Published in final edited form as:

Arch Phys Med Rehabil. 2013 September ; 94(9): 1661–1669. doi:10.1016/j.apmr.2013.03.021.

Development of a Computer-Adaptive Physical Function Instrument for Social Security Administration Disability Determination

Pengsheng Ni, MD, MPH¹, Christine M. McDonough, PT, PhD^{1,2}, Alan M. Jette, PT, PhD¹, Kara Bogusz, BA¹, Elizabeth E. Marfeo, PhD, MPH¹, Elizabeth K. Rasch, PT, PhD³, Diane E. Brandt, PT, MA, PhD³, Mark Meterko, PhD¹, Stephen M. Haley, PT, PhD, and Leighton Chan, MD, MPH³

¹The Health & Disability Research Institute, Department of Health Policy & Management, School of Public Health, Boston University Medical Campus, Boston, MA

²The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Lebanon NH

³Clinical Research Center, Rehabilitation Medicine Department, National Institutes of Health, Bethesda, MD

Abstract

Objectives—To develop and test an instrument to assess physical function (PF) for Social Security Administration (SSA) disability programs, the SSA-PF. Item Response Theory (IRT) analyses were used to 1) create a calibrated item bank for each of the factors identified in prior factor analyses, 2) assess the fit of the items within each scale, 3) develop separate Computer-Adaptive Test (CAT) instruments for each scale, and 4) conduct initial psychometric testing.

Design—Cross-sectional data collection; IRT analyses; CAT simulation.

Setting—Telephone and internet survey.

© 2013 The American Congress of Rehabilitation Medicine. Published by Elsevier Inc. All rights reserved.

CORRESPONDING AUTHOR CONTACT: Pengsheng Ni, MD, MPH, Health & Disability Research Institute, 715 Albany Street T5W, Boston MA 02118, Ph: (617) 638-1989, Fax: (617) 638-1999, psni@bu.edu.

COMMERCIAL SUPPORT/CONFLICTS STATEMENT:

We certify that no party having a direct interest in the results of the research supporting this article has or will confer a benefit on us or on any organization with which we are associated AND, if applicable, we certify that all financial and material support for this research and work are clearly identified in the title page of the manuscript (Pengsheng Ni, Christine M. McDonough, Alan M. Jette, Kara Bogusz, Elizabeth E. Marfeo, Elizabeth K. Rasch, Diane E. Brandt, Mark Meterko, Stephen M. Haley, Leighton Chan).

PRESENTATION OF THIS MATERIAL:

“Development of a Computer Adaptive Test to Assess Physical Capabilities for Work Disability Determination” presented to the Disability Forum of the American Public Health Association, Washington, DC, November 1, 2011.

“Innovations in Self-Reported Function and Disability Assessment.” Presented at the Psychiatric Research Center, Dartmouth Medical School, Lebanon, NH. April, 2012.

An additional presentation was made of this material to the 4th Congress of the International Association of Bodily Impairment (AIDC), Montreal, Quebec on September 12, 2012.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Participants—Two samples: 1,017 SSA claimants, and 999 adults from the US general population.

Interventions—None.

Main Outcome Measure—Model fit statistics, correlation and reliability coefficients,

Results—IRT analyses resulted in five unidimensional SSA-PF scales: Changing & Maintaining Body Position, Whole Body Mobility, Upper Body Function, Upper Extremity Fine Motor, and *Wheelchair Mobility* for a total of 102 items. High CAT accuracy was demonstrated by strong correlations between simulated CAT scores and those from the full item banks. Comparing the simulated CATs to the full item banks, very little loss of reliability or precision was noted, except at the lower and upper ranges of each scale. No difference in response patterns by age or sex was noted. The distributions of claimant scores were shifted to the lower end of each scale compared to those of a sample of US adults.

Conclusions—The SSA-PF instrument contributes important new methodology for measuring the physical function of adults applying to the SSA disability programs. Initial evaluation revealed that the SSA-PF instrument achieved considerable breadth of coverage in each content domain and demonstrated noteworthy psychometric properties.

Keywords

Disability Evaluation; Disabled Persons; Insurance; Disability; Questionnaires; Psychometrics; Models; Theoretical; Statistical; Health Status Indicators

In 2011, the Social Security Administration (SSA)'s disability programs, which include Social Security Disability Insurance and Supplemental Security Income provided financial support to over 12 million Americans.¹ As the number of disability claims continues to rise, efforts to improve the efficiency and accuracy of the SSA's disability determination process are increasingly important. The current process aims to assess whether the claimant's impairments rise to the level of SSA's statutory definition of disability; the "inability to do any substantial gainful activity by reason of any medically determinable physical or mental impairment..."² This medical definition diverges substantially from current models, which define disability as a gap between the capability of the individual and the environment in which he/she functions.³⁻⁶ There is evidence that disability determination is shifting away from medical criteria toward functional information.⁷ In its 2007 report entitled, "Improving the Social Security Disability Decision Process," the Institute of Medicine (IOM) Committee noted that as medical treatment and assistive technologies advance, the medical criteria used in the SSA's disability adjudication process have become less useful as markers of disability.⁸ The IOM report recommended the development of alternative approaches, including the creation of standardized functional assessment instruments.

Clinician-administered and/or performance-based functional assessments are time consuming, costly, and impractical to implement in a program where over 3 million disability determinations are made each year.⁹ Although self-report instruments are more feasible to administer, there are concerns over potential reporting bias or inaccuracies. Comprehensive, fixed-length functional assessment instruments exist, but they are

burdensome and impractical for widespread use in the SSA's disability programs. Item-response theory (IRT) and computer adaptive test (CAT)-based assessment may be a promising means to provide the SSA with standardized functional assessment tools that are feasible for widespread implementation, and when combined with other medical evidence and workplace information, can inform the SSA's disability determination process.¹⁰

IRT calibration methods order questions or "items" along a unidimensional continuum allowing for the comprehensive assessment of individuals with a wide range of different medical impairments and varying degrees of severity. Such IRT-based calibration supports the development of CAT algorithms guiding the iterative selection of the most appropriate subset of questions from a larger instrument based on a claimants' response(s) to previous questions(s). Thus, the efficiency of IRT/CAT-based instruments makes it feasible to implement comprehensive functional assessment on a large-scale such as that of the SSA's disability programs.

In a previous study, we developed an item pool to assess the physical function of SSA disability claimants and conducted factor analyses that revealed its underlying multi-factor structure¹¹. This article describes the use of Item Response Theory (IRT) analyses to 1) calibrate the functional items on a continuum from low to high ability on each of the 5 identified sub-domains in a sample of SSA claimants, 2) assess the fit of the items within each scale, and 3) develop separate Computer-Adaptive Test (CAT) instruments for each scale. It further describes initial psychometric evaluation of the instrument using CAT simulation analyses to assess reliability and accuracy of the resulting SSA Physical Function (SSA-PF) instrument; and to compare the distributions of claimants' scores for each scale with those of a normative sample of adults living in the United States (US).

METHODS

The Boston University Internal Review Board approved all of the procedures used in this study. The sampling and data collection methods are provided in detail in a companion paper.¹¹ Brief descriptions follow.

Claimant and Normative Samples

The claimant sample was selected from a pool of 10,000 applicants for Social Security Disability Insurance or Supplemental Security Income. After geographic stratification by the 10 SSA regions and urban/rural location, participants were randomly selected for participation in the study. Study information and consent materials were sent to 7,800 claimants. The survey research firm, Westat Inc., screened potential participants by telephone using the following eligibility criteria: 21 years of age or older; filed the claim on his/her own behalf; able to read and understand English; and an allegation of a physical or physical and mental health condition.

To obtain norms for scale scores, a second sample was developed using unique methods developed by YouGov Inc.¹² These procedures included drawing a sample of 1,000 US adults from an ongoing opt-in internet panel of over 1 million respondents. Proximity

sample matching methods were used to match the US adult population on age, sex, racial/ethnic background, and education.

Data Collection Procedures

The claimant sample responded to the 139-item pool and demographic questions via telephone interview or self-administration using the internet. Participants were asked to report their “usual ability during a typical day,” without help from another person, equipment or devices not mentioned in the question, for example, “Are you able to use a lever handle to open a door?” Response options included; “Yes, without difficulty; yes, with a little difficulty; yes with some difficulty; yes, with a lot of difficulty; unable to do; and I don’t know.” The item pool also included items specific to wheelchair, and assistive device use. Responses to screener questions were used to select appropriate administration of the wheelchair module and adaptive device/equipment questions. The normative sample completed the same demographic questions and 139 items as the claimant sample via self-administered internet survey.

Analysis

Descriptive statistics of item category frequency and proportions were calculated for all discrete variables.

Claimant Sample

Item Calibration and Fit—We reversed the item response scores so that higher category scores indicated higher functioning and used Samejima’s Graded Response model¹³ to calibrate the claimant data organized in four item pools derived from prior factor analyses (Changing & Maintaining Body Position, Whole Body Mobility, Upper Body Function, Upper Extremity Fine Motor) and the Wheelchair Mobility item pool.¹¹ The model estimated a discrimination parameter that reflected the degree of the relationship between the latent factor underlying the scale and the item responses, and three or four threshold parameters, which indicated the score at which the subject would have a 50% probability of endorsing at or above a particular response category. The marginal maximum likelihood estimation procedure using IRTPRO software was used to calibrate the item parameters.¹⁴ We examined item fit by $S-X^2$ (Pearson’s chi-square) based on the summary score. We compared the expected and observed item frequency distribution in each summary score level and calculated overall fit statistics.¹⁵⁻¹⁷ Due to the multiple comparisons, we chose a p value <0.01 to indicate a significant misfit. The item fit was calculated using IRTFIT.¹⁸

Differential Item Functioning (DIF)—DIF indicates that participants in different subgroups with the same ability level have different probabilities of response to that item. We carried out the DIF analysis using the ordinal logistic regression model in SAS,¹⁹ in which the item response was modeled on the summed score of the all items in the item bank and on the variable of interest (i.e., age (<51 , or ≥ 51), sex). A finding of statistical significance for this background variable indicated that there was uniform DIF. A significant interaction effect (ie. a background variable and the summary score) indicated non-uniform DIF. The likelihood ratio test was used for model comparison; Bonferroni corrected p -values were used for significance testing ($0.05/\#$ of items); and the R-square change with Jodoin

and Gierl's criteria was used to quantify the effect size of the DIF.²⁰ Items with non-uniform DIF were removed from the final item banks. *CAT Construction:* Customized CAT algorithms were developed at Boston University for each of the five domains, and programmed to begin administration with an item from the middle of the difficulty range. The person score and standard error (SE) were then estimated using weighted likelihood estimation analysis.²¹ Each subsequent question was selected based on maximum item information matrix at the current score level, and the score and SE were recalculated after each response until a pre-specified minimum SE or maximum number of items had been administered. Scores were transformed so that the mean was 50 and standard deviation was 10, and higher scores represented higher function.

Psychometric Evaluation: Breadth of coverage, Reliability, Precision, and Accuracy

To conduct an initial assessment of the properties of the SSA-PF instrument,, breadth of coverage for each item bank by examining the score distribution and mapping each item response category's expected value onto the sample's score on each scale. Next we simulated CAT scores by feeding the claimant's actual answer to each item selected by the CAT algorithm. This allowed comparison of scores and SEs produced by the full item bank to those of simulated 5- or 10-item CATs. Pearson correlation coefficients were calculated to represent the accuracy of the CAT scores compared to those of the full item bank. SEs of the scores were evaluated throughout the range of scores for each CAT to indicate precision. We estimated conditional reliability for scores throughout the range of each scale as $1/(1 + (SE)^2)$. Reliabilities <0.70 were deemed insufficient.²² Floor and ceiling effects were evaluated by identifying participants who endorsed the highest or lowest response category for all items..

Normative Sample: DIF Testing and Score Estimation

We tested for DIF between the normative and claimant sample responses to items in each domain using ordinal logistic regression. After removing items with DIF, Then we estimated scores for the normative sample using weighted maximum likelihood estimation based on the claimant sample calibrations.

Analyses were conducted separately for the claimant and normative samples. We compared scores between claimants and the normative participants by graphically displaying sample score distributions for each scale and creating sample score profiles for 2 cases. In addition, to link the scores for the samples based on US adults for future CAT implementation and score reporting, we normed the items using weighted maximum likelihood estimates for each person in the normative sample.^{21,31}

RESULTS

This study included 1017 participants in the SSA Claimant Calibration study and 999 participants in the Normative Sample Calibration sample. Table 1 displays the background characteristics of the SSA Claimant and Normative Study Samples. The mean age of both samples was 49 years and a majority of participants in both samples were male.

Claimant Sample

Item Calibration and Fit—IRT analyses were conducted separately on the 5 unidimensional SSA-PF domains which were identified from the confirmatory factor analyses. IRT results were as follows: *Changing & Maintaining Body Position* (23 items), *Whole Body Mobility* (16 items), *Upper Body Function* (23 items), *Upper Extremity Fine Motor* (29 items). We also calibrated an 8 item Wheelchair Mobility scale. The sample size of Wheelchair Mobility scale was too small to calibrate in the 2-parameter model,²³ so we applied the Graded Response Model with equal discrimination parameters across the items.

Based on the $p < 0.01$ of $S-X^2$, only 2 items demonstrated significant misfit and were subsequently removed: “Are you able to sew on a button?” and, “Are you able to reach into a low cupboard?” Due to the content similarity between the assistive device items and non-assistive device items, we added 5 assistive device items in the Whole Body Mobility scale. To put those assistive device items on the same metric as the existing items in the scale, we anchored the calibration of the assistive device items on existing item parameters; this resulted in 102 items in the final item bank.

IRT assumes local independence of items, meaning that for respondents with the same level of function, any pair of items are statistically independent of each other. In previous work, local dependent item pairs were identified in the following scales: Changing and Maintaining Body Position (7 dependent item pairs); Whole Body Mobility (2 dependent item pairs); and Upper Body Function: (4 dependent item pairs).¹⁰ To keep a sufficient number of items in each item bank and to prevent over-estimating score precision, we treated dependent item pairs as ‘enemy items’ in the CAT algorithm, which prevents both items from a dependent item pair from being administered for the same person. Because the local dependent items would inflate the item discrimination parameters estimation, we calibrated the local dependent items separately, resulting in two or three sets of calibrations. To avoid capitalizing on chance, we selected the item parameters with the lowest discrimination parameter.²⁴

DIF—There was no significant DIF for age and sex subgroups or for race, education or survey administration method (telephone interview v. internet self-administration). There was one item which showed DIF across the claimant and normative samples: Are you able to reach behind you to get your seatbelt? Hint: A shoulder harness seatbelt). This item was removed when we estimated the normative sample scores.

Psychometric Evaluation: Breadth of coverage, Reliability, Precision, and Accuracy

Table 2 displays the accuracy of simulated 5- and 10-item CATs compared with the full SSA-PF item banks in the claimant sample. Correlations of the 5-item CATs with the total item banks reached or exceeded 0.95 while the correlations for the 10-item banks with the item banks reached or exceeded 0.98. Table 3 displays the breadth of coverage for the 5- and 10-item simulated CATs and the full SSA-PF item banks in each sub-domain. Among claimants, there were minimal floor effects across all sub-domains except for the Upper Body Function scale where 4-7% of the sample was at the floor of the distribution. There were minimal ceiling effects across scales, except for the Upper Extremity Fine Motor scale

where 10-21% of the sample of claimants were at the ceiling of the distribution. Figure 1 provides the distribution of SSA-PF response categories across each scale for each sub-domain and illustrates the relative paucity of items above one standard deviation of the SSA-PF scale in the domain of Upper Extremity Fine Motor Function and the small number of response categories in the lower end of the Upper Body Function sub-domain.

Figures 2a-2e display the distribution of scores for the claimant and normative samples for each of the 5 sub-domain scales along with the reliability values for the full item bank compared with the reliability achieved for the simulated 5- and 10-item CATs. Figures 2a-2e illustrate the dramatic positive shift of the distribution of physical function for the normative sample to the higher end of each scale compared with that seen for the claimant sample. The reliability graphs for each sub-domain scale revealed a high degree of measurement reliability across the middle range of the distribution for each scale. In Whole Body Mobility, the score range of 35-75 revealed a reliability ≥ 0.9 for the full item bank which covers about 93% of the claimant population. The score range of 35-76, 23-61, 29-77 in Upper Body Function, Upper Extremity Fine Motor and Changing & Maintaining Body Position scales achieved reliabilities of ≥ 0.9 for 93%, 86%, 98% of the claimant population, respectively. For the 8-item Wheelchair Mobility scale, the score range of 33-64 achieved a reliability ≥ 0.85 under full item bank which would cover about 87% of the claimant population.

Figure 3a illustrates the functional profile of a 42 year old male claimant who had an upper extremity amputation and compares his functional profile to his age and sex specific norm. The profile shows that while his Upper Body and Upper Extremity Fine Motor functioning was well below the norm, he functioned much closer to his norm in the areas of Changing and Maintaining Body Position and Whole Body Mobility. In contrast, Figure 3b displays the physical function profile of a 56 year old woman with chronic low back pain. In this profile, we see substantial functional limitation in three of the four physical function scales, with the Fine Motor Function scale illustrating a more normal level of function compared to age and sex matched norms for that scale.

DISCUSSION

Developing a standardized multidimensional physical function instrument appropriate for use in the SSA's disability programs is an ambitious yet important goal. IRT/CAT-based methodology is ideally suited for the complexity of assessing a broad range of functional abilities of claimants seeking disability benefits, and overcomes many of the challenges associated with traditional functional assessment measures. A CAT approach allows for assessment of a wide range of functional levels instead of using multiple measures targeting different levels or specific health conditions. The multidimensional structure of the CAT instrument system allows the user to create functional profiles to quickly and easily illustrate the physical functional status of a claimant. A CAT can employ filter questions to select appropriate domains and items that match an individual's sex, living situation, and/or method of locomotion, thus avoiding inappropriate or redundant questions.

The SSA-PF instrument demonstrated strong psychometric properties across all scales in initial testing in a sample of adult claimants who applied to SSA disability programs. High levels of accuracy of simulated 5- and 10-item CATs were found compared to the overall SSA-PF item banks. This study found very little loss of reliability for simulated CATs compared to the full item bank, except at the highest and lowest ranges there were fewer questions.

This study was not designed to specifically address validity of the SSA-PF instrument. Nonetheless, the findings from this study provide preliminary indications of the SSA-PF's content validity. In developing the SSA-PF, we used qualitative methods and wrote new items using feedback from stakeholder interviews, expert panel input and cognitive interviewing.¹¹ We conducted comprehensive factor analyses that revealed a factor structure of the SSA-PF instrument that was consistent with themes revealed in the development of our initial content model.¹⁰ The calibration study demonstrated the reliability of the instrument and allowed us to identify and remove misfitting items and those that displayed substantial DIF. Finally, the substantial group differences in the functional status distribution of a sample of current SSA claimants as compared with a normative sample of US adults provides preliminary evidence supporting the SSA-PF instrument's validity. Research is under way to prospectively examine the validity of the SSA-PF CATs.

Other research on CAT applications in health care research indicate that CAT instruments offer important measurement advantages such as decreased floor and ceiling effects and increased precision across a wider range of the scale.²⁵⁻³⁰ The SSA-PF CAT instrument, unlike other physical function CAT instruments, was developed to comprehensively assess the full range of domains of physical function using content that targets work-relevant functional capabilities.¹¹ CAT instruments such as those developed by the PROMIS network provide a more generic physical function scale.³¹ There is overlap in content and structure of the SSA-PF instrument and the PROMIS physical function domain framework, which includes Upper Extremities (dexterity); Lower Extremities (mobility); Central (twisting, bending, etc.); and Activities: Instrumental Activities of Daily²⁵ subdomains which overlap with the SSA-PF Upper Extremity Fine Motor, Whole Body Mobility, and Upper Body Function scales respectively. However, our item development process revealed specific work-relevant aspects of physical functioning important for inclusion in the SSA-PF instrument, specifically, prolonged, repetitive, and rotational activities as well as aspects of lifting and mobility such as the height at which the task is conducted. The items in the SSA-PF scales were developed to assess physical functioning important for work and are therefore related to, but distinct from, those in the PROMIS physical function scale.

Another important difference between the PROMIS and SSA-PF scales is their structure. Because of the unidimensional structure of the PROMIS scale, one score is used to represent overall function including content from all of the subdomains included in the physical function domain. Research investigating the PROMIS physical function scale has identified differences in physical function for those with upper extremity and lower extremity conditions, and reported results supporting the future development of separate scales for upper and lower extremity function.³² The structure of SSA-PF instrument provides a score for each of 5 scales, allowing the user to obtain a physical function profile for each claimant,

including a separate score for upper extremity function (Upper Extremity Fine Motor). As demonstrated in Figure 3a and 3b, claimant profiles can differ substantially from each other and from the norm in ways that could be informative for the SSA disability determination process.

Since no gold standard exists for determination of work disability, there are tradeoffs in the choice between obtaining functional information using self-report CAT instruments and traditional functional capacity evaluation approaches conducted by clinicians. Because CAT instruments entail relatively minimal administrative burden, we believe that this will enable their use as a complement to current methods. We expect that the combination of currently used medical information and CAT-based functional profiles as well as normed scores will confer the advantages of rich functional as well as medical information.

Looking at the appropriateness of the instrument in a rehabilitation context was beyond the scope of the study, however, given the role of a range of rehabilitation professionals in work disability and vocational rehabilitation this area may hold potential for future research.

Limitations

The results of this study indicated the items fit the model well, and the scales had sufficient reliability and coverage. In the claimant sample, although there were little to no ceiling and floor effects for the other scales, the Upper Body Function scale displayed some gaps at the floor of the distribution and the Upper Extremity Fine Motor scale, with 10-21% of the claimant sample at the ceiling, demonstrated potential limitations in discriminating between the highest levels of function. Although, coverage at the floor and mid-ranges of function would seem critical for informing decision making within SSA disability program applications, we believe that characterization at the higher functional levels is also important. The sample profiles for a claimant with upper extremity amputation and a claimant with low back pain (Figures 3a-b) demonstrate the potential for relatively high functioning in some scales and low functioning in others. We expect that providing information across the full range of functioning will support increased understanding of claimant functioning and decision-making for SSA disability programs.

In our earlier confirmatory factor analyses, we observed medium correlations (average correlation is 0.65) across the factors in SSA-PF domain, the multidimensional CAT could be an alternative for score estimation. It has been shown that multidimensional CATs appear to have both precision and efficiency advantages compared with separate unidimensional CATs and it also could minimize response burden but retain measurement sensitivity.³³⁻³⁵

In order to allow meaningful interpretation of claimant scale scores relative to the function of US adults, we plan to implement the CATs using scores based on the normative sample. Therefore, we normed the scores using weighted maximum likelihood methods.²¹ Although this approach may introduce bias, alternatives such as multiple-group IRT analysis are associated with inflated precision estimates. Based on our experience and that of Rose et al.,³¹ normed score estimates for future SSA-PF CAT applications will be based on the more conservative weighted maximum likelihood methods.

The sample sizes for Wheelchair Mobility scale and walking aid items were relatively small, which could have decreased the stability of the item parameters and increased the variation in person score estimations. Therefore, future work should include adequate representation of these participants.

Although our study found that the SSA-PF scales demonstrated strong psychometric properties, we recognize that validity can never be neatly summarized by one statistic, and that additional validation studies are needed to demonstrate SSA-PF instrument's concurrent and discriminant validity, and test-retest reliability.

CONCLUSIONS

In summary, the SSA-PF instrument contributes important new methodology for measuring the functional status of adults applying for disability benefits. Initial evaluation revealed that the SSA-PF instrument achieved considerable breadth of coverage in each content domain and demonstrated noteworthy psychometric properties. The use of CAT technology minimizes assessment burden while allowing for the comprehensive assessment of the functional abilities of adult claimants, thus providing a feasible methodology for achieving widespread standardized functional assessment into the SSA disability determination process. Our simulation study suggested that with the use of CATs, especially those that consist of 10 items, very little information is lost. Future research needs to assess the SSA-PF instrument's validity and utility for use within the SSA disability program.

Acknowledgments

This research was supported by SSA-NIH Interagency Agreements under NIH Contract # HHSN269200900004C, NIH Contract # HHSN269201000011C, and NIH Contract # HHSN269201100009I and through the NIH intramural research program. Dr. McDonough was funded by a New Investigator Fellowship Training Award from the foundation for Physical Therapy.

ABBREVIATIONS

CAT	Computer adaptive testing or computer-adaptive test
IRT	Item response theory
IOM	Institute of Medicine
PROMIS	Patient-Reported Outcomes Measurement Information System
SSA	Social Security Administration
SSA-PF	Social Security Administration Physical Function Instrument
SE	Standard error
US	United States

REFERENCES

1. Table 8. Annual Statistical Report on the Social Security Disability Insurance Program, 2011, Office of Policy, released July. 2012. Available at http://www.ssa.gov/policy/docs/statcomps/di_asr/2011/index.html

2. The Social Security Administration. § 416.905. Basic definition of disability for adults 45 FR 55621, Aug. 20, 1980, as amended at 56 FR 5553, Feb. 11, 1991; 68 FR 51164, Aug. 26, 2003.
3. Brandt DE, Houtenville AJ, Huynh MT, Chan L, Rasch EK. Connecting contemporary paradigms to the Social Security Administration's Disability Evaluation Process. *Journal of Disability Policy Studies*. Sep; 2011 22(2):116–128. 2011.
4. Nagi, S. Some conceptual issues disability and rehabilitation. In: Sussman, M., editor. *Sociology and rehabilitation*. American Sociology Association; Washington, DC: 1965. p. 100-113.
5. Verbrugge LM, Jette AM. The disablement process. *Social Science & Medicine*. Jan; 1994 38(1):1–14. [PubMed: 8146699]
6. World Health Organization. *International Classification of Functioning, Disability and Health (ICF)*. Geneva: 2001.
7. Social Security Advisory Board. *Charting the Future of Social Security's Disability Programs: The Need for Fundamental Change*. Washington, D.C.: 2001.
8. Institute of Medicine. *Improving the Social Security disability decision process*. The National Academy Press; Washington, DC: 2007.
9. Brandt D, Houtenville H, Huynh M, Chan L, Rasch E. Connecting contemporary paradigms to the Social Security Administration's Disability Evaluation Process. *Journal of Disability Policy Studies*. 2011; 22:116–128.
10. Jette A, Haley S. Contemporary measurement techniques for rehabilitation outcome assessment. *Journal of Rehabilitation Medicine*. 2005; 37:339–345. [PubMed: 16287664]
11. McDonough CM, Jette AM, Ni P, Bogusz K, Marfeo EE, Brandt DE, Chan L, Meterko M, Haley SM, Rasch EK. Development of a Self-Report Physical Function Instrument for Work Disability Assessment: Item Pool Construction and Factor Analysis. *Archives of Physical Medicine and Rehabilitation Medicine*. 2012; XX:XX–XX.
12. Rivers, D. *Sample matching: Representative sampling from internet panels. A white paper on the advantages of the sample matching methodology*. Palo Alto, CA:
13. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969; 17
14. Thissen, D., editor. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Minneapolis, MN: 2009. The MEDPRO project: An SBIR project for a comprehensive IRT and CAT software system—IRT software.
15. Orlando M, Thissen D. New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*. 2000; 24:50–64.
16. Orlando M, Thissen D. Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*. 2003; 27:289–298.
17. Kang T, Chen T. Performance of the Generalized S-X2 Item Fit Index for Polytomous IRT Models. *Journal of Educational Measurement*. 2008; 45(4):391–406.
18. Bjorner, JB.; Smith, KJ.; Stone, C.; Sun, X. IRTFIT: A macro for item fit and local dependence tests under IRT models. *QualityMetric*; Lincoln, RI: 2007.
19. SAS Institute. *SAS users guide, version 9.1*. SAS institute, Inc; Cary, NC: 2004.
20. Jodoin MG, Gierl MJ. Evaluating Type. I error and power rates using an effect size measure with the Logistic Regression Procedure for DIF detection. *Apply Measure Education*. 2001; 14(4):329–349.
21. Warm TA. Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*. 1989; 54:427–450.
22. Mâsse LC, Heesch KC, Eason KE, Wilson M. Evaluating the properties of a stage-specific self-efficacy scale for physical activity using classical test theory, confirmatory factor analysis and item response modeling. *Health Educational Research*. 2006; 21:i33–i46.
23. Thissen D, Reeve BB, Bjorner JB, Chang CH. Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*. 2007; 16:109–116. [PubMed: 17294284]

24. Varni J, Stucky BD, Thissen D, DeWitt EM, Irwin D, Lai JS, Yeatts K, DeWalt DA. PROMIS Pediatric Pain Interference Scale: An Item Response Theory Analysis of the Pediatric Pain Item Bank. *Journal of Pain*. 2010; 11:1109–1119. [PubMed: 20627819]
25. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, DeVellis R, DeWalt D, Fries JF, Gershon R, Hahn EA, Lai J, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R, on behalf of the PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*. Nov; 2010 63(11): 1179–94. [PubMed: 20685078]
26. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai J, Cella D, on behalf of the PROMIS Cooperative Group. Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007; 45:S22–S31. [PubMed: 17443115]
27. Nowinski CJ, Victorson D, C Cavazos JE, Gershon R, Cella D. Neuro-QOL and the NIH Toolbox: implications for epilepsy. *Therapy*. 2010; 7:533–540. [PubMed: 21552344]
28. Haley SM, Siebens H, Coster WJ, Tao W, Black-Schaffer RM, Gandek B, Sinclair SJ, Ni PS. Computerized Adaptive Testing for Follow-Up After Discharge From Inpatient Rehabilitation: I. Activity Outcomes. *Arch Phys Med Rehabil*. 2006; 87:1033–42. [PubMed: 16876547]
29. Tulskey DS, Kisala PA, Victorson D, Tate D, Heinemann AW, Amtmann D, Cella D. Developing a Contemporary Patient-Reported Outcomes Measure for Spinal Cord Injury. *Arch Phys Med Rehabil*. 2011; 92(10 Suppl 1):S44–51. [PubMed: 21958922]
30. Jette AM, McDonough CM, Ni P, Haley SM, Hambleton RK, Olarsch S, Hunter DJ, Felson DT. A functional difficulty and functional pain instrument for hip and knee osteoarthritis research. *Arthritis Research & Therapy*. 2009; 11(4):R107. [PubMed: 19589168]
31. 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). *Journal of Clinical Epidemiology*. 2008; 61:17–33. [PubMed: 18083459]
32. Hung M, Clegg DO, Greene T, Saltzman CL. Evaluation of the PROMIS physical function item bank in orthopaedic patients. *J Orthop Res*. 2011; 29(6):947–53. [PubMed: 21437962]
33. Haley SM, Ni P, Ludlow LH, Fragala-Pinkham MA. Measurement precision and efficiency of multidimensional computer adaptive testing of physical functioning using the pediatric evaluation of disability inventory. *Arch Phys Med Rehabil*. 2006; 87:1223–9. [PubMed: 16935059]
34. Allen DD, Ni P, Haley SM. Efficiency and sensitivity of multidimensional computerized adaptive testing of pediatric physical functioning. *Disabil Rehabil*. 2008; 30:479–84. [PubMed: 18297502]
35. Wang WC, Chen PH. Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Appl Psychol Meas*. 2004; 28:295–316.

Figure 1.

Distribution of Physical Function Items/Categories by Content SubDomain

Figure 2.

a-e Distribution of Physical Function Person Scores and Reliability of 5 item, 10 item, and Full Item Bank by Subdomain for SSA Claimant (N=1017) and Normative (N=999) Samples

a. Changing & Maintaining Body Position

Note: Claimant distribution shown in medium grey, Normative in light grey

b. Whole Body Mobility

Note: Claimant distribution shown in medium grey, Normative in light grey

c. Upper Body Function

Note: Claimant distribution in medium grey, Normative in light grey

d. Upper Extremity Fine Motor

Note: Claimant distribution shown in medium grey, Normative in light grey

e. Wheelchair Mobility

Note: Claimant distribution shown in medium grey, Normative in light grey

Figure 3.
a and b Claimant Functional Profiles Compared to Age and Sex Specific Normative Scores.
a. Profile of a Claimant with an Arm amputation
b. Profile of a Claimant with Chronic Low Back Pain

Table 1
Background Characteristics of the SSA Claimant and Normative Study Samples

Values are expressed as Mean $\pm$ SD or N (%)

Characteristic	SSA Claimants		Normative Sample		
	(N=1017)	%	(N=999)	%	
<i>Age mean $\pm$ SD</i>	49.65 $\pm$ 9.85		49.72 $\pm$ 16.12		t(1616)=-0.12 p=0.9053
<i>Gender</i>			996		
Female	474	46.61	480	48.19	X ² (1)=0.51
Male	543	53.39	516	51.81	p=0.48
<i>Race</i>					
White	597	58.7	782	78.28	
Black/African American	323	31.76	110	11.01	X ² (2)=140
Other	63	6.2	105	10.51	p<0.0001
Missing	34	3.34	2	0.20	
<i>Education</i>					
Less than high school	199	19.57	40	4	
High School/GED	397	39.03	361	36.14	X ² (2)=136.53
Greater than high school	419	41.2	591	59.16	p<0.0001
Missing	2	0.2	7	0.7	

Table 2
Accuracy of Simulated 5- and 10- item Physical Function CATs Using Claimant Data:
Pearson Correlation Coefficients by Content Subdomains

Subdomain	N	5-item CAT	10-item CAT
Changing & Maintaining Body Position	1017	0.95	0.99
Upper Body Function		0.96(1016) *	0.99(1015) *
Upper Extremity Fine Motor	1017	0.95	0.98
Whole Body Mobility		0.96(1004) *	0.99(997) *
Wheelchair Mobility		0.96(88)	-

* Because of missing values, sample sizes varied across content subdomains

Table 3
Breadth of Coverage for simulated 5- and 10-item Physical Function CATs and Full Item Bank based on Claimant Data Presented by Content Subdomain

Subdomain	Mode	N	Mean (SD), Range	%Ceiling	%Floor
Changing and Maintaining Body Position	5-item CAT	1017	50.08(10.43),(16.45,85.55)	0.69%	0.39%
	10-item CAT	1017	50.07(10.24),(11.56,88.57)	0.39%	0%
	Full item bank	1017	50.07(10.38),(11.28,90.04)	0.29%	0%
Whole Body Mobility	5-item CAT	1004	48.5(10.88),(12.77,88.56)	0.7%	2.09%
	10-item CAT	997	48.44(10.85),(9.89,92.48)	0.3%	1.49%
	Full item bank	1017	48.51(10.87),(9.9,92.51)	0.3%	1.39%
Upper Body Function	5-item CAT	1016	50.45(9.73),(29.52,81.42)	0.88%	7.77%
	10-item CAT	1015	50.03(10.22),(27.36,83.27)	0.69%	4.03%
	Full item bank	1017	50.06(10.22),(27.06,83.8)	0.39%	3.64%
Upper Extremity Fine Motor	5-item CAT	1017	49.08(8.81),(19.04,61.37)	21.73%	0.2%
	10-item CAT	1017	49.81(9.76),(17.31,67.52)	10.13%	0.1%
	Full item bank	1017	49.89(9.67),(16.3,67.94)	8.85%	0.1%
Wheelchair Mobility	5-item CAT	88	49.6(10.59),(18.8,79.58)	5.22%	0.87%
	Full item bank	115	50.05(11.11),(18.48,80.36)	4.35%	0.87%