Published in final edited form as:
Arthritis Rheumatol. 2014 May ; 66(5): 1378–1387. doi:10.1002/art.38342.

EXTENDING THE FLOOR AND THE CEILING FOR ASSESSMENT OF PHYSICAL FUNCTION

James F. Fries, MD[1], **Bharathi Lingala, PhD**[1], **Liseth Siemons, MSc**[1,2], **Cees A. W. Glas, PhD**[2], **David Cella, PhD**[3], **Yusra N Hussain, MD**[1], **Bonnie Bruce, DrPH, MPH, RD**[1], and **Eswar Krishnan, MD, MPhil.**[1]

[1]Stanford University, CA USA [2]University of Twente, Enschede, The Netherlands [3]Northwestern University, IL, USA

Abstract

Objective—The objective of the current study was to improve the assessment of physical function by improving the precision of assessment at the floor (extremely poor function) and at the ceiling (extremely good health) of the health continuum.

Methods—Under the NIH PROMIS program, we developed new physical function floor and ceiling items to supplement the existing item bank. Using item response theory (IRT) and the standard PROMIS methodology, we developed 30 floor items and 26 ceiling items and administered them during a 12-month prospective observational study of 737 individuals at the extremes of health status. Change over time was compared across anchor instruments and across items by means of effect sizes. Using the observed changes in scores, we back-calculated sample size requirements for the new and comparison measures.

Results—We studied 444 subjects with chronic illness and/or extreme age, and 293 generally fit subjects including athletes in training. IRT 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 two to four at the floor and a factor of two at the ceiling.

Conclusion—Extending the range of physical function measurement can substantially improve measurement quality, can reduce sample size requirements and improve research efficiency. The paradigm shift from Disability to Physical Function includes 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.

INTRODUCTION

Over recent decades, quantitative assessment of functional disability has greatly assisted study of chronic diseases and their treatments. Over time, more complex and more precise

Address for correspondent/reprint requests: Eswar Krishnan MD, M.Phil, Stanford ARAMIS Program, 1000 Welch Road, Suite 203 Palo Alto CA 94304, T. 650-71258004; Fax: 888-419-2656, e.krishnan@stanford.edu.

Competing Interests: The authors declare that there are no other financial interests and that none of the authors have any conflict of interest or the appearance of a conflict of interest with regard to the work.

measures have been developed, including the Health Assessment Questionnaire Disability Index (HAQ-DI) (1) and the PF-10 derived from the SF-36 (2), which assess a range of functional ability captured as patient-reported outcomes. These instruments, although more quantitative than their predecessors and proven useful in clinical trials and observational studies, (3–5) have disadvantages. The World Health Organization (WHO) definition of “ health” as “not merely the absence of disease, but complete physical, psychological, and social well-being” (6) foresaw a need not only to measure impairments worse than the population mean but also function above that mean. Thus, the domain of “Disability” requires redefinition as “Physical Function (PF)” (7, 8) and the focus of health status measurement has become measurement of physical function as opposed to assessment of disability. In the first cycle of the NIH Patient Reported Outcomes Medical Information System (PROMIS) program, we developed a Core Physical Function item bank (PROMIS PF n=154) containing 124 new items, “legacy” items from PF-10 (n=10, subscale within Medical Outcomes Study 36-item Short Form Health Survey) and “legacy” items from the HAQ-DI (n= 20) (9). The new items and instruments derived from them outperform prior instruments both in terms of efficiency and of precision, allowing studies to be performed with smaller sample sizes (10, 11).

Although it represented an advancement, some limitations from legacy measures persisted in the PROMIS physical function item bank, the most important being insensitivity to changes at the extremes of physical function. The floor and ceiling effects meant that large changes in true physical function among the frail and the robust, respectively, were necessary before these changes were reflected in the physical function metric (8). For instance, in a prospective observational study of 6,436 patients with rheumatoid arthritis who were longitudinally followed for 32,324 person-years with 64,647 HAQ-DI measurements, and an average of 19 measurements per person, 10% scored zero, signifying no disability (12). Subtle but clinically critical changes thus 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 the extremes may be missed.

Assessment of the extremes of function requires a sufficient number of validated items for each extreme, tested with sufficient numbers of subjects at the functional extremes, and a broader measurement metric to provide stable estimates. This paper describes the development of new physical function floor and ceiling items to supplement our PF item bank using item response theory (IRT) and standard PROMIS methodology. We applied these items in a prospective observational study setting and showed that: (1) addition of floor and ceiling items to existing core physical function item banks increase the statistical power of research across the full spectrum of human abilities; (2) use of these measures enable more precise study of populations at the extremes such as institutionalized patients or trained athletes; and (3) that use of the new physical function item bank requires smaller sample sizes to achieve a given level of precision than were necessary with previous instruments.

METHODS

Theoretical framework

Stanford has been a Primary Research Site of PROMIS over the past nine years with a particular focus on improving assessment of physical function (9). PROMIS is an NIH Roadmap Program (http://nihpromis.org) tasked with improving the infrastructure of clinical research by using IRT (13–19) and computerized adaptive testing (CAT) (11, 19–22). IRT allows improved measurement through selection and optimization of the best available items and aggregation of these to develop better instruments. The PROMIS approach to item bank development includes identification of candidate items, item improvement, qualitative item evaluation, study of clarity, translatability, importance, and quantitative validation.

Items and instruments

The existing PROMIS instruments have 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 then taken through a series of review processes described previously (15, 21). The PROMIS protocol used for developing a set of psychometrically optimal items has been described (11). Qualitative evaluation of these new floor and new ceiling items has been reported (8). We found that items easier or harder than existing core items could be constructed, understood by subjects, and efficiently administered and scored. The level of item difficulty was increased at the ceiling by a factor of five with addition of 26 new items of higher difficulty and reduced at the floor by a factor of 4 with addition of 30 new floor items of lesser difficulty. Tables 1 and 2 list the newly developed items, the PROMIS PF-20 items (evolved from the HAQ-DI), and the Legacy PF-10 items derived from the SF-36 used in the present study. Since the items were developed using IRT, we were able to aggregate individual items to create instruments that could be subject to further analyses. For the present analyses we present the raw scores for the sake of statistical clarity, although the PROMIS convention is to present physical function scores only in terms of T statistic distribution with a mean of 50 and standard deviation of 10.

IRT analyses

Item and test information curves were studied to examine whether the ten best new floor and ceiling items did indeed assess the extremes of 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 where such measurement is reliable*. Analyses were performed with the statistical program MIRT (23), using a multidimensional generalized partial credit model for repeated measures (24), taking into account the dependency between responses at baseline and at 12 months. Since the same test was 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 developed which are applicable at the extremes but also that subjects be included who were sufficiently disabled or sufficiently healthy to accurately assess the new items. We did not administer floor items to ceiling populations or vice versa since asking a marathoner whether he can squeeze a person's hand or a nursing home resident whether she can run five miles was considered potentially offensive. Since not all impaired persons or of fit persons are the same, we sought diversity in subjects to improve the generalizability of findings. The administration consisted of a 12-month longitudinal study of 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 with disease duration and aging (11, 25).

Floor Population

Floor items were tested in 444 subjects known to have poor functional status. The floor populations were 65% female, aged 86 years on average, over 90% white, and averaged 16 years of education. They had baseline PROMIS PF-20 (the PROMIS successor to the HAQ-DI) raw scores averaging 38.5 on a 0–100 scale, where 0 is highly able and 100 is severely disabled. Sixteen 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. This average baseline raw score is similar to patients with severe rheumatoid arthritis or severe osteoarthritis (1, 3, 26). The floor participants were administered newly developed floor items (n=30), the PROMIS PF-20 item form, and the Legacy PF-10, for a total of 60 items (Table 1) at baseline and after 12 months.

Ceiling population

The 293 ceiling subjects consisted of 107 vigorous exercisers including ultra-marathoners, 147 apparently healthy seniors, and 39 patients with no more than mild symptoms of arthritis. They were an average of 60 years of age, had 17 years of education, were over 85% white, and 60% male. They had PROMIS PF-20 Physical Function raw scores averaging 3.4 (on the 0–100 scale) indicating excellent physical function. These ceiling subjects completed the items in the PROMIS PF-20 and the Legacy PF-10 as well as the new 26 ceiling items (Table 2) at baseline and at 12 months. 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 at the item level and at the instrument level was used to determine whether the expanded item range increased or decreased the "effect sizes" of simulated "instruments" in populations at the extremes and in broader populations containing both average and extreme subjects. The groups are termed "simulated" since results are obtained from administration of items as a group of 60 items, whereas reference instruments are usually administered 10 or 20 items at a time. Effect sizes at the item level were used to select the 10 and 20 best individual floor and ceiling items. We did not further study items with effect sizes close to zero since there was no expected improvement with their addition, and since we wished to compare instruments holding item numbers constant at 10 or at 20.

RESULTS

Difficulty of new floor and new ceiling items at baseline

Using data from baseline administration, we cross-sectionally compared the difficulty of the new items and the reference items in both the floor and ceiling groups to confirm that the new items were indeed easier or more difficult than the reference items. All comparisons revealed a clear and consistent separation of old core items and new items, both toward the floor and toward the ceiling (Tables 1 and 2, Figure 1).

Item-level effect sizes for floor population

Table 1 shows item-level baseline and 12-month final scores in the floor population as well as the 12-month changes in scores and standard deviations. Item-level effect sizes are also shown, computed as the change between baseline and final means divided by the pooled standard deviation for each item. Within each instrument (new, PROMIS PF-20, and Legacy PF-10), items were sorted by item effect sizes. These items were likely to make the largest contributions to the instrument-level effect size, which in turn indicates the discrimination from the entire instrument. The direction of change for the large majority of items showed decreases over time, with only a few items with small effect sizes suggesting improvement. From 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. The 10 of the 30 tested floor items with the lowest effect sizes ranged from 0.05 to 0.0 and the direction of change was opposite in four of ten items which were discarded from further analyses because of the inconsistencies. Item-level effect sizes for the PROMIS PF-20 items ranged from 0.11 to 0.0, and the twelfth and seventeenth items had the direction of change reversed. For the Legacy PF-10 effect sizes ranged from 0.05 to 0.0, and the direction of change was reversed from the other two instruments in six out of ten items.

Instrument-level effect sizes for floor population

Six “instruments” were simulated from floor population data (Table 2). These are the Legacy PF-10, the PROMIS PF-10 (a subset of the PROMIS PF-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 PF-10 and the 10 new floor items with highest effect sizes. Differences between these simulated “instruments” were assessed by means, standard deviation of means, p-values from pair-wise t-tests, standardized response mean (SRM), Cohen’s effect size, minimum detectable difference (MDD), and sample-size requirements.

Baseline and final values were statistically different from each other for each simulated “instrument” except for the Legacy PF-10, which also showed a difference in the direction of change. There are several plausible explanations for this finding: the Legacy PF-10 has only 10 items, only three response items rather than five, and these ten items are known to be more sensitive toward the middle of a normal population than at the extremes. For all of the other "instruments" statistical differences were observed between baseline and final values, and all of the statistics tested were consistent. Twenty-item instruments out-performed the ten-item instruments, and the 30-item scale was roughly similar to the 20-item scales. The PROMIS PF-10 and PF-20 item scales detected change at p<0.01, and the 10 and 20 item floor instruments and the PROMIS PF-10/new floor-10 detected change at p<0.001. Sample size requirements in this floor population were reduced by a factor of 0.25 to 0.5 (from about 150 per arm in the PROMIS PF-10 and -20 item instruments to about 115 in the instruments containing new floor items). Sample sizes were those required to reach a MDD of 2.5%.

Item-level effect sizes for ceiling population

Baseline and 12-month scores in the ceiling population and the 12-month change scores and their standard deviations are shown in Table 3. Each of the 30 new ceiling items detected progression after 12 months. The effect size of the change score ranged from 0.25 to 0.0 with a median of 0.10. With the same PROMIS PF-20 instrument used in the floor population studies all items predicted progression, with effect sizes of 0.18 to 0.00 with a median of about 0.05. All Legacy PF-10 items also predicted progression, with item-level effect sizes ranging from 0.17 to 0.04 and a median of 0.13. These effect sizes are approximately double those seen in with the new items designed to evaluate the floor population.

Instrument-level effect sizes for ceiling population

As was done for the floor population, six "instruments" were simulated using data from the ceiling population. The Legacy PF-10, the PROMIS PF-10, and the PROMIS PF-20 described above. The new ceiling-10, new ceiling-20, and new ceiling-30 instruments included the 10, 20, and 30 ceiling items, respectively, with the largest item-level effect sizes. The Legacy PF-10 had a lower p-value than the PROMIS PF-10 and PF-20. This occurred because the Legacy PF-10 performs at its best at 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 including increase to five response options. There was less change over time in the ceiling population than in the floor population, and differences across instruments are harder to detect. Overall, the hybrid instrument of the PROMIS PF-10 and the ceiling PF-10 outperformed other simulated instruments, with the lower sample size requirements, increased effect sizes, and lower p-values. All instrument results were adjusted to an MDD level of 2.5%.

Information curves and population distribution

Figure 1 summarizes the difficulty levels of individual items which provided the highest effect sizes in floor and ceiling populations, respectively, in the Legacy PF-10 and the 10 new floor and 10 new ceiling items. The percentages responding "unable to do" and the percent responding "with no difficulty" are given. The results show that the new floor and ceiling items function better at the extremes of the scale than the PF-10 items, expanding the measurement range. Figure 2 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 floor and ceiling items extend the range toward the floor by one to two standard deviations and the range toward the ceiling by about two standard deviations.

DISCUSSION

Quantitative assessment of functional disability has greatly assisted study of chronic diseases and their treatments. However, currently used assessments fail to discriminate well at either extreme of physical ability. We demonstrate that items which 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 its potential utility 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 ten best items from both floor and ceiling resulted in additional slight improvement. The addition of more items with effect sizes near zero added little, and paradoxical effects and less consistency resulted. An adequate Physical Function Core Item Bank for study of a wide range of function might contain about 200 items.

With the optimized instrument, small discrepancies are observed between the population and instrument measurement range. The new floor items are located even more towards the extreme than the floor population, leaving room for the measurement of physical function levels in even worse performing populations. In contrast, the distribution of the ceiling population is located nearer the extreme than the range measured by the ceiling items. This may indicate that a ceiling effect is present with the optimized instrument, even though the inclusion of the new ceiling items diminished this effect relative to that observed with the PROMIS PF-20 or the Legacy PF-10.

The importance of addressing floor and ceiling effects are evident in the case of rheumatoid arthritis; about 10% of patients who clearly suffer from age-related and disease-related limitations do not have measurable disability using the HAQ-DI (12). This issue is brought to a sharper relief in a random sample of general population where 75% do not report any disability, and the 75th percentile of HAQ-DI was zero among men and women aged 30–55 (25). Moreover, other interesting subject groups cannot be adequately evaluated with currently used instruments. If health is “not merely the absence of disease but complete physical, social, and psychological well-being” (WHO) then large numbers of persons with impairments must, with appropriate interventions, be able to achieve functional abilities above the population mean. Similar challenges arise at the floor 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 may have major meaning.

The impact of the availability of adequate measures for floor and ceiling populations will be most felt in CAT applications. IRT yields improved items but often is used with too few items that cover the extremes, and may be plagued with questionnaire burden if extremes are covered. CAT improves the efficiency of study at the extremes and reduces questionnaire burden but is often used with too narrow an item bank, which decreases its ability to reach precision at the extremes. Extension of item-coverage toward the floor and the ceiling 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 the extremes will be adequately populated with subjects; in smaller populations with a high percent of persons at an extreme, as in our nursing home residents or ultra-marathoners; and for construction of short-forms to be used in relatively narrow populations not at the floor or ceiling. For the latter application, a test sub-population is given the CAT, and the items most frequently selected define items needed for construction of static forms focused at narrow ranges of subjects.

There are important limitations in the present study. Although the items were developed according to the PROMIS protocols, it may not be relevant in all situations. Indeed many of the items are specific to athletes and thus may be of limited value for routine clinical care. Secondly, the testing was performed in a sample of individuals identified from pre-existing cohort studies. These people were not enrolled based on random sampling from the underlying population. It is inevitable that the validation population lack as much heterogeneity as one would wish for, consequently making it difficult to study the nuances of the population performance of the items. Thirdly, we administered ceiling items to high functioning young individuals and floor items to the elderly frail and sick which raises the question about the performance of these items in young but sick patients and old but robust people.

A major contribution of improved health outcome assessment may be the ability to conduct research with lower costs and fewer subjects (27). The costs of medical research are driven to a large degree by the number of subjects required to achieve desired statistical power. The period to complete enrollment, 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 extremes of ability enhances research infrastructure with greater measurement precision but 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 performing due diligence psychometric testing, these short forms could be used in specific situations such as the military or community dwelling senior citizens. Nonetheless, the concept of tailored short forms and CAT methodology are relatively new in health status assessment unlike in educational testing. Furthermore, these methodologies have not been well operationalized in real world situations and acceptance and endorsements by key stakeholders such as the Food and Drug Administration are still pending.

Acknowledgments

This work was supported by the Patient-Reported Outcomes Measurement Information System (PROMIS), Stanford University Primary Research Site - PI: James F. Fries, MD, #2U01AR052158.

PROMIS II was funded by cooperative agreements with a Statistical Center (Northwestern University, PI: David Cella, PhD, 1U54AR057951), a Technology Center (Northwestern University, PI: Richard C. Gershon, PhD, 1U54AR057943), a Network Center (American Institutes for Research, PI: Susan (San) D. Keller, PhD, 1U54AR057926) and thirteen Primary Research Sites which may include more than one institution (State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD and Arthur A. Stone, PhD, 1U01AR057948; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, 1U01AR057954; University of Washington, Seattle, PIs: Dagmar Amtmann, PhD and Karon Cook, PhD, 1U01AR052171; University of North Carolina, Chapel Hill, PI: Darren A. DeWalt, MD, MPH, 2U01AR052181; Children's Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, 1U01AR057956; Stanford University, PI: James F. Fries, MD, 2U01AR052158; Boston University, PIs: Stephen M. Haley, PhD and David Scott Tulsky, PhD (University of Michigan, Ann Arbor), 1U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD and Brennan Spiegel, MD, MSHS, 1U01AR057936; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, 2U01AR052155; Georgetown University, PIs: Carol. M. Moinpour, PhD (Fred Hutchinson Cancer Research Center, Seattle) and Arnold L. Potosky, PhD, U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan DeWitt, MD, MSCE, 1U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, 1U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, 2U01AR052186). NIH Science Officers on 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. See the Web site at www.nihpromis.org for additional information on the PROMIS initiative.

References

1. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980; 23:137–145. [PubMed: 7362664]
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; 30:473–483. [PubMed: 1593914]
3. Bruce B, Fries J. The Stanford health assessment questionnaire (HAQ): a review of its history, issues, progress, and documentation. J Rheumatol. 2003; 30:167–178. [PubMed: 12508408]
4. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. J Rheumatol. 1982; 9:789–793. [PubMed: 7175852]
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. [PubMed: 10335742]
6. World Health Organization. Constitution of the World Health Organization. Geneva: WHO; 1948.
7. Hays, RD.; Spritzer, KL.; Amtmann, D.; Lai, J-S.; DeWitt, EM.; Rothrock, N.; DeWalt, DA.; Riley, WT.; Fries, JF.; Krishnan, E. Upper extremity and mobility subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS®) adult physical functioning item bank. Archives of Physical Medicine and Rehabilitation. 2013. http://dx.doi.org/10.1016/j.apmr.2013.05.014
8. Bruce B, Fries JF, Lingala B, Hussain YN, EK. Development and Assessment of Floor and Ceiling Items for the PROMIS Physical Function Item Bank. Arthritis Res Ther. 2013 in press.
9. 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, JEW. Preliminary Evaluations of a Physical Function Item Bank Support the Methods and Advantages of the Patient Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiology. 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:R505–513. [PubMed: 15535828]
13. Bruce B, Fries JF, Ambrosini D, Lingala B, Gandek B, Rose M, Ware JE Jr. Better assessment of physical function: item improvement is neglected but essential. Arthritis Res Ther. 2009; 11:R191. [PubMed: 20015354]

14. Fries JF, Krishnan E, Rose M, Lingala BBB. Improved responsiveness of physical function (disability) scales based upon item response theory (IRT). Arthritis & Rheumatism. 2009; 60(suppl 10):S229.

15. Cella D, Chang CH. A discussion of item response theory and its applications in health status assessment. Med Care. 2000; 38:II66–72. [PubMed: 10982091]

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:S53–57. [PubMed: 16273785]

17. Hays RD, Morales LS, Reise SP. Item Response Theory and health outcomes measurement in the 21st century. Med Care. 2000; 38(9 Suppl):1128–1142.

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:S32–38. [PubMed: 17443117]

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; 36:2061–2066. [PubMed: 19738214]

20. Ware JE Jr, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlof CG, Tepper S, Dowson A. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Qual Life Res. 2003; 12:935–952. [PubMed: 14651413]

21. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol. 2007; 34:1426–1431. [PubMed: 17552069]

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; 50:71–78.

23. Glas CA, van der Linden WJ. Marginal likelihood inference for a model for item responses and response times. Br J Math Stat Psychol. 63:603–626. [PubMed: 20109271]

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. Arthritis & Rheumatism. 2004; 50:953–960. [PubMed: 15022339]

26. Fries JF, Spitz PW, Mitchell DM, Roth SH, Wolfe F, Bloch DA. Impact of specific therapy upon rheumatoid arthritis. Arthritis Rheum. 1986; 29:620–627. [PubMed: 3718554]

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. [PubMed: 21914216]

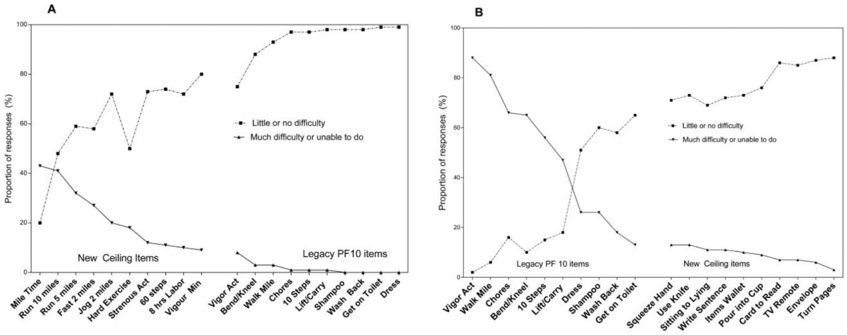

Figure 1.

Levels of difficulty of individual items in the Legacy PF-10 and the 10 new floor and 10 new ceiling items. For the respective populations new ceiling and new floor items are improved over reference items. In the ceiling population, new items are more difficult and in the floor population new items are considerably easier than reference items.

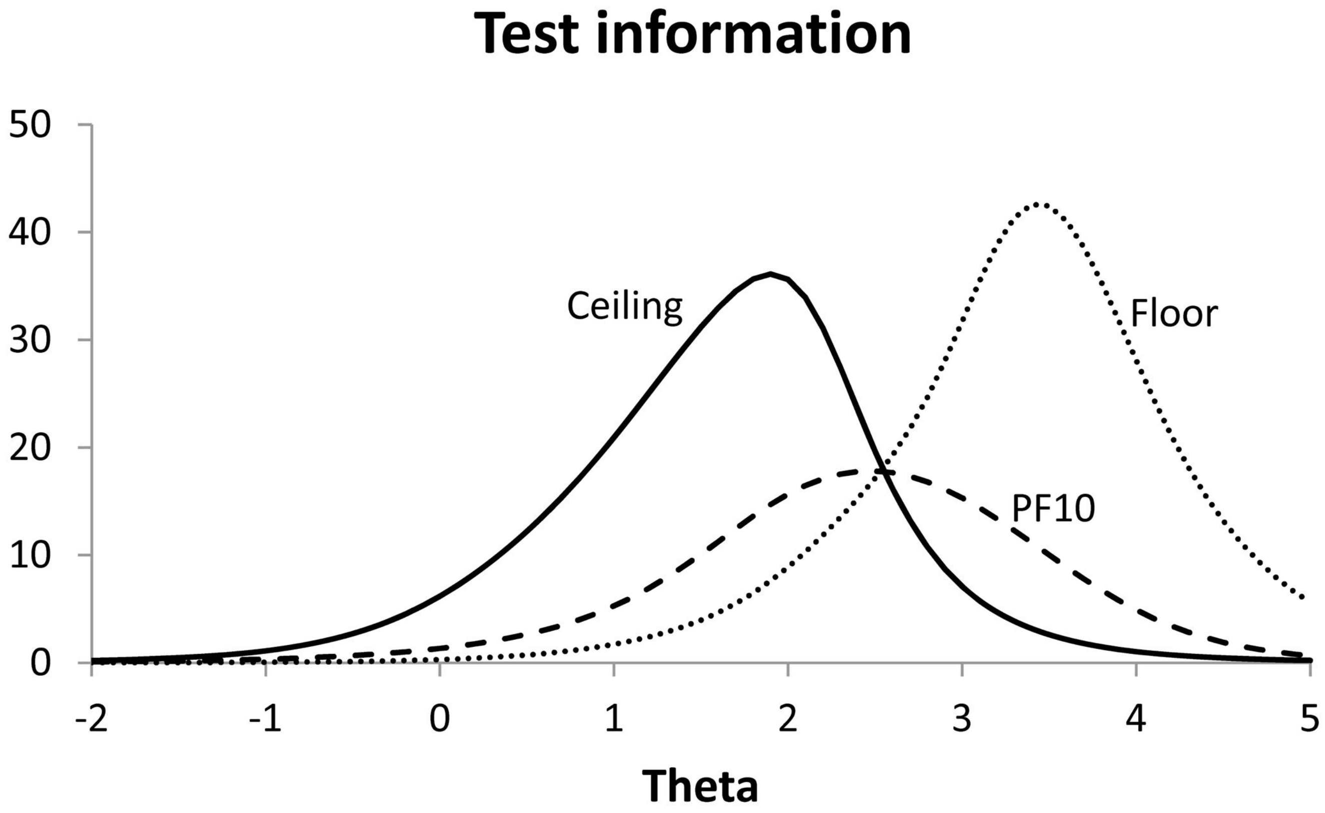

Figure 2.

Baseline test information curves of the instruments containing the 10 best floor items (left curve), the 10 PF-10 items (center curve), and the 10 best ceiling items (right curve). Theta scores (scaled around 0) correspond to the level of physical function, where higher scores represent worse function.

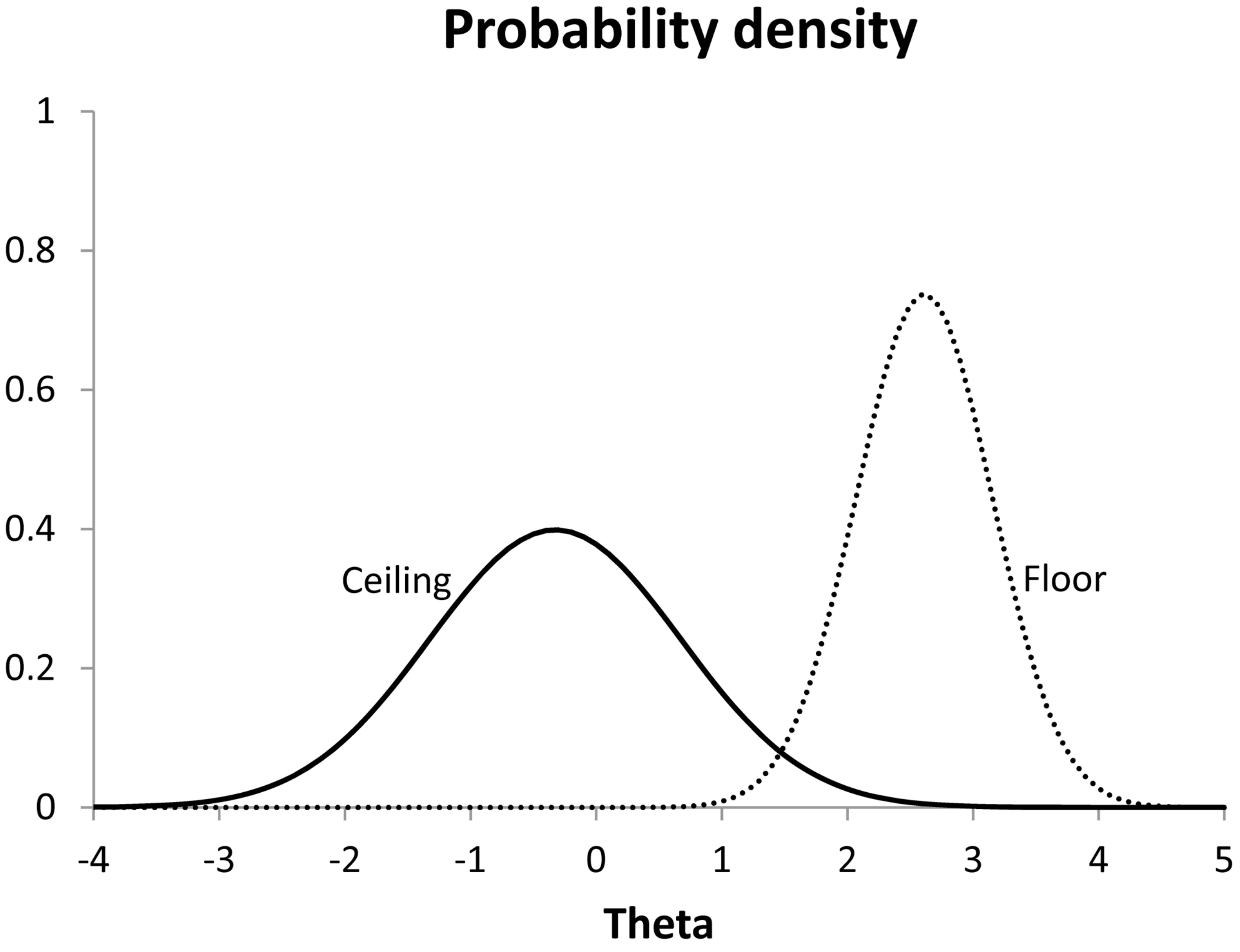

Figure 3.

Frequency distributions of the floor and ceiling populations in relationship to the continuum of physical function. Theta scores (scaled around 0) correspond to the level of physical function, where higher scores represent worse function.

Table 1

Baseline, 12-month, and Change Scores in Floor Population

	Baseline (Mean (SD))	12-month Follow Up (Mean (SD))	12-Month Change Score (Mean (SD))	Item Level Effect Size (Mean Change/ Pooled SD)
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)	0.1(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)	0.1(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)	0.7(1)	0.1(0.7)	0.09
Move from sitting on the bed to lying down?	0.7(1)	0.8(1)	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(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(0.5)	0.07
Loosen a screw using a manual screwdriver?	1.1(1.3)	1.2(1.3)	0.1(1)	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)	0.1(0.7)	0.06
Chew and eat your food as quickly as five years ago?	1(1.2)	1(1.3)	0.1(1)	0.06
What is the farthest distance you can walk by yourself?	1.9(1.3)	2(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(0.7)	0.05
Put on your shoes?	0.9(1.1)	1(1.2)	0.1(0.8)	0.05
Dress yourself in less than ten minutes?	1.4(1.3)	1.4(1.4)	0.1(1)	0.05
Do you feel exhausted?	1.7(0.9)	1.7(0.9)	0(0.8)	0.04
Dress and groom yourself as quickly as five years ago?	2.1(1.4)	2.2(1.4)	0(1.1)	0.03
Take a letter out of an envelope?	0.3(0.7)	0.3(0.7)	0(0.5)	0.03
Put on your socks?	1.1(1.2)	1.1(1.2)	0(0.8)	0.02
Move from street to sidewalk without a curb cut?	1.2(1.2)	1.3(1.2)	0(0.8)	0.02
Push the buttons on a television remote control?	0.3(0.8)	0.3(0.7)	0(0.6)	0.02
Move about your residence?	0.7(1)	0.7(1)	0(0.7)	0.02
Compared to 5 years ago, is your normal walking speed:	2.4(0.8)	2.4(0.8)	0(0.6)	0.02
In the past year, amount of unintentional weight loss?	0.6(0.9)	0.6(0.9)	0(1)	0.01
Walk a block as quickly as you did five years ago?	2.8(1.3)	2.8(1.3)	0(1)	0.01
Walk up or down inclines?	1.5(1.2)	1.5(1.2)	0(0.9)	0.00
PROMIS PF-20 Items				
Are you able to wash & dry your body	0.8(1.1)	0.9(1.2)	0.1(0.8)	0.11
Are you able to hold a plate full of food	0.6(1)	0.7(1.1)	0.1(0.8)	0.10
Are you able to 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(1.2)	3.1(1.1)	0.1(0.9)	0.10

	Baseline (Mean (SD))	12-month Follow Up (Mean (SD))	12-Month Change Score (Mean (SD))	Item Level Effect Size (Mean Change/ Pooled SD)
Are you able to shampoo your hair	0.8(1.3)	0.9(1.3)	0.1(0.8)	0.08
Are you able to do chores like vacuuming or yardwork?	2.2(1.4)	2.3(1.3)	0.1(1)	0.07
Are you able to 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)	3.4(0.9)	0.1(0.8)	0.06
Are you able to dress yourself?	1(1.2)	1.1(1.2)	0.1(0.8)	0.06
Are you able to dry your back with a towel?	0.9(1.2)	0.9(1.2)	0(0.8)	0.04
Are you able to get in/out of a car?	1.1(1)	1(0.9)	0(0.7)	0.03
Walking more than a mile?	2.9(1.3)	2.9(1.4)	0(0.9)	0.03
Are you able to squeeze a new tube of toothpaste?	0.4(0.9)	0.5(0.8)	0(0.7)	0.03
Bending, kneeling, or stooping?	2.3(1.1)	2.3(1.1)	0(1)	0.03
Climbing one flight of stairs?	1.9(1.3)	1.9(1.4)	0(0.9)	0.02
Are you able to run a short distance to catch a bus?	2.8(1.4)	2.8(1.4)	0(1)	0.01
Are you able to transfer from a bed to a chair and back?	0.7(1)	0.7(1)	0(0.7)	0.01
Are you able to wash your back?	1.5(1.3)	1.6(1.2)	0(0.9)	0.01
Are you able to get on/off a toilet?	0.8(1)	0.8(0.9)	0(0.7)	0.00
Legacy PF-10 Items				
Climb one flight of stairs?	1(0.8)	0.9(0.8)	0(0.7)	0.05
Walking more than a mile?	1.5(0.7)	1.5(0.7)	0(0.7)	0.04
Vigorous activities, like running or strenuous sports?	1.8(0.5)	1.8(0.5)	0(0.6)	0.04
Bending, kneeling, or stooping?	1.2(0.7)	1.2(0.7)	0(0.7)	0.03
Climbing several flights of stairs?	1.4(0.7)	1.4(0.7)	0(0.6)	0.02
Lifting or carrying groceries?	1(0.7)	0.9(0.7)	0(0.6)	0.02
Walking several hundred yards?	1(0.8)	1(0.8)	0(0.7)	0.01
Walking one hundred yards?	0.8(0.8)	0.8(0.8)	0(0.6)	0.01
Bathing and dressing yourself?	0.5(0.7)	0.5(0.7)	0(0.6)	0.00
Moderate activities, like moving a table or golf?	1.3(0.7)	1.3(0.7)	0(0.6)	0.00

PF: Physical function, SD: standard deviation

Table 2

Baseline, 12-month, and Change Scores in Ceiling Population

	Baseline (Mean (SD))	12-Month Follow Up (Mean (SD))	12-Month Change Score (Mean (SD))	Item Level Effect Size (Mean Change/ Pooled SD)
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 (1.2)	0.3 (0.8)	0.23
Climb five 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)	0.1 (0.6)	0.15
Doing heavy work around the house?	0.5 (0.8)	0.6 (1)	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)	0.1 (0.7)	0.14
Past week, total time on vigorous physical activity?	0.7 (1)	0.8 (1.2)	0.1 (0.9)	0.13
Doing eight hours of physical labor?	0.9 (1.1)	1 (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 of/drive way	0.6 (1)	0.8 (1.2)	0.1 (0.8)	0.11
Transfer a full load of clothes from a washer to dryer	0 (0.1)	0 (0.3)	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)	0.1 (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 1000 vertical feet on a trail in an hour?	0.6 (1)	0.7 (1)	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 two miles?	0.9 (1.5)	1 (1.4)	0.1 (0.7)	0.06
Run at a fast pace for two 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 five 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 (0.6)	0.04
How many minutes does it take for you to walk one mile?	0.7 (0.8)	0.8 (0.9)	0 (0.7)	0.02
Run ten miles?	1.8 (1.7)	1.8 (1.7)	0 (0.8)	0.00
PROMIS PF-20 Items				
Lifting or carrying groceries?	0.1 (0.3)	0.1 (0.4)	0.1 (0.3)	0.18
Are you able to 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 (0.4)	0.11

	Baseline (Mean (SD))	12-Month Follow Up (Mean (SD))	12-Month Change Score (Mean (SD))	Item Level Effect Size (Mean Change/ Pooled SD)
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
Are you able to push open a heavy door?	0.1 (0.4)	0.1 (0.5)	0 (0.4)	0.07
Chores such as vacuuming or yardwork?	0.2 (0.5)	0.2 (0.6)	0 (0.5)	0.05
Doing two hours of physical labor?	0.4 (0.8)	0.4 (0.9)	0 (0.6)	0.05
Are you able to dry your back with a towel?	0 (0.3)	0 (0.2)	0 (0.3)	0.04
Are you able to sit on the edge of a bed?	0 (0.2)	0 (0.2)	0 (0.3)	0.03
Are you able to get on/off a toilet?	0 (0.3)	0 (0.2)	0 (0.3)	0.03
Are you able to transfer from a bed to a chair and back?	0 (0.3)	0 (0.2)	0 (0.3)	0.03
Are you able to run a short distance to catch a bus?	0.2 (0.7)	0.3 (0.7)	0 (0.7)	0.02
Are you able to squeeze a new tube of toothpaste?	0 (0.3)	0 (0.2)	0 (0.3)	0.02
Are you able to get in/out of a car?	0.1 (0.3)	0.1 (0.3)	0 (0.4)	0.01
Are you able to hold a plate full of food?	0 (0.3)	0 (0.2)	0 (0.3)	0.01
Are you able to shampoo your hair?	0 (0.3)	0 (0.2)	0 (0.3)	0.01
Are you able to dress yourself?	0.1 (0.3)	0.1 (0.3)	0 (0.3)	0.00
Are you able to wash and dry your body?	0 (0.3)	0 (0.2)	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 (0.2)	0.1 (0.3)	0 (0.3)	0.17
Bathing and dressing yourself?	0 (0.1)	0 (0.2)	0 (0.2)	0.14
Walking one hundred yards?	0 (0.1)	0.1 (0.3)	0 (0.2)	0.14
Walking more than a mile?	0.1 (0.3)	0.1 (0.4)	0 (0.3)	0.13
Climbing several flights of stairs?	0.1 (0.4)	0.2 (0.4)	0 (0.4)	0.13
Moderate activities, like moving a table or golf?	0.1 (0.4)	0.1 (0.4)	0 (0.4)	0.11
Lifting or carrying groceries?	0 (0.2)	0.1 (0.3)	0 (0.3)	0.10
Walking several hundred yards?	0 (0.2)	0.1 (0.3)	0 (0.3)	0.04
Bending, kneeling, or stooping?	0.2 (0.4)	0.2 (0.5)	0 (0.4)	0.04

PF: Physical function, SD: standard deviation

Table 3

Physical function instruments, metric raw scores, and sample size requirements for floor and ceiling items. While the PROMIS convention is to express T scores with a mean of 50 and range 0 to 100, raw scores are provided here for clarity.

	Baseline Mean (SD)	Follow Up Mean (SD)	Change Score Mean (SD)	p-value Pair-wise t-test	SRM[1]	Cohen's Effect Size[2]	Min. Detect. Diff.[3]	Sample Size (n)[4]
Physical Function - Floor Respondents								
Instruments								
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	22.1 (19)	24.7 (20)	2.6 (10)	<0.001	0.26	0.13	1.35	130
Floor top 20	22.8 (18)	24.9 (19)	2.1 (9)	<0.001	0.23	0.11	1.23	108
Floor top 30	26.7 (18)	27.9 (18)	1.2 (8)	0.002	0.15	0.07	1.09	86
PROMIS SF-10 and Floor 10	33.3 (19)	35.3 (20)	2.0 (9)	<0.001	0.22	0.10	1.21	106
Physical Function - Ceiling Respondents								
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)	5.1 (10)	0.9 (7)	0.03	0.13	0.10	1.15	64
PROMIS SF-20	3.4 (7)	3.9 (8)	0.5 (7)	0.19	0.08	0.06	1.07	55
Ceiling top 10	12.5 (17)	16.3 (20)	3.9 (10)	<0.001	0.39	0.21	1.64	128
Ceiling top 20	13.1 (15)	15.9 (18)	2.8 (9)	<0.001	0.31	0.16	1.46	101
Ceiling top 30	16.9 (18)	19.1 (20)	2.1 (9)	<0.001	0.24	0.11	1.45	99
PROMIS SF-10 and Ceiling 10	8.3 (11)	10.7 (14)	2.4 (8)	<0.001	0.31	0.18	1.26	76

[1] SRM (Standardized Response Mean): mean change of the score/SD of change

[2] Cohen's Effect Size = Change in Mean/pooled SD

[3] Minimum Detectable Change at 80% power in the outcomes with Floor n=444; ceiling n=293

[4] The sample size requirement to detect the Difference in Population Means of 2.5 based on the observed SD of the change.

SF: short form, PF: physical function