ARTHRITIS & RHEUMATOLOGY
Vol. 66, No. 5, May 2014, pp 1378–1387
DOI 10.1002/art.38342

Extending the Floor and the Ceiling for Assessment of Physical Function

James F. Fries,[1] Bharathi Lingala,[1] Liseth Siemons,[2] Cees A. W. Glas,[3] David Cella,[4] Yusra N. Hussain,[1] Bonnie Bruce,[1] and Eswar Krishnan[1]

Objective. **To improve the assessment of physical function by enhancing precision of physical function assessment as it pertains to subjects at extreme ends of the health continuum (i.e., subjects with extremely poor function ["floor"] or extremely good health ["ceiling"]).**

Methods. **Under the Patient-Reported Outcomes Measurement Information System (PROMIS) (a National Institutes of Health initiative), we developed new items to assess floor and ceiling physical function in order to supplement the existing item bank. Using item response theory and standard PROMIS methodology, we developed 31 floor items and 31 ceiling items and administered the items during a 12-month prospective, observational study of 737 subjects whose health status was at either extreme. Effect size was calculated and change over time was compared across anchor instruments and across items. Using the observed changes in scores, we back-calculated sample size requirements for the new and comparison measures.**

Results. **We studied 444 subjects who had been diagnosed as having a chronic illness and/or were of old age and 293 generally fit subjects (including athletes in training). Item response theory analyses confirmed that the new floor and ceiling items outperformed reference items ($P < 0.001$). The estimated post hoc sample size requirements were reduced by a factor of 2–4 for the floor population and a factor of 2 for the ceiling population.**

Conclusion. **Extending the range of items by which physical function is measured can substantially improve measurement quality, reduce sample size requirements, and improve research efficiency. The paradigm shift from assessing disability to assessing physical function focuses assessment on the entire spectrum of physical function, signals improvement in the conceptual base of outcome assessment, and may be transformative as medical goals more closely approach societal goals for health.**

In recent decades, quantitative assessment of functional disability has greatly benefited the study of chronic diseases and their treatments. Over time, more complex and precise measures have been developed, including the Health Assessment Questionnaire (HAQ) disability index (DI) (1) and the Short Form 36 (SF-36) physical functioning (PF) domain (2), which measure functional ability through patient-reported outcomes. These instruments, although more quantitative than their predecessors and useful in clinical trials and observational studies (3–5), have disadvantages. "Health," as defined by the World Health Organization (WHO), is "not merely the absence of disease, but complete physical, psychological, and social well-being" (6); this definition indicates the need not only to measure impairments that are worse than those experienced by the population on average, but also to measure the functional status of subjects whose health is above average. Thus, the "disability" domain is best redefined as "physical function" (7,8), since the focus has shifted

Supported by the NIH (grant 2U01-AR-052158 to the Patient-Reported Outcomes Measurement Information System [PROMIS] at Stanford University [principal investigator: Dr. Fries]). PROMIS II was supported by the NIH at several centers and research sites (grants 1U54-AR-057951, 1U54-AR-057943, 1U54-AR-057926, 1U01-AR-057948, 1U01-AR-057954, 1U01-AR-052171, 2U01-AR-052181, 1U01-AR-057956, 1U01-AR-057929, 1U01-AR-057936, 2U01-AR-052155, U01-AR-057971, 1U01-AR-057940, 1U01-AR-057967, and 2U01-AR-052186).

[1]James F. Fries, MD, Bharathi Lingala, PhD, Yusra N. Hussain, MD, Bonnie Bruce, DrPH, MPH, RD, Eswar Krishnan, MD, MPhil: Stanford University, Palo Alto, California; [2]Liseth Siemons, MSc: Stanford University, Palo Alto, California, and University of Twente, Enschede, The Netherlands; [3]Cees A. W. Glas, PhD: University of Twente, Enschede, The Netherlands; [4]David Cella, PhD: Northwestern University, Chicago, Illinois.

Address correspondence to Eswar Krishnan, MD, MPhil, Stanford University, ARAMIS Program, 1000 Welch Road, Suite 203, Palo Alto, CA 94304. E-mail: e.krishnan@stanford.edu.

Submitted for publication June 5, 2013; accepted in revised form December 26, 2013.

from assessment of disability to assessment of physical function.

In the first cycle of the Patient-Reported Outcomes Measurement Information System (PROMIS) (a National Institutes of Health [NIH] initiative), we developed a core physical function item bank called PROMIS PF (n = 154 items), which contained 124 new items, 10 legacy items from the physical functioning 10-item scale (PF-10) of the SF-36 health survey, and 20 items from the HAQ (9). The new items and instruments derived from previously established instruments outperformed the previously established instruments in terms of efficiency and precision, allowing studies to be performed with smaller sample sizes (10,11).

Although it represented an advancement, some limitations associated with the legacy items persisted in the PROMIS PF item bank, the most important being insensitivity to changes at the extreme ranges of physical function. This meant that large changes in true physical function among the frail and the robust were necessary before these changes were reflected in the physical function metric (8). For example, in a prospective, observational study of 6,436 patients with rheumatoid arthritis who were longitudinally followed up for 32,324 person-years (providing 64,647 HAQ DI measurements, with an average of 19 measurements per person), 10% of the patients scored 0, signifying no disability (12). Thus, subtle but clinically critical changes to patient health were not documented, limiting the use of the item bank in the broader population. Further, in longitudinal studies (including clinical trials), potentially important changes in physical function among subjects whose physical function was at an extreme may be missed.

Assessment of subjects on the extreme ends of the physical function spectrum requires a sufficient number of validated items for each extreme, testing among a sufficient number of subjects whose health status is at a functional extreme, and a broader measurement metric to provide stable estimates. This report describes the development of new physical function items related to extremely poor function ("floor") and extremely good health ("ceiling") using item response theory and standard PROMIS methodology to supplement the PROMIS PF item bank (PROMIS II). We applied these items in a prospective observational study and found the following: 1) the addition of floor and ceiling items to existing core physical function item banks increased the statistical power of the research across the full spectrum of human ability, 2) use of these measures enabled more precise study of subjects whose health was at the extreme ends of physical function (such as institutionalized patients or trained athletes), and 3) use of the new physical function item bank required smaller sample sizes than were necessary with previous instruments to achieve a given level of precision.

SUBJECTS AND METHODS

Theoretical framework. Stanford University has been a primary research site of PROMIS over the past 9 years, with a particular focus on improving assessment of physical function (9). PROMIS is an NIH Roadmap Program (online at http://nihpromis.org) tasked with improving the infrastructure of clinical research by using item response theory (13–19) and computerized adaptive testing (CAT) (11,19–22). Members of PROMIS II are listed in Appendix A. Item response theory allows improved measurement through selection and optimization of the best available items and aggregation of these items to develop better instruments. The PROMIS approach to item bank development includes identification of candidate items, item improvement, and qualitative item evaluation, as well as study of clarity, translatability, importance, and quantitative validation.

Items and instruments. Information on the existing PROMIS instruments has been reported in detail elsewhere (1–3,9–11). Candidate floor and ceiling items were submitted and evaluated by item content experts, using modified Delphi methods, and then reviewed using previously described processes (15,21). The PROMIS protocol used for developing a set of psychometrically optimal items has been previously described (11), and qualitative evaluation of these new floor and ceiling items has been reported (8). We found that items indicating tasks that were easier or harder than the tasks indicated by existing core items could be constructed, understood by subjects, and efficiently administered and scored. The level of difficulty was increased by a factor of 5 for the ceiling items (with the addition of 31 new items describing tasks of higher difficulty) and reduced by a factor of 4 for the floor items (with the addition of 31 new floor items describing tasks of lesser difficulty). Tables 1 and 2 list the newly developed items, the PROMIS SF-20 items (which in turn were designed to be similar to the HAQ), and the legacy PF-10 items derived from the SF-36. Since the items were developed using item response theory, we were able to aggregate individual items to create new instruments, and then subject these new instruments to further analyses. In the present study, we included the raw scores for statistical clarity, although the PROMIS convention is to present physical function scores only in terms of T statistic distribution with a mean ± SD of 50 ± 10.

Item response theory analyses. Item and test information curves were examined to determine whether the 10 best new floor and ceiling items did indeed assess extremes on the physical function scale more precisely than the legacy PF-10 items. We examined whether the inclusion of the floor and ceiling items expanded the breadth of the physical function assessment. Item information curves can show the contribution of individual items to the measurement of physical function. Test information curves can demonstrate the range of physical function in which such measurement is reliable. Analyses were

Table 1. Baseline, 12-month, and change scores in the "floor" population (subjects with extremely poor health)*

	Baseline score, mean ± SD	12-month followup score, mean ± SD	12-month change score, mean ± SD	Effect size†
New floor items				
Hold a card or letter in order to read it	0.3 ± 0.7	0.4 ± 0.8	0.1 ± 0.7	0.12
Squeeze another person's hand	0.6 ± 0.9	0.7 ± 1.0	0.1 ± 0.7	0.11
Cut your toenails	2.5 ± 1.4	2.6 ± 1.4	0.1 ± 0.8	0.10
Pour liquid into a cup	0.5 ± 0.9	0.6 ± 1.0	0.1 ± 0.7	0.10
Dial a number on the keypad of a cell phone	0.7 ± 1.1	0.8 ± 1.1	0.1 ± 0.8	0.09
Put on a sweater or t-shirt over your head	0.8 ± 1.1	0.9 ± 1.1	0.1 ± 0.8	0.09
Write a simple sentence using a pen or pencil	0.6 ± 1.0	0.7 ± 1.0	0.1 ± 0.7	0.09
Move from sitting on the bed to lying down	0.7 ± 1.0	0.8 ± 1.0	0.1 ± 0.7	0.09
Move about in a dark room or hallway without falling	1.3 ± 1.3	1.4 ± 1.3	0.1 ± 0.9	0.08
Type a sentence on a computer keyboard	1.0 ± 1.3	1.1 ± 1.4	0.1 ± 0.9	0.08
Turn pages in a book	0.3 ± 0.6	0.3 ± 0.6	0.0 ± 0.5	0.07
Loosen a screw using a manual screwdriver	1.1 ± 1.3	1.2 ± 1.3	0.1 ± 1.0	0.07
Get items in and out of a wallet	0.6 ± 0.9	0.7 ± 0.9	0.1 ± 0.7	0.07
Fasten buttons on a shirt or blouse	1.1 ± 1.2	1.2 ± 1.2	0.1 ± 0.8	0.07
Use a knife and fork	0.5 ± 0.9	0.6 ± 1.0	0.1 ± 0.7	0.06
Chew and eat your food as quickly as 5 years ago	1.0 ± 1.2	1.0 ± 1.3	0.1 ± 1.0	0.06
What is the farthest distance you can walk by yourself	1.9 ± 1.3	2.0 ± 1.3	0.1 ± 0.7	0.05
In the past year, how many times did you fall	0.7 ± 0.8	0.7 ± 0.8	0.0 ± 0.7	0.05
Put on your shoes	0.9 ± 1.1	1.0 ± 1.2	0.1 ± 0.8	0.05
Dress yourself in <10 minutes	1.4 ± 1.3	1.4 ± 1.4	0.1 ± 1.0	0.05
Do you feel exhausted	1.7 ± 0.9	1.7 ± 0.9	0.0 ± 0.8	0.04
Dress and groom yourself as quickly as 5 years ago	2.1 ± 1.4	2.2 ± 1.4	0.0 ± 1.1	0.03
Take a letter out of an envelope	0.3 ± 0.7	0.3 ± 0.7	0.0 ± 0.5	0.03
Put on your socks	1.1 ± 1.2	1.1 ± 1.2	0.0 ± 0.8	0.02
Move from street to sidewalk without a curb cut	1.2 ± 1.2	1.3 ± 1.2	0.0 ± 0.8	0.02
Push the buttons on a television remote control	0.3 ± 0.8	0.3 ± 0.7	0.0 ± 0.6	0.02
Move about your residence	0.7 ± 1.0	0.7 ± 1.0	0.0 ± 0.7	0.02
Compared to 5 years ago, what is your normal walking speed	2.4 ± 0.8	2.4 ± 0.8	0.0 ± 0.6	0.02
In the past year, amount of unintentional weight loss	0.6 ± 0.9	0.6 ± 0.9	0.0 ± 1.0	0.01
Walk a block as quickly as you did 5 years ago	2.8 ± 1.3	2.8 ± 1.3	0.0 ± 1.0	0.01
Walk up or down inclines	1.5 ± 1.2	1.5 ± 1.2	0.0 ± 0.9	0.00
PROMIS SF-20 items				
Wash and dry your body	0.8 ± 1.1	0.9 ± 1.2	0.1 ± 0.8	0.11
Hold a plate full of food	0.6 ± 1.0	0.7 ± 1.1	0.1 ± 0.8	0.10
Push open a heavy door	1.7 ± 1.2	1.8 ± 1.2	0.1 ± 0.9	0.10
Doing two hours of physical labor	3.0 ± 1.2	3.1 ± 1.1	0.1 ± 0.9	0.10
Shampoo your hair	0.8 ± 1.3	0.9 ± 1.3	0.1 ± 0.8	0.08
Do chores like vacuuming or yard work	2.2 ± 1.4	2.3 ± 1.3	0.1 ± 1.0	0.07
Sit on the edge of a bed	0.3 ± 0.8	0.4 ± 0.8	0.1 ± 0.6	0.07
Lifting or carrying groceries	1.7 ± 1.3	1.8 ± 1.3	0.1 ± 0.9	0.07
Vigorous activities, like running, lifting heavy objects	3.4 ± 1.0	3.4 ± 0.9	0.1 ± 0.8	0.06
Dress yourself	1.0 ± 1.2	1.1 ± 1.2	0.1 ± 0.8	0.06
Dry your back with a towel	0.9 ± 1.2	0.9 ± 1.2	0.0 ± 0.8	0.04
Get in/out of a car	1.1 ± 1.0	1.0 ± 0.9	0.0 ± 0.7	0.03
Walking more than a mile	2.9 ± 1.3	2.9 ± 1.4	0.0 ± 0.9	0.03
Squeeze a new tube of toothpaste	0.4 ± 0.9	0.5 ± 0.8	0.0 ± 0.7	0.03
Bending, kneeling, or stooping	2.3 ± 1.1	2.3 ± 1.1	0.0 ± 1.0	0.03
Climbing one flight of stairs	1.9 ± 1.3	1.9 ± 1.4	0.0 ± 0.9	0.02
Run a short distance to catch a bus	2.8 ± 1.4	2.8 ± 1.4	0.0 ± 1.0	0.01
Transfer from a bed to a chair and back	0.7 ± 1.0	0.7 ± 1.0	0.0 ± 0.7	0.01
Wash your back	1.5 ± 1.3	1.6 ± 1.2	0.0 ± 0.9	0.01
Get on/off a toilet	0.8 ± 1.0	0.8 ± 0.9	0.0 ± 0.7	0.00
Legacy PF-10 items				
Climb one flight of stairs	1.0 ± 0.8	0.9 ± 0.8	0.0 ± 0.7	0.05
Walking more than a mile	1.5 ± 0.7	1.5 ± 0.7	0.0 ± 0.7	0.04
Vigorous activities, like running or strenuous sports	1.8 ± 0.5	1.8 ± 0.5	0.0 ± 0.6	0.04
Bending, kneeling, or stooping	1.2 ± 0.7	1.2 ± 0.7	0.0 ± 0.7	0.03
Climbing several flights of stairs	1.4 ± 0.7	1.4 ± 0.7	0.0 ± 0.6	0.02
Lifting or carrying groceries	1.0 ± 0.7	0.9 ± 0.7	0.0 ± 0.6	0.02
Walking several hundred yards	1.0 ± 0.8	1.0 ± 0.8	0.0 ± 0.7	0.01

Table 1. *(Cont'd)*

	Baseline score, mean ± SD	12-month followup score, mean ± SD	12-month change score, mean ± SD	Effect size†
Walking 100 yards	0.8 ± 0.8	0.8 ± 0.8	0.0 ± 0.6	0.01
Bathing and dressing yourself	0.5 ± 0.7	0.5 ± 0.7	0.0 ± 0.6	0.00
Moderate activities, like moving a table or golf	1.3 ± 0.7	1.3 ± 0.7	0.0 ± 0.6	0.00

* SF = short form; PF = physical function.
† The effect size is of individual items and was calculated as the mean change divided by the pooled SD.

performed with the Multidimensional Item Response Theory statistical software program (23), using a multidimensional generalized partial credit model for repeated measures (24) and taking into account the dependency between responses at baseline and at 12 months. Since the same tests were administered at both time points, item parameters were kept constant. Parameters were estimated using the marginal maximum likelihood method.

Study subjects and item administration. The Stanford Institutional Review Board approved the study, and all subjects provided written informed consent. This study required not only that items be applicable to subjects at extreme ends of the physical functioning scale, but also that subjects who were sufficiently disabled or sufficiently healthy be included in order to accurately assess the new items. We did not administer floor items to ceiling populations or vice versa, since asking a marathon runner whether he or she can squeeze a person's hand or a nursing home resident whether he or she can run 5 miles was considered potentially offensive. Since not all impaired persons or fit persons are the same, we sought diversity among subjects to improve the generalizability of findings. This 12-month longitudinal study measured the sensitivity of the items and instruments to detect changes in functional abilities over time. The intervention, therefore, was time, based on the consistent observation that declines in physical function invariably occur as diseases progress and subjects age (11,25).

Floor population. Floor items were administered to 444 subjects known to have poor functional status. Sixty-five percent of the subjects were women; the average age was 86 years. More than 90% of the subjects were white, and they had an average of 16 years of education. The average raw score at baseline on the PROMIS SF-20 (the PROMIS successor to the HAQ DI) was 38.5 on a scale of 0–100, where 0 is highly able and 100 is severely disabled. Sixteen of the subjects were nursing home residents interviewed in their home, 206 were participating in longitudinal studies of aging, and 222 had either moderate-to-severe rheumatoid arthritis or osteoarthritis. The average baseline raw score was similar to scores attained by patients with severe rheumatoid arthritis or severe osteoarthritis (1,3,26). Participants were administered the candidate floor items (n = 31), the PROMIS SF-20, and the legacy PF-10 (61 total items) (Table 1) at baseline and at 12 months.

Ceiling population. Ceiling items were administered to 293 subjects (107 vigorous exercisers [including ultra-marathon runners], 147 apparently healthy seniors, and 39 patients with no more than mild symptoms of arthritis). Sixty percent of these subjects were men; the average age was 60 years. More than 85% of the subjects were white, and they had an average of 17 years of education. The average raw score on the PROMIS SF-20 was 3.4, indicating excellent physical function. These subjects completed the items in the PROMIS SF-20 and the legacy PF-10 as well as the 31 candidate ceiling items at baseline and at 12 months (Table 2). The new ceiling items included difficult tasks such as "do 8 hours of physical labor," "run 10 miles," and "climb 15 flights of stairs."

Measurement of change over time. Cohen's effect size was calculated for individual items and for the instruments as a whole in order to determine whether the expanded item range increased or decreased the effect size of simulated instruments in populations of subjects whose health status was at an extreme end of the spectrum as well as in broader populations containing both subjects of average health and subjects whose health was on the extreme end of the spectrum. The term "simulated" is used because all 61 items were administered together, while reference instruments are generally composed of 10 or 20 items. Effect sizes calculated for the individual items were used to select the 10 and 20 best individual floor and ceiling items. We did not further study items that had an effect size close to 0, since no improvement was expected with their addition to the item bank.

RESULTS

Difficulty of tasks indicated by new floor and new ceiling items at baseline. Using data from baseline administration of the items, we performed a cross-sectional comparison of the difficulty of tasks described by the new items and the reference items in both the floor and ceiling groups to confirm that the tasks indicated by the new items were indeed easier or more difficult than those indicated by the reference items. All comparisons revealed a clear and consistent separation of old core items from new items, both among the floor population and the ceiling population (Tables 1 and 2 and Figure 1).

Item-level effect sizes for the floor population. Table 1 shows item-level mean ± SD baseline scores and scores at 12 months as well as the 12-month changes in scores in subjects with poor functional status. Item-level effect sizes are also shown, calculated as the change

Table 2. Baseline, 12-month, and change scores in the "ceiling" population (subjects with extremely good health)*

	Baseline score, mean ± SD	12-month followup score, mean ± SD	12-month change score, mean ± SD	Effect size†
New ceiling items				
Climb 10 flights of stairs (40 steps)	0.5 ± 0.9	0.8 ± 1.1	0.2 ± 0.8	0.25
Climb 15 flights of stairs (60 steps)	0.8 ± 1.1	1.0 ± 1.2	0.3 ± 0.8	0.23
Climb 5 flights of stairs (20 steps)	0.2 ± 0.6	0.4 ± 0.9	0.1 ± 0.6	0.20
Exercise hard for half an hour	0.4 ± 0.8	0.5 ± 0.9	0.1 ± 0.7	0.15
Climb a ladder to trim a tree	0.3 ± 0.7	0.4 ± 1.0	0.1 ± 0.6	0.15
Doing heavy work around the house	0.5 ± 0.8	0.6 ± 1.0	0.1 ± 0.7	0.14
Paint a room	0.3 ± 0.7	0.4 ± 0.9	0.1 ± 0.7	0.14
Row a rowboat	0.5 ± 0.9	0.6 ± 1.0	0.1 ± 0.7	0.14
Past week, total time on vigorous physical activity	0.7 ± 1.0	0.8 ± 1.2	0.1 ± 0.9	0.13
Doing 8 hours of physical labor	0.9 ± 1.1	1.0 ± 1.2	0.1 ± 0.7	0.12
Take a 20-minute brisk walk, without stopping to rest	0.1 ± 0.5	0.2 ± 0.7	0.1 ± 0.7	0.12
Trim a hedge	0.2 ± 0.7	0.3 ± 0.8	0.1 ± 0.6	0.11
Shovel fresh snow and clear 30 feet off driveway	0.6 ± 1.0	0.8 ± 1.2	0.1 ± 0.8	0.11
Transfer a full load of clothes from a washer to dryer	0.0 ± 0.1	0.0 ± 0.3	0.0 ± 0.3	0.11
What is your best time for running one mile now	1.8 ± 1.1	1.9 ± 1.1	0.1 ± 0.7	0.10
Strenuous activities such as backpacking, skiing, tennis	0.8 ± 1.1	0.9 ± 1.2	0.1 ± 0.8	0.10
Hand wash and wax a car	0.2 ± 0.6	0.3 ± 0.7	0.1 ± 0.5	0.10
Exercise for an hour	0.4 ± 0.8	0.4 ± 0.9	0.1 ± 0.6	0.10
Hang a heavy painting or picture on your wall	0.5 ± 0.8	0.6 ± 1.0	0.1 ± 0.8	0.09
Push and move an empty refrigerator	0.8 ± 1.1	0.7 ± 1.1	−0.1 ± 0.8	−0.08
Climb 1,000 vertical feet on a trail in an hour	0.6 ± 1.0	0.7 ± 1.0	0.1 ± 0.7	0.07
Dig a hole in the dirt with a shovel	0.4 ± 0.8	0.4 ± 0.9	0.1 ± 0.6	0.07
Run or jog slowly for 2 miles	0.9 ± 1.5	1.0 ± 1.4	0.1 ± 0.7	0.06
Run at a fast pace for 2 miles	1.4 ± 1.5	1.5 ± 1.5	0.1 ± 0.9	0.05
Past week, how many times vigorous physical activity	1.5 ± 1.2	1.6 ± 1.1	0.1 ± 0.9	0.05
Push a car in neutral gear	0.8 ± 1.1	0.8 ± 1.1	−0.1 ± 0.9	0.05
Change a flat tire	0.8 ± 1.2	0.8 ± 1.2	0.1 ± 0.7	0.04
Run 5 miles	1.4 ± 1.7	1.5 ± 1.7	0.1 ± 0.8	0.04
Move a full garbage/recycle bin	0.2 ± 0.6	0.2 ± 0.7	0.0 ± 0.6	0.04
How many minutes does it take for you to walk one mile	0.7 ± 0.8	0.8 ± 0.9	0.0 ± 0.7	0.02
Run 10 miles	1.8 ± 1.7	1.8 ± 1.7	0.0 ± 0.8	0.00
PROMIS SF-20 items				
Lifting or carrying groceries	0.1 ± 0.3	0.1 ± 0.4	0.1 ± 0.3	0.18
Wash your back	0.2 ± 0.5	0.1 ± 0.3	0.1 ± 0.5	0.15
Vigorous activities, like running or strenuous sports	0.8 ± 1.1	0.9 ± 1.2	0.1 ± 0.9	0.12
Climbing one flight of stairs	0.1 ± 0.3	0.1 ± 0.4	0.0 ± 0.4	0.11
Walking more than a mile	0.2 ± 0.6	0.2 ± 0.8	0.1 ± 0.6	0.09
Bending, kneeling, or stooping	0.4 ± 0.7	0.4 ± 0.8	0.1 ± 0.6	0.08
Push open a heavy door	0.1 ± 0.4	0.1 ± 0.5	0.0 ± 0.4	0.07
Chores such as vacuuming or yard work	0.2 ± 0.5	0.2 ± 0.6	0.0 ± 0.5	0.05
Doing 2 hours of physical labor	0.4 ± 0.8	0.4 ± 0.9	0.0 ± 0.6	0.05
Dry your back with a towel	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.3	0.04
Sit on the edge of a bed	0.0 ± 0.2	0.0 ± 0.2	0.0 ± 0.3	0.03
Get on/off a toilet	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.3	0.03
Transfer from a bed to a chair and back	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.3	0.03
Run a short distance to catch a bus	0.2 ± 0.7	0.3 ± 0.7	0.0 ± 0.7	0.02
Squeeze a new tube of toothpaste	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.3	0.02
Get in/out of a car	0.1 ± 0.3	0.1 ± 0.3	0.0 ± 0.4	0.01
Hold a plate full of food	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.3	0.01
Shampoo your hair	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.3	0.01
Dress yourself	0.1 ± 0.3	0.1 ± 0.3	0.0 ± 0.3	0.00
Wash and dry your body	0.0 ± 0.3	0.0 ± 0.2	0.0 ± 0.3	0.00
Legacy PF-10 items				
Vigorous activities, like running or strenuous sports	0.4 ± 0.6	0.6 ± 0.7	0.2 ± 0.6	0.24
Climbing one flight of stairs	0.0 ± 0.2	0.1 ± 0.3	0.0 ± 0.3	0.17
Bathing and dressing yourself	0.0 ± 0.1	0.0 ± 0.2	0.0 ± 0.2	0.14
Walking 100 yards	0.0 ± 0.1	0.1 ± 0.3	0.0 ± 0.2	0.14
Walking more than a mile	0.1 ± 0.3	0.1 ± 0.4	0.0 ± 0.3	0.13
Climbing several flights of stairs	0.1 ± 0.4	0.2 ± 0.4	0.0 ± 0.4	0.13
Moderate activities, like moving a table or golf	0.1 ± 0.4	0.1 ± 0.4	0.0 ± 0.4	0.11

Table 2. *(Cont'd)*

	Baseline score, mean ± SD	12-month followup score, mean ± SD	12-month change score, mean ± SD	Effect size†
Lifting or carrying groceries	0.0 ± 0.2	0.1 ± 0.3	0.0 ± 0.3	0.10
Walking several hundred yards	0.0 ± 0.2	0.1 ± 0.3	0.0 ± 0.3	0.04
Bending, kneeling, or stooping	0.2 ± 0.4	0.2 ± 0.5	0.0 ± 0.4	0.04

* SF = short form; PF = physical function.
† The effect size is of individual items and was calculated as the mean change divided by the pooled SD.

between mean baseline scores and mean final scores divided by the pooled standard deviation of each item. Within each instrument (the new item bank, the PROMIS SF-20, and the legacy PF-10), items were sorted by effect sizes. These items were likely to make the largest contributions to the effect sizes of the individual instruments. The direction of change for the majority of items showed decreases over time, with only a few items with small effect sizes suggesting improvement. Among the new items, the 20 items with the highest effect sizes were selected. These effect sizes ranged from 0.12 to 0.05, which are small effect sizes for an instrument, but are considerable for a single item. Of the 31 tested floor items, the effect sizes for the 10 items that had the lowest effect sizes ranged from 0.05 to 0.0 and the direction of change was opposite for 4 of these 10 items (which were also excluded from further analyses because of inconsistencies). Item-level effect sizes for the PROMIS SF-20 items ranged from 0.11 to 0.0. For the legacy PF-10, effect sizes ranged from 0.05 to 0.0.

Instrument-level effect sizes for the floor population. Six instruments were simulated from floor population data. These included the legacy PF-10, the PROMIS SF-10 (a subset of the PROMIS SF-20), and instruments compiled from the 10 floor items with the largest effect sizes, the 20 floor items with the largest effect sizes, and the 30 floor items with the largest effect sizes. For further comparison, we simulated a 20-item instrument composed of the PROMIS SF-10 and the 10 new floor items with the highest effect sizes. Mean ± SD scores, *P* values (calculated by pairwise *t*-test), the standardized response mean (SRM), Cohen's effect size, minimum detectable difference, and sample size requirements were determined for each of the simulated instruments (Table 3).

For the simulated instruments, differences between baseline and final scores were statistically significant, except for the legacy PF-10, which also had a different direction of change. There are several plausible explanations for this finding: the legacy PF-10 has only

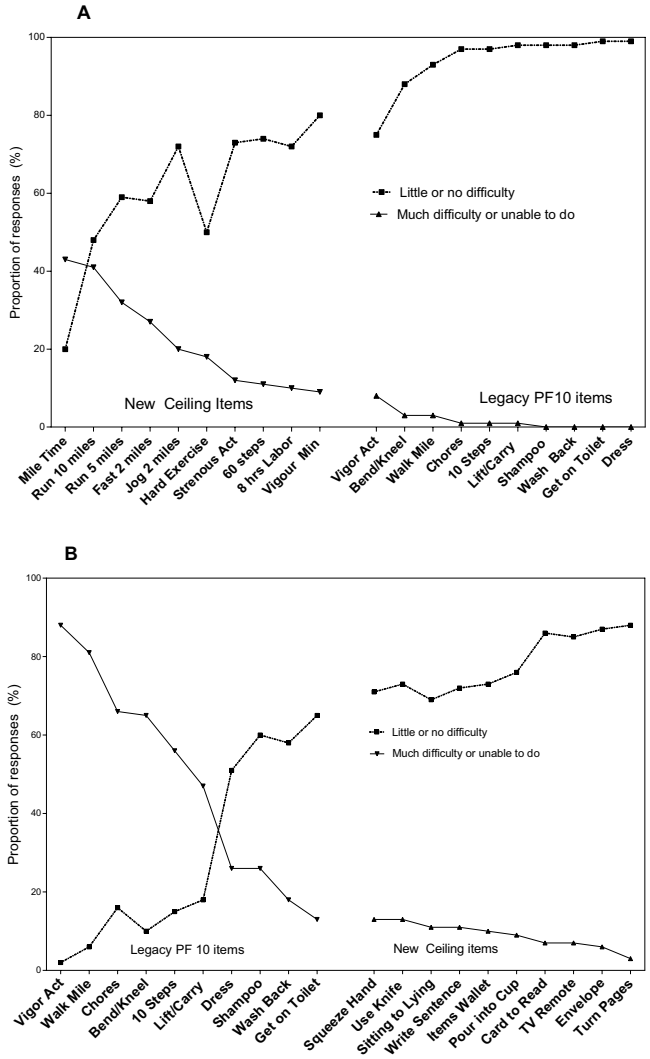

Figure 1. Level of difficulty of tasks indicated by newly developed physical function (PF) items (measuring extremely good health function ["ceiling"] and extremely poor health function ["floor"]) and items in the legacy PF-10 item bank. **A,** Among the ceiling population, tasks indicated by the new items were found to be more difficult as compared to tasks indicated by legacy PF-10 items. **B,** Among the floor population, tasks indicated by the new items were found to be considerably easier than those indicated by the legacy PF-10 items.

Table 3. Physical function instruments, metric raw scores, and sample size requirements for "floor" items (measuring extremely poor function) and "ceiling" items (measuring extremely good function)*

	Baseline score, mean ± SD	Followup score, mean ± SD	Change score, mean ± SD	*P*	SRM	Cohen's effect size	Minimum detectable difference†	Sample size requirement‡
Instruments tested in the floor population								
Legacy PF-10	57.6 ± 26	56.9 ± 27	−0.7 ± 19	0.46	0.04	0.03	2.52	451
PROMIS SF-10	44.3 ± 23	45.8 ± 23	1.5 ± 12	0.01	0.13	0.07	1.55	171
PROMIS SF-20	38.5 ± 21	39.7 ± 21	1.2 ± 10	0.01	0.12	0.06	1.37	135
Floor: top 10 items	22.1 ± 19	24.7 ± 20	2.6 ± 10	<0.001	0.26	0.13	1.35	130
Floor: top 20 items	22.8 ± 18	24.9 ± 19	2.1 ± 9.0	<0.001	0.23	0.11	1.23	108
Floor: top 30 items	26.7 ± 18	27.9 ± 18	1.2 ± 8.0	0.002	0.15	0.07	1.09	86
PROMIS SF-10 and floor top 10 items	33.3 ± 19	35.3 ± 20	2.0 ± 9.0	<0.001	0.22	0.10	1.21	106
Instruments tested in the ceiling population								
Legacy PF-10	5.2 ± 10	7.3 ± 15	2.1 ± 11	<0.001	0.19	0.17	1.81	154
PROMIS SF-10	4.2 ± 8.0	5.1 ± 10	0.9 ± 7.0	0.03	0.13	0.10	1.15	64
PROMIS SF-20	3.4 ± 7.0	3.9 ± 8.0	0.5 ± 7.0	0.19	0.08	0.06	1.07	55
Ceiling: top 10 items	12.5 ± 17	16.3 ± 20	3.9 ± 10	<0.001	0.39	0.21	1.64	128
Ceiling: top 20 items	13.1 ± 15	15.9 ± 18	2.8 ± 9.0	<0.001	0.31	0.16	1.46	101
Ceiling: top 30 items	16.9 ± 18	19.1 ± 20	2.1 ± 9.0	<0.001	0.24	0.11	1.45	99
PROMIS SF-10 and ceiling top 10 items	8.3 ± 11	10.7 ± 14	2.4 ± 8.0	<0.001	0.31	0.18	1.26	76

* While the Patient-Reported Outcomes Measurement Information System (PROMIS) convention is to express T scores with a mean of 50 (range 0–100), raw scores are provided for clarity. The standardized response mean (SRM) is the mean change of the score divided by the standard deviation of the change in score. Cohen's effect size is the mean change divided by the pooled SD. *P* values were calculated by pairwise *t*-test. PF = physical function; SF = short form.

† A sample size of 444 subjects in the floor population and 293 subjects in the ceiling population was used to calculate the minimum detectable difference at 80% power.

‡ Number of subjects needed to enable detection of a difference in the population means of 2.5 based on the observed standard deviation of the change.

10 items, there are only 3 response options (rather than 5), and the 10 items are known to be more sensitive toward the middle of a normal population than at the extremes. For all of the other instruments, statistically significant differences were observed between baseline and final values, and all of the statistics tested were consistent. The 20-item instruments outperformed the 10-item instruments, and the results gathered using the 30-item scale were roughly similar to the results gathered using the 20-item scales. With the PROMIS SF-10 and SF-20 item scales change was detected at a *P* value of 0.01, and the 10-item and 20-item floor instruments and the instrument composed of the PROMIS SF-10 and the top 10 new floor items enabled detection of change at $P < 0.001$. Sample size requirements in this floor population were reduced by a factor of 2–4 (from ~150 per arm in the PROMIS SF-10 and PROMIS SF-20 item instruments to ~115 in the instruments containing new floor items). Sample sizes were those required to reach a minimum detectable difference of 2.5%.

Item-level effect sizes for the ceiling population. Mean ± SD baseline scores, 12-month scores, and 12-month change scores in the ceiling population are shown in Table 2. With each of the 31 new ceiling items change in health status was detected after 12 months. The effect sizes of the change scores ranged from 0.25 to 0.0 (median 0.10). Using the same PROMIS SF-20 instrument that was administered to the floor population, all items predicted progression, with effect sizes of 0.18 to 0.00 (median ~0.05). All legacy PF-10 items also predicted progression, with item-level effect sizes ranging from 0.24 to 0.04 (median 0.13). These effect sizes are approximately double those seen in the new items that were designed to evaluate the floor population.

Instrument-level effect sizes for the ceiling population. As with the floor population, 6 instruments were simulated using data from the ceiling population. These included the legacy PF-10, the PROMIS SF-10, and the PROMIS SF-20 described above, as well as instruments composed of the 10 items with the largest effect sizes, the 20 items with the largest effect sizes, and the 30 items with the largest effect sizes. The legacy PF-10 had a lower *P* value than the PROMIS SF-10 and SF-20. This occurred because the legacy PF-10 performs at its best at

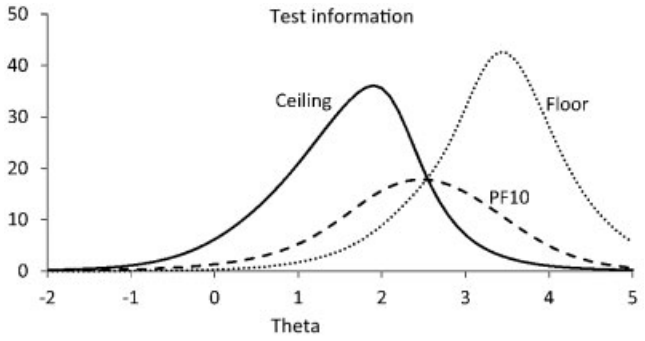

Figure 2. Baseline test information curves of the instruments containing the 10 best "floor" (extremely poor health function) items, the 10 items from the legacy physical functioning (PF-10) item bank, and the 10 best "ceiling" (extremely good health function) items. Theta scores (scaled around 0) correspond to the level of physical function, where higher scores represent worse function.

the population mean (ceiling), whereas the PROMIS instruments perform best in the range of moderate impairment (toward the floor). All of the PROMIS instruments have better psychometrics than the legacy PF-10, due in part to item improvement (which included an increase to 5 response options). There was less change over time in the ceiling population than in the floor population, and differences across instruments were harder to detect. Overall, the hybrid instrument (PROMIS SF-10 and top 10 ceiling items) outperformed other simulated instruments, with increased effect sizes and lower *P* values, and it required a lower sample size. All instrument results were adjusted to a minimum detectable difference of 2.5%.

Information curves and population distribution. Figure 1 summarizes the difficulty level of tasks indicated by the individual items that had the highest effect sizes among floor and ceiling populations. The percentage of subjects responding "little or no difficulty" and the percentage of subjects responding "much difficulty or unable to do" are shown. The results demonstrate that the new floor and ceiling items function better than the legacy PF-10 items in subjects whose health is at an extreme, expanding the measurement range (Figure 2). Figure 3 shows test information curves for the 10 best new ceiling items, the legacy PF-10 items, and the 10 best new floor items. The new items extend the range toward the floor by 1–2 SD and the range toward the ceiling by ~2 SD.

DISCUSSION

Quantitative assessment of functional disability has greatly contributed to the study of chronic diseases and their treatments. However, current assessments fail to properly discriminate between extremes at either end of the physical ability spectrum. We demonstrated that items that sensitively query abilities toward the extremes of physical function can be created and that the new items are reliable, sensitive, and valid. When the new floor and ceiling items were added to the item bank, the assessable range was substantially extended and the potential utility of the item bank was enhanced. We began with the 154 core PROMIS PF items. Addition of 20 new items that assess floor and ceiling function significantly improved precision. Adding the next 10 best floor and ceiling items resulted in slight additional improvement. The addition of more items (as the effect sizes neared 0) added little, and paradoxical effects and less consistency resulted. Therefore, an adequate core item bank for the study of a wide range of physical function might contain ~200 items.

With the optimized instrument, small discrepancies were observed between the population and instrument measurement range. The new floor items describe tasks that are located even more toward the extreme than the capabilities of the floor population, leaving room for the measurement of physical function levels in subjects whose health function is worse than that seen in our floor population. In contrast, our ceiling population functioned nearer the extreme than the range measured by the ceiling items. This indicates that a residual ceiling effect is present with the optimized instrument and that more work is needed in that regard.

The importance of addressing floor and ceiling

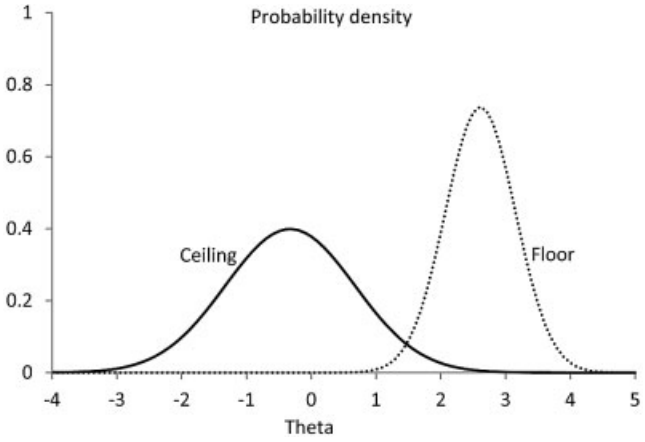

Figure 3. Frequency distributions of the "floor" (extremely poor health function) and "ceiling" (extremely good health function) populations in relation to the physical function continuum. Theta scores (scaled around 0) correspond to the level of physical function, where higher scores represent worse function.

effects is evident in the case of rheumatoid arthritis studies; ~10% of patients who clearly have age-related and disease-related limitations do not have measurable disability when assessed by the HAQ DI (12). This issue is further evident in a random sample of the general population in which 75% of those polled did not report any disability, and among men and women ages 30–55 years, a score of 0 on the HAQ DI placed them in the 75th percentile (25). Moreover, other subject groups cannot be adequately evaluated using current instruments. If health, as described by the WHO, is "not merely the absence of disease, but complete physical, social, and psychological well-being," then a large number of people with impairments must, with appropriate interventions, be able to achieve functional abilities above the population mean. Similar challenges arise in floor populations when the clinical issue involves measuring improvement reliably even if the improvement is only from "squeezing another person's hand" to "squeezing another person's hand firmly." Such improvements seem small, but in some clinical situations they may have significance.

The impact of the availability of adequate measures for floor and ceiling populations will be most felt in computerized adaptive testing. Item response theory yields improved items, but it is often used with too few items that cover the extremes, and subjects may be overwhelmed by the effort required to complete the questionnaire if extremes are covered. Computerized adaptive testing improves the efficiency with which extremes can be studied and reduces the effort required to complete the questionnaire; however, it is often used with too narrow an item bank, which decreases precision in measuring extremes. Extension of item coverage to include floor and ceiling populations is likely to prove useful for most outcome domains, not just physical function evaluation. For example, broadening the range of effective coverage should prove useful in very large populations, where an adequate number of subjects function at an extreme end of the scale, in smaller populations with a high percentage of subjects at an extreme (as with our nursing home residents or ultramarathon runners), and for construction of short forms to be used in relatively narrow populations of subjects whose health is not at the floor or ceiling. For the latter application, a test subpopulation is administered the computerized adaptive test, and the items most frequently selected define items needed for construction of static forms that focus on assessing subjects with a narrow range.

There are important limitations in the present study. Although the items were developed according to PROMIS protocol, they may not be relevant in all situations. Indeed many of the items are specific to athletes and thus may be of limited value in routine clinical care. Second, the testing was performed in a sample of subjects who were selected from preexisting cohort studies. These subjects were not enrolled based on random sampling of the underlying population. Inevitably, the validation population lacked the level of heterogeneity we would have preferred, consequently making it difficult to study the nuances of the performance of the items. Third, we administered ceiling items to high-functioning young individuals and floor items to elderly, frail, and sick individuals, which raises a question about the performance of these items in young, sick subjects and old, robust subjects.

A major contribution of improved health outcome assessment may be the ability to conduct research at a lower cost and with fewer subjects (27). The costs of medical research are driven largely by the number of subjects required to achieve the desired level of statistical power. The enrollment time, the number of study centers needed, supplies, time to complete the project, and other study costs are driven by sample size requirements. Our creation of items that accurately measure physical function at extreme ends of the ability spectrum enhances research infrastructure with greater measurement precision, but it will necessitate careful selection of primary outcomes.

In theory, one could develop tailored short forms that include items from the core item bank as well as either floor items or ceiling items. After performance of due-diligence psychometric testing, these short forms could be used in specific situations, such as among the military or community-dwelling senior citizens. Nonetheless, use of tailored short forms and computerized adaptive technology is a relatively new addition to health status assessment (unlike in educational testing). Furthermore, these methodologies have not been widely used in real-world situations and acceptance and endorsement by key stakeholders, such as the Food and Drug Administration, are still pending.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Fries had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Fries, Lingala, Bruce, Krishnan.

Acquisition of data. Fries, Lingala, Bruce, Krishnan.

Analysis and interpretation of data. Fries, Lingala, Siemons, Glas, Cella, Hussain, Bruce, Krishnan.

REFERENCES

1. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
2. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
3. Bruce B, Fries J. The Stanford Health Assessment Questionnaire (HAQ): a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
4. 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.
5. Ware JE, Jr., Keller SD, Hatoum HT, Kong SX. The SF-36 Arthritis-Specific Health Index (ASHI). I. Development and cross-validation of scoring algorithms. Med Care 1999;37:MS40–50.
6. World Health Organization. Constitution of the World Health Organization. Geneva: WHO; 1948.
7. Hays RD, Spritzer KL, Amtmann D, Lai JS, 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;94:2291–6.
8. Bruce B, Fries J, Lingala B, Hussain YN, Krishnan E. Development and assessment of floor and ceiling items for the PROMIS physical function item bank. Arthritis Res Ther 2013;15:R144.
9. PROMIS web site. URL: www.nihpromis.org.
10. Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, Hays RD. The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). Arthritis Care Res (Hoboken) 2011;63 Suppl 11:S486–490.
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. 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:R505–13.
13. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, et al. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther 2009;11:R191.
14. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness of physical function (disability) scales based upon item response theory (IRT) [abstract]. Arthritis Rheum 2009;60 Suppl: S229.
15. Cella D, Chang CH. A discussion of item response theory and its applications in health status assessment. Med Care 2000;38 Suppl: II66–72.
16. 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:S53–7.
17. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38 Suppl:II28–42.
18. Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. Med Care 2007;45:S32–8.
19. 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;36:2061–6.
20. Ware JE Jr, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlof CG, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual Life Res 2003;12:935–52.
21. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol 2007;34:1426–31.
22. Ware JE, Gandek B, Sinclair SJ, Bjorner JB. Item response theory and computerized adapive testing: implications for outcomes measurement in rehabilitation. Rehabil Psychol 2005;50:71–8.
23. Glas CA, van der Linden WJ. Marginal likelihood inference for a model for item responses and response times. Br J Math Stat Psychol 2010;63:603–26.
24. Te Marvelde JM, Glas CAW, van Landeghem G, van Damme J. Application of multidimensional item response theory models to longitudinal data. Educ Psychol Meas 2006;66:5–34.
25. Krishnan E, Sokka T, Hakkinen A, Hubert H, Hannonen P. Normative values for the Health Assessment Questionnaire disability index: benchmarking disability in the general population. Arthritis Rheum 2004;50:953–60.
26. Fries JF, Spitz PW, Mitchell DM, Roth SH, Wolfe F, Bloch DA. Impact of specific therapy upon rheumatoid arthritis. Arthritis Rheum 1986;29:620–7.
27. 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:R147.

APPENDIX A: PROMIS II

Institutions and principal investigators participating in Patient-Reported Outcomes Measurement Information System (PROMIS) II, in addition to the authors, are as follows: for Northwestern University, David Cella, PhD and Richard C. Gershon, PhD; for the American Institutes for Research, Susan (San) D. Keller, PhD; for the State University of New York, Stony Brook, Joan E. Broderick, PhD and Arthur A. Stone, PhD; for the University of Washington, Seattle, Dagmar Amtmann, PhD, Karon Cook, PhD, Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD; for the University of North Carolina, Chapel Hill, Darren A. DeWalt, MD, MPH; for the Children's Hospital of Philadelphia, Christopher B. Forrest, MD, PhD; for Boston University, Stephen M. Haley, PhD; for the University of Michigan, Ann Arbor, David Scott Tulsky, PhD; for the University of California, Los Angeles, Dinesh Khanna, MD and Brennan Spiegel, MD, MSHS; for the University of Pittsburgh, Paul A. Pilkonis, PhD; for Fred Hutchinson Cancer Research Center, Carol M. Moinpour, PhD; for Georgetown University, Arnold L. Potosky, PhD; for the Children's Hospital Medical Center, Cincinnati, Esi M. Morgan DeWitt, MD, MSCE; for the University of Maryland, Baltimore, Lisa M. Shulman, MD; and for Duke University, Kevin P. Weinfurt, PhD. National Institutes of Health science officers participating to this project have included Deborah Ader, PhD, Vanessa Ameen, MD, Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztein, MD, Ellen Werner, PhD, and James Witter, MD, PhD. This manuscript was reviewed by PROMIS reviewers before submission for external peer review.