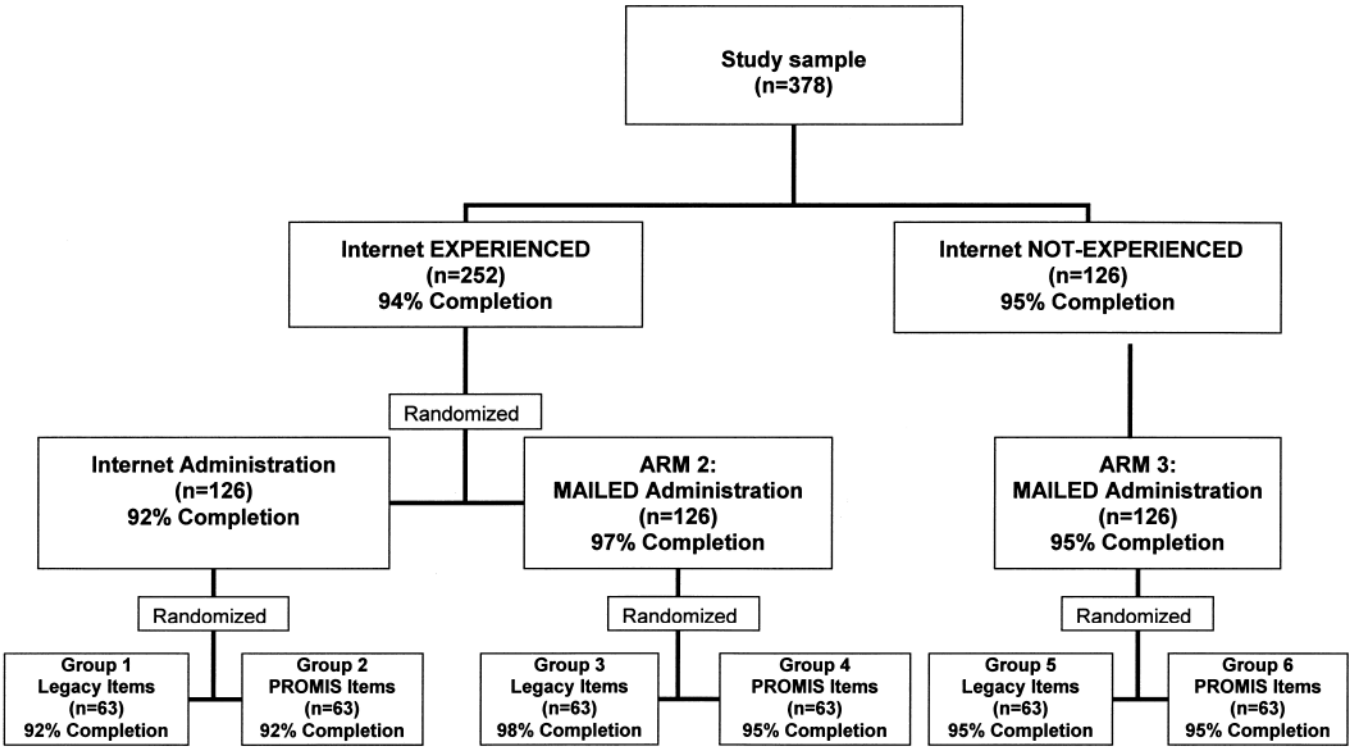


*Figure 1. The* group formation process for examination of Internet versus mailed administration of Legacy versus PROMIS physical function items.

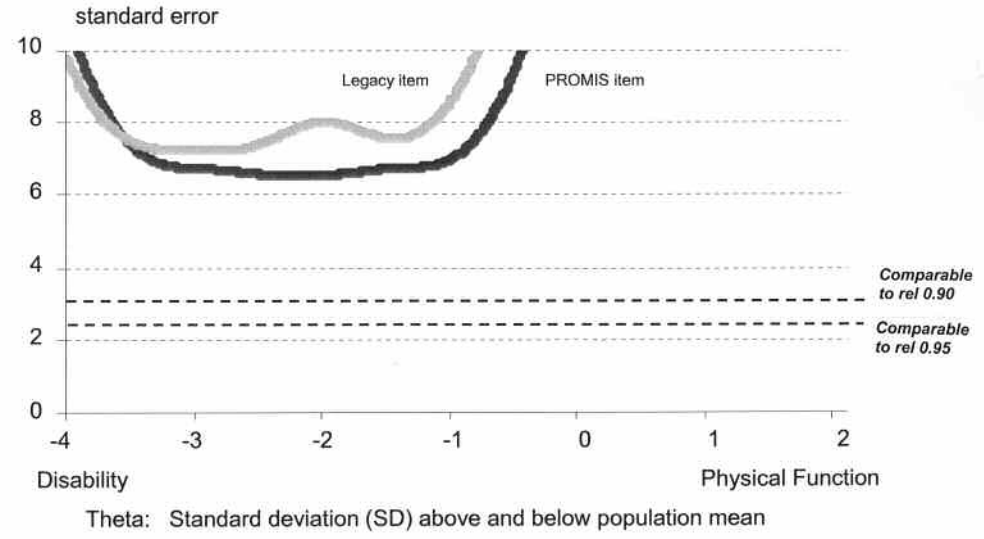


*Figure 2.* Comparison of information content (reliability) of a sample physical function item ("bend down and pick up clothing from the floor") between the Legacy HAQ and the PROMIS HAQ. A better item will have a lower curve (more precision) and a broader range of applicability in terms of theta. Zero represents the population mean.

(data not shown). There was no significant bias across instruments, and mean values were closely similar (Table 2).

*Instrument scoring.* Several scoring algorithms applicable to either the Legacy HAQ or the PROMIS HAQ are currently being evaluated. The Legacy HAQ, as traditionally scored, averages the highest values in 8 categories, each of which contains 2 to 3 items. Alternatively, the number of items can be reduced to 16, with 2 to a category, with some advantages, such as removing the poorly performing bathtub item and shortening the instrument without appreciable loss (preferred). Finally, a simple average of the 20 items may be used. However, this gives results with a smaller effect size as esti-

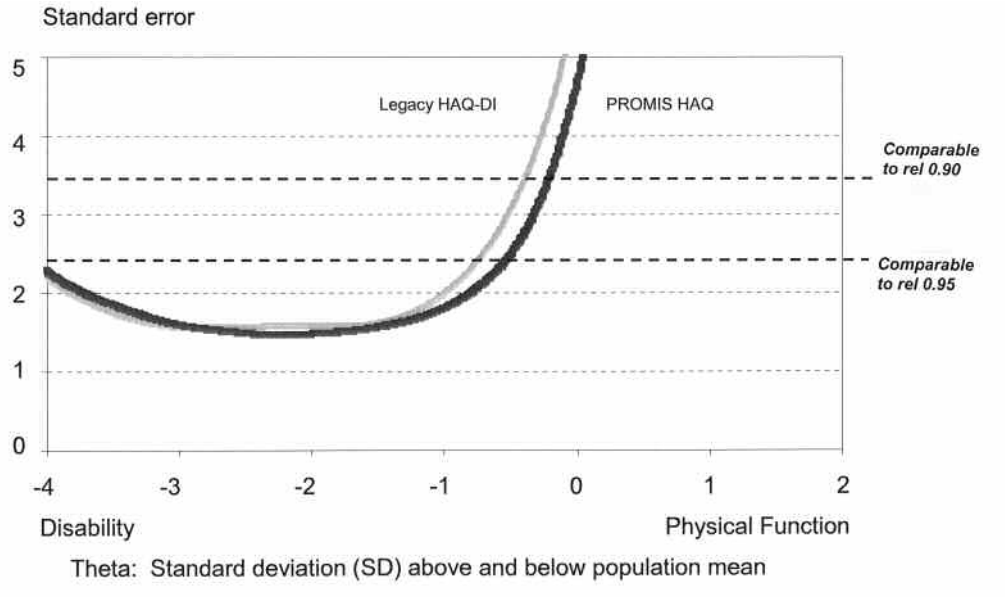


*Figure 3.* Comparison of physical function instrument precision relative to range of measurement between the Legacy HAQ and the PROMIS HAQ. This comparison of instruments shows much greater precision than items, with much of the range having a reliability > 0.95.

*Table 1.* Spearman correlations for PROMIS HAQ items with the highest mean scores within each of the 8 categories. All correlations are $p < 0.0001$.

| | Stand | Carton | Walk | Tub Bath | Reach Up | Open Jars | Chores |
|---|---|---|---|---|---|---|---|
| Dress | 0.63 | 0.61 | 0.61 | 0.67 | 0.62 | 0.56 | 0.67 |
| Stand | 1 | 0.42 | 0.71 | 0.65 | 0.55 | 0.40 | 0.60 |
| Carton | | 1 | 0.37 | 0.56 | 0.65 | 0.69 | 0.61 |
| Walk | | | 1 | 0.65 | 0.49 | 0.37 | 0.64 |
| Tub bath | | | | 1 | 0.62 | 0.46 | 0.75 |
| Reach up | | | | | 1 | 0.57 | 0.64 |
| Open jars | | | | | | 1 | 0.59 |

*Table 2.* Comparison of Legacy and PROMIS physical function scores.

| Feature | PROMIS HAQ, n = 178 | Legacy HAQ, n = 180 | p |
|---|---|---|---|
| Age, yrs* | 73 ± 11 | 70 ± 13 | 0.03 |
| Education level, yrs* | 16 ± 2 | 16 ± 2 | 0.62 |
| Female, % | 66 | 65 | 0.88 |
| White, % | 94 | 94 | 0.84 |
| HAQ score | | | |
| With aids and devices | 25 ± 25 | 24 ± 23 | 0.67 |
| Without aids and devices | 24 ± 24 | 19 ± 20 | 0.06 |
| Subjects with HAQ scores of zero (no disability) | | | |
| With aids and devices, % | 25 | 26 | 0.95 |
| Without aids and devices, % | 25 | 28 | 0.59 |

* Mean ± standard deviation.

mated by comparing the mean value with the SD. Thus, this appears to be the least useful of the scoring algorithms examined.

Combining the PROMIS HAQ and the Legacy HAQ groups from Figure 1 and contrasting the 2 instruments with scoring standardized on a 0–100 scale shows no significant differences. Mean scores were 25 and 24 ($p = 0.67$), and percentages of zero scores were 25% and 26% (Table 2). The aids

and devices questions may play a somewhat larger role in the HAQ than in the PROMIS HAQ, but this difference did not reach statistical significance (Table 2).

*Information content of items and instruments.* Each item was first improved by qualitative methods[10,21] and then assessed quantitatively in large populations using IRT[1]. Figure 2 illustrates improvement in the item "bend down and pick up clothing from the floor" from the Legacy HAQ and the PROMIS HAQ. The scale has zero set at the mean of a normal population, and each integer represents one SD from that mean. The height of the curve above the standard error of zero represents the item information content[19,20]. The information content of the PROMIS item is spread more broadly, meaning that it provides information across a greater range of function. In this instance, it is broader by about one SD, and the curve is lower, indicating that there is more information at every point of functional impairment. This is one of the more improved items. Note, however, that the SE of even the improved items remains greater than 6.

Comparing instruments rather than items, Figure 3 shows that the PROMIS HAQ outperforms the Legacy HAQ, albeit by a relatively small amount, primarily by expanding the breadth of its applicability by about one-fourth of an SD. Most likely, this effect is largely due to the addition of the fifth response option. For these instruments, the SE is as low as 1.5 in some ranges, representing a reliability substantially above 0.95[19,20].

Figure 4 compares information content and reliability across 6 instruments: (from left to right) the PROMIS 10-item static form; the 10-item PF-10; the 20-item Legacy HAQ; the 20-item PROMIS HAQ; the 20-item PROMIS static form; and the PROMIS CAT 10-item simulation. Beginning at the top of Figure 4, the legacy PF-10 is not very sensitive and is narrow in applicability (2 SD at a reliability > 0.90). The Legacy HAQ is more sensitive (3 SD at a reliability > 0.95), but is quite weak in the range of normal function. The PROMIS HAQ is similarly limited in assessment of normal individuals but is a little better than the Legacy HAQ (3.3 SD > 0.95) in patients with more normal function. The PROMIS 20-item short form (4.8 SD > 0.95) is much more broadly applicable than the PROMIS 10-item short form, the Legacy HAQ, and the PROMIS HAQ. The 10-item simulated CAT (4.7 SD > 0.95) is superior to the other instruments on the basis of both information content across the severity spectrum and breadth of applicability.

## DISCUSSION

These studies show that new instruments constructed with qualitative item improvement and modern quantitative IRT assessment outperform Legacy instruments. The PROMIS HAQ, which was evolved from the Legacy HAQ, slightly outperformed its parent, and since it so closely correlates with its parent, it may be considered well-validated and ready for clinical trial use. Therefore, the PROMIS HAQ is already considered in the public domain (available from http://ARAMIS.Stanford.edu). The PROMIS 20-item short form, which uses the 20 best items, is undergoing validity testing, including sensitivity to change. It may be expected to replace the

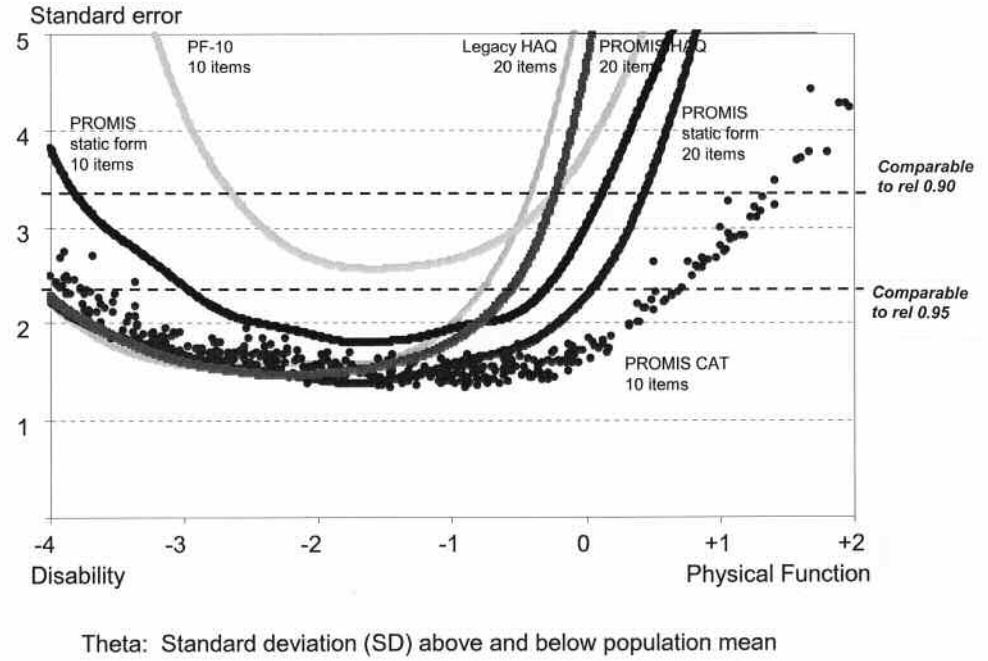


*Figure 4.* Comparison of information content of 6 physical function instruments: Legacy HAQ, PROMIS HAQ, PROMIS 10-item short form, PROMIS 20-item short form, PF-10, and 10-item PROMIS CAT. Instruments with greater information content have curves that are lower and have a greater SD range at a reliability > 0.95. More items are better than fewer, IRT-based (PROMIS) is better than non-IRT-based (Legacy). CAT is better than static.

PROMIS HAQ as validity is established, although some characteristics, such as use of "aids and devices" questions, need to be further explored. The PROMIS 10-item short form should find little use except where a high premium is placed on brevity, such as inclusion in a large-scale population survey. The PROMIS physical function CAT may be expected to be a major improvement on the instruments described here. CAT applications using item banks larger than 20 items and stopping rules at 10 or more items would be likely to outperform the tools described here.

Of interest, the Legacy HAQ's scoring algorithm approximates a crude CAT approach, which may help account for its historical effectiveness. By averaging the highest score in 8 categories rather than using a single average of 20 items, the score is dependent upon only 8 items out of 20, and these items are different in different patients. Use of only these "most abnormal" items raises the average scores and decreases the number of zero values, increasing sensitivity, and at the same time tailoring the item selection to the problem areas of the patient.

Among static instruments, those with more items are likely to outperform shorter instruments. A static form with items selected for IRT characteristics is likely to outperform instruments of the same length (e.g., Legacy instruments) constructed without this knowledge. CAT-based assessment offers superior performance over static forms of the same or even greater length. Refined CAT approaches that use differing stopping rules (e.g., 20 items) or larger original item pools may yield even greater gains. Documentation of responsiveness (sensitivity to change) in longitudinal studies is required to confirm these results.

We recommend that the PROMIS HAQ be considered for any use where the Legacy HAQ is now used. The PROMIS 20-item short form and the PROMIS CAT should be considered as secondary endpoints in these same studies through 2009, when validation studies will be completed. If these are as positive as expected, the PROMIS 20-item short form and PROMIS CAT may be considered as primary endpoints in 2010 and beyond.

## REFERENCES


1. Embretson SE, Reise SP. Item response theory for psychologists. London: Lawrence Erlbaum Associates; 2000.
2. Cella D, Chang CH. A discussion of item response theory and its applications in health status assessment. Med Care 2000;38 Suppl 9:1166-72.
3. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a computer-adaptive test for depression (D-CAT). Qual Life Res 2005;14:2277-91.
4. Steinbrocker O, Traeger CH, Batterman RC. Therapeutic criteria in rheumatoid arthritis. JAMA 1949;140:659-62.
5. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales. J Rheumatol 1982;9:789-93.
6. Meenan RF. The AIMS approach to health status measurement: conceptual background and measurement properties. J Rheumatol 1982;9:785-88.
7. Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 health survey, manual and interpretation guide. Boston: Health Institute, New England Medical Center; 1993.
8. Cella D, Yount S, Rothrock N, et al. The Patient Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. Med Care 2007;45:S3-S11.
9. Bruce B, Fries JF, Ambrosini D, Lingala B, Rose M, Ware JE. Better assessment of physical function: item improvement is neglected but essential. Arthritis Rheum 2006;54:5534.
10. DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45 Suppl 1:S12-21.
11. 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;61:17-33.
12. Van der Linden W, Hambleton R. Handbook of modern item response theory. New York: Springer; 1997.
13. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
14. Fries JF, Bruce B, Rose M. Improving the Health Assessment Questionnaire: the PROMIS HAQ [abstract]. Arthritis Rheum 2007;56 Suppl:S599.
15. Bruce B, Fries J. The Health Assessment Questionnaire (HAQ). Clin Exp Rheumatol 2005;23 Suppl 39:S14-18.
16. Bruce B, Fries JF. The Arthritis, Rheumatism and Aging Medical Information System (ARAMIS): still young at 30 years. Clin Exp Rheumatol 2005;23 Suppl 39:S163-7.
17. Hubert HB, Bloch DA, Oehlert JW, Fries JF. Lifestyle habits and compression of morbidity. J Gerontol A Biol Sci Med Sci 2002;57:M347-51.
18. Fries JF, Singh G, Morfeld D, Hubert HB, Lane NE, Brown BW Jr. Running and the development of disability with age. Ann Intern Med 1994;121:502-9.
19. Bjorner JB, Kosinski M, Ware JE Jr. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the Headache Impact Test (HIT). Qual Life Res 2003;12:913-33.
20. Reeve BB, Hays RD, Bjorner JB, 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 Suppl 1:S22-31.
21. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol 2005;23 Suppl 39:S53-7.