Published in final edited form as:

Qual Life Res. 2017 March ; 26(3): 789–798. doi:10.1007/s11136-016-1477-1.

Improving Measures of Work-Related Physical Functioning

Christine M. McDonough^{1,2}, Pengsheng Ni¹, Kara Peterik¹, Elizabeth E. Marfeo^{1,3}, Molly E. Marino¹, Mark Meterko^{1,4}, Elizabeth K Rasch⁵, Diane E. Brandt⁵, Alan M Jette¹, and Leighton Chan⁵

¹Boston University School of Public Health, Health and Disability Research Institute, Boston, MA

²Geisel School of Medicine at Dartmouth, Department of Orthopaedic Surgery

³Tufts University, Medford, MA

⁴U.S. Department of Veterans Affairs Center for Healthcare Organization and Implementation Research, Boston, MA

⁵National Institutes of Health Rehabilitation Medicine Department, Bethesda, MD

Abstract

Purpose—To expand content of the physical function domain of the Work Disability Functional Assessment Battery (WD-FAB), developed for the US Social Security Administration's (SSA) disability determination process.

Methods—Newly developed questions were administered to 3,532 recent SSA applicants for work disability benefits and 2,025 US adults. Factor analyses and item response theory (IRT) methods were used to calibrate and link the new items to existing WD-FAB, and computer-adaptive test simulations were conducted.

Results—Factor and IRT analyses supported integration of 44 new items into 3 existing WD-FAB scales and the addition of a new 11-item scale (Community Mobility). The final physical function domain consisting of: Basic Mobility (56 items), Upper Body Function (34 items), Fine Motor Function (45 items), and Community Mobility (11 items) demonstrated acceptable psychometric properties.

Conclusions—The WD-FAB offers an important tool for enhancement of work disability determination. The FAB could provide relevant information about work-related functioning for initial assessment of claimants, identifying denied applicants who may benefit from interventions to improve work and health outcomes; enhancing periodic review of work disability beneficiaries; and assessing outcomes for policies, programs and services targeting people with work disability.

Keywords

Disability Evaluation; Work Disability; Questionnaires

Corresponding author: Christine M. McDonough, PhD, PT, MS, Boston University School of Public Health, 715 Albany Street - T5W, Boston, MA 02118, Phone: (617) 638-1982, Fax: (617) 638-1999, cmm@bu.edu.

Conflict of Interest: The authors declare that they have no conflict of interest

INTRODUCTION

Work disability is a major public health problem that is associated with poverty, lack of access to health care, and limitations in other important aspects of social participation[1,2]. The U.S. Social Security Administration's Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) programs are the largest federal source of support for people with disabilities. In 2014, SSA provided payments of over 11.4 billion dollars to just over 10.2 million people[3]. Specific recommendations have been made for improvement of the SSA disability determination process. For example, the National Academy of Medicine Committee on Improving the SSA Disability Determination process has advocated for the use of functional assessment instruments as part of work disability determination[4]. Within SSDI, musculoskeletal and connective tissue disorders were the primary reason for disability for disabled workers and disabled widows in 2014[3], highlighting the importance of physical function in work disability related programs.

In 2011, the Work Disability Functional Assessment Battery (WD-FAB) was created to allow the SSA to collect more systematic and comprehensive information about claimants' functioning. This instrument was developed using item response theory (IRT) methods. Such measures require the creation of a comprehensive set of questions that define each relevant construct -- in this case, physical function domains relevant to one's ability to work. IRT modelling is used to assess the fit of the items on one hierarchical scale, from low to high functioning, and then to assign, or calibrate each item to a location on the scale based on the information it provides. This allows scale score estimation to be determined from any subset of questions, and because they are based on the calibrated item bank, scale scores from various subsets of items are comparable. In addition, computerized adaptive testing (CAT) administration methods can be used to administer an IRT instrument, allowing an algorithm to tailor item selection in real time, selecting the item that will provide the most information at the respondent's estimated functional level.

IRT-based instruments can be updated and expanded, or replenished, by adding new items and calibrating them onto the existing scoring metric[5]. Replenishment is commonly used in educational testing and emerging in health measurement. The initial WD-FAB physical function domain (WD-FAB 1.0) developed in 2011 included 4 scales: Changing and Maintaining Body Position, which includes the ability to assume, maintain and transfer among various positions such as lying, kneeling, sitting, squatting and standing; Whole Body Mobility, which includes the ability to move around from one place to another including crawling, walking and running; Upper Body Function, which entails reaching, lifting, pulling, pushing, and carrying; and Upper Extremity Fine Motor, which includes manipulation of objects requiring dexterity.

Our prior work supported the potential of the WD-FAB scales for measuring functioning for claimants[6–9]. However, because driving and transportation use are important activities relative to work[1], we sought to add content related to this domain. In addition, in the context of SSA application, having more content and items is desirable. Therefore, the purpose of this study was to expand the breadth of content covered in the WD-FAB 1.0 physical function scales. Specifically, we sought to add more difficult items in Upper

Extremity Fine Motor Function, easier items in Upper Body Function, to add more items to other scales, and to provide new items addressing the subdomain of Community Mobility which we defined as driving or using mass transportation to get around one's community.

METHODS

Study procedures were approved by the XXXXXXXX Institutional Review Board and informed consent was obtained from all individual participants included in the study.

Instruments

Developing WD-FAB 2.0—For replenishment of WD-FAB scales, we recruited content experts with expertise in measurement and treatment of physical function limitations and disability to expand content coverage and address identified floor and ceiling effects. The International Classification of Functioning, Disability, and Health (ICF) was used as the theoretical framework for the original scale development and replenishment of the WD-FAB 1.0 physical function scales[6,10,11]. Existing items were coded according to the ICF framework and hierarchically ordered using IRT calibration in the initial field study[6,7].

For the Community Mobility subdomain, content experts provided information on known transportation and driving assessment instruments. A comprehensive literature search was conducted in collaboration with a research librarian in PubMed, CINAHL, Web of Science and Summon including driving and public transportation related search terms, as well as a systematic search of the Instruments field in CINAHL to retrieve measurement instruments listed in published studies. In addition, a comprehensive search of participation measures was conducted to identify functional items specific to driving and public transportation.

The design of new WD-FAB physical functional items was consistent with the existing WD-FAB item bank. Specifically, item stems asked about the respondent's ability to complete an activity, such as "Are you able to walk the aisles of a grocery store using a shopping cart?" Response options include: 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 Community Mobility items consisted of items using both 4 and 5-level agreement response options. The new items were administered to persons with physical limitations during cognitive interviewing to elicit the meaning of the item and the reasons for the response option selected. Cognitive interviews informed revision or deletion of items.

The replenishment design leveraged IRT methods by administering only a subset of existing items –called "anchor items" -- in conjunction with the new items to allow for co-calibration and integration into the WD-FAB 1.0 scales. The final instrument administered in the field calibration study included the following number of new/anchor items, respectively: Whole Body Mobility: 8/12; Changing & Maintaining Body Position: 12/13; Upper Body Function: 20/13; Fine Motor Function: 22/14; and 20 new items for Community Mobility.

Demographics—Demographic characteristics of participants included: age, sex, race, ethnicity, marital status, education level, health condition category, and work history.

Sampling Methods

We recruited 2 samples: 3,532 recent SSA claimants and a general working age sample of 2025 U.S. adults (aged 21–66 years). We randomly selected claimants from a pool of 137,000 applicants who applied for SSA Disability Insurance or Supplemental Security Income programs within a 2 month time period. We randomly sampled claimants after geographic stratification by the 10 SSA regions and by urban/rural status based on address. The research firm Westat, Inc. mailed study invitation letters and consent forms to 16,500 claimants, and then contacted them via the telephone. Inclusion criteria were: age 21 or older as of disability application date, filed an initial application for disability in either the title 2 (SSDI) or title 16 (SSI) program within the last 2 months, resided within the 50 states or Washington D.C, filed the disability claim on his or her own behalf, and ability to read and understand English.

We recruited the general working age sample through an opt-in internet panel of > 1 million people maintained by the survey research firm YouGov, Inc. The proximity matching method developed by YouGov[12] was used to obtain a sample of 2,025 matched to the U.S. population on age, sex, race, ethnicity, US census region urban/rural classifications based on zip code of residence, and education level. We oversampled by race and interviewed 412 African Americans, 268 Asian American and 447 Hispanic/Latino Americans to allow analysis of response patterns by racial/ethnic groups.

Data collection Methods—For the claimant sample, a trained interviewer administered the instrument by phone, or, if preferred, the claimant self-administered the instrument via the internet. Procedures were monitored for quality by Westat via direct observation of selected interviews. The general working age sample self-administered the instrument via the internet.

Analytic Methods

The analysis was first conducted using the newly developed items along with anchor items from the WD-FAB 1.0 item banks to enable co-calibration in the claimant sample and then replicated in the U.S. general sample to allow score comparisons between claimants and US adults. For these analyses, we used data from claimants who reported a physical or physical and mental health conditions. We calculated summary statistics for demographic variables. “I don’t know” responses were treated as missing responses and frequency for these responses was calculated for each item.

Examination of WD-FAB Structure—We conducted exploratory Factor Analyses (EFA) to examine the structure of the new items within the framework of our existing WD-FAB scales and the ICF-based content model. We compared factor loading patterns and decided on the number of factors by balancing the interpretability of the factor content with model parsimony. We considered items loadings of greater than 0.30 sufficient to retain the item under investigation and used the higher loading to designate factor assignment for dual-loading items[13]. In those instances where dual loadings were of similar magnitude, the content model was used to guide item categorization. We employed confirmatory factor analyses (CFAs) to test unidimensionality for each factor. We determined acceptable model

fit as $RMSEA \leq 0.08$, CFI & TLI ≥ 0.90 [14–16]. Item content was considered in removing items with the goal of balancing statistical fit with parsimony and adequate content coverage.

Item Calibration, Fit and CAT development—Responses were coded so that higher scores represent high functioning. The Graded Response Model [17] was used to calibrate the items. Item fit was assessed by examining the difference between the expected and observed number of subjects in each category at summed raw score level using Pearson's Chi-Square ($s-\chi^2$) [18], and misfit items were identified using the Bonferroni correction p-value. In addition the item discrimination parameter was used to evaluate items for inclusion. IRTPRO software was used to perform marginal maximum likelihood procedure for item calibration [19].

CAT algorithms were created for each of the scales using weighted likelihood estimation to estimate the score and standard error. We programmed the algorithm to select the initial item at mid-level difficulty, calculate the score estimate, administer subsequent items with the optimal information yield for that score, and then recalculate the score based on the subsequent response. The stopping rule used in the algorithm required a minimum of 5 items, maximum of 10, and reliability ≥ 0.90 . Summary scores were transformed to T-scores with mean=50, SD=10, with lower scores indicating lower function.

Differential Item Functioning—Differential Item Functioning (DIF) occurs when people at the same estimated ability or functional level in a particular content domain respond differently to the same item based on some other variable. We examined DIF in three ways. We tested for DIF in this sample by age (age (<55 vs. 55), gender, and race/ethnicity (White vs. African American, Asian American and non-white Hispanic Latino). We also used DIF analysis to examine whether there were differences in response patterns for the claimant sample compared to the general population sample to allow for calibration of the items across the two samples. Finally, DIF analyses were conducted of the newly developed items compared to the WD-FAB's existing items for concurrent calibration of the items.

We employed IRT methods to analyze DIF. This involved a two-step DIF analysis procedure. First, we applied Langer's method [20–22] to screen for DIF and non-DIF items. Then we examined the difference in the Item Characteristic Curves (ICCs) using the weighted Area Between the expected score Curves (wABC) to examine DIF impact [23]. Five-category items with $wABC > 0.3$ and 4-category items with $wABC > 0.24$ were identified as demonstrating DIF. Items that demonstrated DIF were considered carefully for removal or retention with separate calibrations specific to the relevant demographic subgroup.

Linking the WD-FAB 2.0 Items to the WD-FAB 1.0 Scales—We used concurrent calibration methods to estimate the new item parameters onto each existing WD-FAB 1.0 scale. We applied a two-group IRT model and used the original WD-FAB sample as the reference population (with mean=0 and standard deviation=1). After that, the new item parameters were positioned on the original scale. We applied separate calibrations with the Stocking-Lord method to link the original WD-FAB scales onto general population scales.

Psychometric Evaluation of the WD-FAB 2.0—To examine the properties of the WD-FAB 2.0 physical function scales, distribution of scores for claimant and working-age adult samples were examined, and percent at the lowest and highest scores were calculated. Reliability of scores from responses to the full item bank were calculated as (score variance-average of squared standard error)/score variance[24]. Simulated CAT scores were created by using the CAT algorithm to select the first item, and using claimant responses to questions provided for each item selected by the algorithm. Reliability of simulated CAT scores were compared to the reliability for the full item bank.

RESULTS

The claimant sample included 3,532 adults who had recently applied for SSA disability benefits (Table 1); 55% were female and the mean age was 48.3 (SD 11.2) years. There were 2,025 people in the general working age sample; 55% were female and the mean age was 42.2 (SD 12.8). The mean (SD) for “I don’t know” responses was 1.09% (SD: 1.11%) for the claimant sample and 2.16% (SD: 1.34%) for the general sample.

WD-FAB Structure

The structure supported by the EFA and CFA analyses included 4 domains: Basic Mobility, Upper Body Function, Fine Motor Function, and Community Mobility. Whereas the previous WD-FAB structure contained a separate domain for Whole Body Mobility (e.g. walking, running, and climbing) and one for Changing & Maintaining Body Position (e.g. kneeling, squatting, and standing), the results of this analysis supported a more parsimonious structure by including these items into one Basic Mobility domain defined as the ability to assume, maintain and transfer among various body positions and the ability to move around from one place to another. The structure that emerged in these analyses retained separate Upper Body Function and Fine Motor Function domains as defined in the original WD-FAB. We found a distinct domain representing Community Mobility defined as using transportation, including public transportation and driving. The fit statistics for CFAs met our criteria: RMSEA $\leq$ 0.08, CFI & TLI $\geq$ 0.90 (Table 2).

Figure 1 summarizes the flow of new and anchor items into the analysis, and the final number of items in each subdomain.

Item Calibration, Fit, Linking and CAT

After we examined the WD-FAB structure and removed misfitting items, we added 14, 13 and 17 new items into the original Basic Mobility, Upper Body Function and Upper Extremity Fine Motor item banks respectively. We identified 16, 8 and 11 common items, in the domains noted, respectively, from the prior and current claimant samples; and, conducted the concurrent calibration to put the new WD-FAB items onto the original WD-FAB scale. We calculated the linking coefficients based on identified 26, 16 and 22 common items across the claimant sample and general working age sample, and linked the original scale onto the general working age scale. We transformed the general working age scale onto the T-scale with mean = 50 and standard deviation =10, where a higher score reflects better functioning. Item parameters (Appendix A) and the number of items at each analytic stage

for each WD-FAB scales (Appendix B) are provided for each scale as supplemental material.

IRT analyses supported integration of 44 new items into 3 revised scales: Basic Mobility (56 items), Upper Body Function (34 items), Fine Motor Function (45 items), and the addition of a new scale: Community Mobility (11 items). The Community Mobility scale consists of a filter question to identify the mode(s) of transportation the respondent uses, so that the driving items will be administered only to those who drive, and the public transportation items only to those who use public transportation.

DIF analyses

In the Basic Mobility scale there were 3 items exhibiting DIF by gender, 1 by race, and 1 by age. We found DIF by gender for 1 item in the Upper Body Function scale and 4 items in the Fine Motor Function scale. One Basic Mobility item exhibited DIF in the in response patterns for the WD-FAB 1.0 and WD-FAB 2.0 items, and this item was removed. The items exhibiting DIF were retained based on content coverage considerations with different calibration values.

Psychometric Evaluation: Breadth of Coverage, Test Information and Reliability

Table 3 provides a summary of claimant score distributions, percent at the lowest (floor) and highest (ceiling) scores, and marginal reliability for the full item banks and simulated CATs. Scores were transformed to T-scores with mean=50, SD=10 based on the general working age sample's score distribution, with lower scores indicating lower function. The claimant sample demonstrated scores for all scales centered at greater than one standard deviation below the mean of the general working age sample. Ceiling effects were less than 15% for all scales except for the Driving scale in Community Mobility, for which 77.55% of claimants were at the highest possible score, indicating no difficulty.

Figures 2 illustrates the distributions of the expanded scales in the claimant and general working age samples.

Simulated CAT reliability ranged from $r=0.76$ for the Public Transportation scale of Community Mobility to $r=0.93$ for the Upper Body Function scale. Calculation of reliability for the Driving scale in Community Mobility was not possible because of the low variance in the claimant sample.

DISCUSSION

Accuracy in measuring function related to one's ability to work is central to public confidence in a work disability benefits system. In the U.S., national disability programs are challenged to adjudicate millions of work disability claims each year in a timely and accurate manner. The WD-FAB was developed to provide the SSA and other interested parties a comprehensive and efficient approach to profiling a person's function related to their ability to work. The WD-FAB is grounded by a conceptual framework, enhanced by a thorough literature review and expert consultation, and constructed using contemporary IRT methods to yield an instrument that can be administered using CAT techniques.

The current study expanded the original WD-FAB to increase its scope and coverage of a wide range of physical functioning. A new Community Mobility subdomain was added to the WD-FAB based on the respondent's use of public transportation and driving a car. The number of items in each scale have been substantially increased, allowing more comprehensive coverage across the range of work-relevant physical functioning. These scales can be used to create multidimensional functional profiles of persons applying for work disability benefits and can be compared to scores for a general U.S. working age adults. Such profiles could be used to inform work disability adjudication decisions, treatment planning for rehabilitation and subsequent return to work decisions, as well as identifying work environments where work requirements match the functional profile of the individual.

This field calibration study of 3,532 recent claimants and 2,025 working-age adults demonstrated that IRT methods can be used to expand existing measures of functioning to improve their psychometric and conceptual properties. These methods allowed for the expansion of content at the floor for the Upper Body Function and at the ceiling for the Fine Motor Function scale. At the same time, new content was added to the combined Basic Mobility scale. Finally, the replenishment study allowed for the development of a completely new scale to address the important area of functioning related to driving and use of public transportation. The initial properties of the scales are promising for systematic measurement of work-relevant functioning. Although future research is needed to further support various aspects of validity, the scales demonstrated high reliability and minimal ceiling and floor effects.

There are limitations to the new community mobility scale that need to be addressed. Because so many claimants reported no difficulty in driving, there was little variance in these scores. Therefore, this scale will be of little use at the current time for assessment of a claimant's functional status. One approach to addressing this limitation would be to use a filter question to ascertain any difficulty in driving to trigger administration of the driving scale. We believe that Community Mobility will be a useful scale for the assessment of people with work disability with this additional screening question. Our review of the literature in this area found very few instruments that measured functioning related to driving and transportation use. There were comprehensive performance-based driving tests for use by clinicians, and the vision-focused Driving Habits questionnaire, which did not address functioning related to community mobility for a broader population. Although some existing participation scales do address driving and/or public transportation use, they typically use one global question and integrate the community mobility information into a broader score. Future research will be required to investigate the reliability and validity of the Community Mobility scale with the proposed filter.

Results of our DIF analyses indicated that very few items demonstrated differential response associated with age, race, ethnicity or gender. This is a noteworthy contribution based on our a priori over-sampling of minority respondents. The results indicate that the large majority of items may be administered across racial groups without concern for differential interpretation. In cases where DIF was found, the decision to remove or to retain the item

with calibration adjustment depended on the content of each item. In this study, most items were retained using specified calibration values for different subgroups displaying DIF.

CONCLUSIONS

The WD-FAB's Basic Mobility, Upper Body Function, and Fine Motor physical function scales provide a systematic and psychometrically sound approach to profiling a person's physical functioning related to their ability to work. The WD-FAB is grounded by a conceptual framework, enhanced by a thorough literature review and expert consultation, and constructed using contemporary IRT methods, and can be efficiently administered using CAT techniques.

To support public health, clinical and occupational research and practice, there is a critical need for comprehensive functional assessment instruments that are not disease specific and can cover a range of content and be improved over time as relevant content changes. IRT methods, existing scales were successfully expanded and an additional scale was developed to fill gaps in content coverage.

The FAB could provide relevant information about work-related functioning for initial assessment of claimants, identifying denied applicants who may benefit from interventions to improve work and health outcomes; enhancing periodic review of work disability beneficiaries; and assessing outcomes for policies, programs and services targeting people with work disability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Acknowledgement:

This study was supported by Social Security Administration-National Institutes of Health Interagency Agreements under the National Institutes of Health (contract nos. HHSN269200900004C, HHSN269201000011C, HHSN269201100009I, HHSN269201200005C), and by the National Institutes of Health Intramural Research Program.

References

1. Institute of Medicine (IOM). The Future of Disability in America. Washington, DC: The National Academies Press; 2007.
2. Institute on Disability. 2015 Annual Disability Statistics Compendium. Durham, NH, USA: University of New Hampshire; 2016.
3. Social Security Administration. Annual statistical report on the Social Security Disability Insurance program. 2014. [Accessed April 21, 2016]
4. IOM (Institute of Medicine). Improving the Social Security disability decision process. Washington, D.C: 2007.
5. Haley S, Ni P, Jette A, et al. Replenishing a computerized adaptive test of patient-reported daily activity functioning. *Quality of Life Research*. 2009; 18(4):461–471. [PubMed: 19288222]

6. McDonough CM, Jette AM, Ni P, et al. Development of a self-report physical function instrument for disability assessment: item pool construction and factor analysis. *Arch Phys Med Rehabil.* 2013; 94(9):1653–1660. [PubMed: 23542402]
7. Ni PS, McDonough CMJA, Bogusz K, Marfeo EE, Rasch EK, Brandt DE, Meterko M, Haley SM, Chan L. Development of a computer-adaptive physical function instrument for Social Security Administration disability determination. *Arch Phys Med Rehabil.* 2013; 94:1661–1669. [PubMed: 23578594]
8. Marino ME, Meterko M, Marfeo EE, et al. Work-related measures of physical and behavioral health function: Test-retest reliability. *Disability and health journal.* 2015; 8(4):652–657. [PubMed: 25991419]
9. Meterko M, Marfeo EE, McDonough CM, et al. Work disability functional assessment battery: feasibility and psychometric properties. *Arch Phys Med Rehabil.* 2015; 96(6):1028–1035. [PubMed: 25528263]
10. Marfeo EEHS, Jette AM, Eisen SE, et al. A conceptual foundation for measures of physical function and behavioral health function for Social Security work disability evaluation. *Arch Phys Med Rehabil.* 2013; 94:1645–1652. [PubMed: 23548543]
11. World Health Organization. *International Classification of Functioning, Disability and Health (ICF).* Geneva: 2001.
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. Thompson, B. *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association; 2004.
14. Hu, L., Bentler, PM., editors. *Evaluating Model Fit.* Thousand Oaks, CA: Sage Publications; 1995. Hoyle, RH., editor. *Structural Equation Modeling: Concepts, Issues and Applications.*
15. Hu LT, Bentler P. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling.* 1999; 6:1–55.
16. Steiger JH. Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research.* 1990; 25:173–180. [PubMed: 26794479]
17. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr.* 1969; 17(Suppl):1–100.
18. 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.
19. Thissen, D. The MEDPRO project: an SBIR project for a comprehensive IRT and CAT software system-IRT software. Paper presented at: GMAC Conference on Computerized Adaptive Testing; 2009; Minneapolis, MN.
20. IRTPRO. Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. [computer program]. Lincolnwood, IL: Scientific Software International; 2011.
21. Langer, M. A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation. Chapel Hill: University of North Carolina; 2008.
22. Woods C, Cai L, Wang M. The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-Group IRT. *Educational and Psychological Measurement.* 2013; 73:532–547.
23. Edelen M, Stucky B, Chandra A. Quantifying 'problematic' DIF within an IRT framework: application to a cancer stigma index. *Qual Life Res.* 2015; 24(1):95–103. [PubMed: 24214177]
24. Green BF, Bock RD, Humphreys LG, Linn RL, Reckase MD. Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement.* 1984; 21:347–360.

Appendix A. Item Parameters

Notes:

Slope = item discrimination parameter

B1,...,B4 = item threshold parameters. The number of threshold parameters equal to number of categories minus 1. For the Community Mobility Driving scale, because of the skewed distribution of the sample distribution across response categories, we merged the first with the second and fourth with the fifth categories. Therefore the number of categories was decreased from 5 to 3, and the number of threshold parameters was decreased from 4 to 2.

DIF: For example, item 29 in basic mobility was indicated as gender DIF item. We will use 29.a item parameters for male subjects and 29.b item parameters for female subjects.

Basic Mobility						
Item	SLOPE	B1	B2	B3	B4	DIF
1	2.69839	-3.00297	-2.03314	-1.19328	-0.61392	
2	2.07909	-1.42781	-0.79169	-0.049	0.65299	
3	2.21365	-2.18301	-1.57384	-0.68832	0.04545	
4	2.15493	-1.17014	-0.73944	0.03101	0.59952	
5	2.83069	-2.40643	-2.03305	-1.19678	-0.50432	
6	2.18294	-1.58873	-1.12415	-0.36146	0.43503	
7	2.45354	-2.18765	-1.76049	-0.97799	-0.36308	
8	3.21552	-2.09218	-1.6786	-0.96784	-0.35111	
9	3.14384	-2.88461	-2.06843	-1.2396	-0.56428	
10	3.00161	-2.29532	-1.80792	-1.05407	-0.41944	
11	2.54771	-2.16396	-1.32446	-0.48185	0.22171	
12	2.29622	-2.67268	-1.81314	-0.9496	-0.24614	
13	2.78885	-2.251	-1.78347	-0.99322	-0.36303	
14	3.53963	-2.42193	-1.64425	-0.84366	-0.21233	
15	2.58498	-1.85718	-1.43893	-0.73261	-0.04082	
16	2.9422	-1.98268	-1.72715	-1.06136	-0.31741	
17	1.61652	-2.37274	-1.17566	-0.05998	0.77565	
18	2.17787	-2.26651	-1.653	-0.82386	-0.04333	
19	2.10116	-2.12273	-1.65058	-0.84678	-0.12832	
20	2.50785	-2.43802	-1.62465	-0.82788	-0.11469	
21	3.53013	-2.58328	-1.81369	-0.98891	-0.35002	
22	2.43416	-1.3257	-0.77658	-0.03347	0.62329	
23	1.76543	-1.5118	-0.68871	0.14829	0.91753	
24	3.87464	-2.20525	-1.69476	-0.94637	-0.31936	
25	3.09107	-2.11934	-1.87885	-1.1966	-0.54712	
26	2.68848	-2.52118	-1.93365	-1.02775	-0.26809	
27	3.03729	-1.83151	-1.42702	-0.82515	-0.25644	
28	3.00097	-1.93477	-1.27446	-0.5585	0.14326	
29.a	1.78484	-1.91641	-1.21665	-0.27865	0.59254	male
29.b	1.58167	-1.90307	-0.99246	0.03467	0.839	female

Basic Mobility						
Item	SLOPE	B1	B2	B3	B4	DIF
30	2.07069	-2.50323	-1.76225	-0.82988	-0.00504	
31	2.36114	-1.94777	-1.31894	-0.63895	0.03634	
32.a	2.6053	-1.7017	-1.16406	-0.48366	0.24725	male
32.b	2.03731	-1.33538	-0.78054	-0.00861	0.68254	female
33	2.93551	-2.72081	-1.99145	-1.2203	-0.54272	
34	3.16454	-2.39522	-1.75207	-1.02041	-0.3818	
35	2.44649	-2.94052	-2.04048	-1.14579	-0.40858	
36	3.14193	-2.40369	-1.7196	-0.91755	-0.25258	
37	3.74431	-2.26427	-1.71115	-1.02962	-0.43129	
38	2.81963	-2.22481	-1.42326	-0.66462	0.02829	
39.a	2.53111	-2.36985	-1.59276	-0.66698	0.10438	male
39.b	2.0748	-1.92704	-0.97082	-0.00372	0.81386	female
40	1.99048	-2.84283	-1.94219	-0.90534	-0.11753	
41	3.62219	-2.46485	-1.96277	-1.18914	-0.50544	
42	3.21136	-2.73874	-2.14831	-1.42679	-0.7327	
43	3.21377	-2.38061	-1.67807	-0.87804	-0.20335	
44	3.16069	-2.21285	-1.45482	-0.70317	-0.04068	
45.a	2.88185	-2.13365	-1.50994	-0.83484	-0.07254	male
45.b	2.1394	-1.82457	-1.12557	-0.35524	0.34974	female
Upper Body Function						
Item	SLOPE	B1	B2	B3	B4	DIF
1	2.30362	-0.96759	-0.08485	0.84003	1.61716	
2	1.99523	-1.29983	-0.4545	0.45632	1.23827	
3	2.42742	-1.03238	-0.50877	0.24058	0.88908	
4	3.02645	0.62167	1.14517	1.66805	2.21399	
5	3.23787	-0.82883	-0.37767	0.28466	0.88402	
6	2.88737	0.1673	0.87294	1.51428	2.13348	
7	2.84091	-0.89254	-0.30564	0.35647	0.96903	
8	2.68337	-0.00307	0.53353	1.08526	1.56422	
9	2.08301	-0.7841	-0.21812	0.53603	1.13279	
10	2.51894	0.08606	0.84756	1.53176	2.21784	
11	2.68031	-0.8464	-0.3578	0.34963	0.97025	
12	3.57733	-0.27615	0.2446	0.86274	1.49964	
13	3.44454	-0.65559	-0.1079	0.57411	1.20374	
14	2.61166	-0.58043	-0.19995	0.39544	0.99458	
15	2.40955	0.23264	1.0335	1.7314	2.36776	
16	3.30835	-0.05802	0.3931	0.95297	1.55754	
17	3.00948	0.15326	0.8179	1.42089	1.96462	

Upper Body Function

Item	SLOPE	B1	B2	B3	B4	DIF
18	3.0347	0.43194	1.08622	1.66981	2.14975	
19	3.36489	0.35114	0.86566	1.42126	1.99599	
20	2.5778	0.42424	1.00089	1.49161	2.02095	
21	3.38507	-0.27353	0.13472	0.72634	1.32594	
22	2.83488	-0.9331	-0.14896	0.58895	1.25279	
23	1.86755	-2.9382	-1.73373	-0.62655	0.20129	
24	2.56831	-1.42253	-0.88195	-0.04684	0.64142	
25	2.758	-1.03257	-0.36803	0.33618	1.00597	
26	3.08212	-0.7685	0.07858	0.84112	1.48685	
27	3.00168	-0.06693	0.7518	1.48367	2.18689	
28	2.5144	-0.86256	-0.19785	0.52189	1.22565	
29	3.37132	-0.69643	-0.15289	0.50411	1.16454	
30	3.26859	-0.85351	-0.27479	0.39859	1.04119	
31	3.73751	-0.65599	0.10285	0.82904	1.4577	
32.a	3.39322	-0.90041	-0.27931	0.43485	1.11038	male
32.b	3.03868	-1.2588	-0.62005	0.13615	0.78611	female
33	2.78047	0.05252	0.77401	1.40216	2.07392	
34	2.60944	-1.38582	-0.54428	0.34305	1.02165	

Upper Extremity Fine Motor

Item	SLOPE	B1	B2	B3	B4	DIF
1	2.69839	-3.00297	-2.03314	-1.19328	-0.61392	
2	2.07909	-1.42781	-0.79169	-0.049	0.65299	
3	2.21365	-2.18301	-1.57384	-0.68832	0.04545	
4	2.15493	-1.17014	-0.73944	0.03101	0.59952	
5	2.83069	-2.40643	-2.03305	-1.19678	-0.50432	
6	2.18294	-1.58873	-1.12415	-0.36146	0.43503	
7	2.45354	-2.18765	-1.76049	-0.97799	-0.36308	
8	3.21552	-2.09218	-1.6786	-0.96784	-0.35111	
9	3.14384	-2.88461	-2.06843	-1.2396	-0.56428	
10	3.00161	-2.29532	-1.80792	-1.05407	-0.41944	
11	2.54771	-2.16396	-1.32446	-0.48185	0.22171	
12	2.29622	-2.67268	-1.81314	-0.9496	-0.24614	
13	2.78885	-2.251	-1.78347	-0.99322	-0.36303	
14	3.53963	-2.42193	-1.64425	-0.84366	-0.21233	
15	2.58498	-1.85718	-1.43893	-0.73261	-0.04082	
16	2.9422	-1.98268	-1.72715	-1.06136	-0.31741	
17	1.61652	-2.37274	-1.17566	-0.05998	0.77565	
18	2.17787	-2.26651	-1.653	-0.82386	-0.04333	
19	2.10116	-2.12273	-1.65058	-0.84678	-0.12832	

Upper Extremity Fine Motor

Item	SLOPE	B1	B2	B3	B4	DIF
20	2.50785	-2.43802	-1.62465	-0.82788	-0.11469	
21	3.53013	-2.58328	-1.81369	-0.98891	-0.35002	
22	2.43416	-1.3257	-0.77658	-0.03347	0.62329	
23	1.76543	-1.5118	-0.68871	0.14829	0.91753	
24	3.87464	-2.20525	-1.69476	-0.94637	-0.31936	
25	3.09107	-2.11934	-1.87885	-1.1966	-0.54712	
26	2.68848	-2.52118	-1.93365	-1.02775	-0.26809	
27	3.03729	-1.83151	-1.42702	-0.82515	-0.25644	
28	3.00097	-1.93477	-1.27446	-0.5585	0.14326	
29.a	1.78484	-1.91641	-1.21665	-0.27865	0.59254	male
29.b	1.58167	-1.90307	-0.99246	0.03467	0.839	female
30	2.07069	-2.50323	-1.76225	-0.82988	-0.00504	
31	2.36114	-1.94777	-1.31894	-0.63895	0.03634	
32.a	2.6053	-1.7017	-1.16406	-0.48366	0.24725	male
32.b	2.03731	-1.33538	-0.78054	-0.00861	0.68254	female
33	2.93551	-2.72081	-1.99145	-1.2203	-0.54272	
34	3.16454	-2.39522	-1.75207	-1.02041	-0.3818	
35	2.44649	-2.94052	-2.04048	-1.14579	-0.40858	
36	3.14193	-2.40369	-1.7196	-0.91755	-0.25258	
37	3.74431	-2.26427	-1.71115	-1.02962	-0.43129	
38	2.81963	-2.22481	-1.42326	-0.66462	0.02829	
39.a	2.53111	-2.36985	-1.59276	-0.66698	0.10438	male
39.b	2.0748	-1.92704	-0.97082	-0.00372	0.81386	female
40	1.99048	-2.84283	-1.94219	-0.90534	-0.11753	
41	3.62219	-2.46485	-1.96277	-1.18914	-0.50544	
42	3.21136	-2.73874	-2.14831	-1.42679	-0.7327	
43	3.21377	-2.38061	-1.67807	-0.87804	-0.20335	
44	3.16069	-2.21285	-1.45482	-0.70317	-0.04068	
45.a	2.88185	-2.13365	-1.50994	-0.83484	-0.07254	male
45.b	2.1394	-1.82457	-1.12557	-0.35524	0.34974	female

Community Mobility Driving

Item	SLOPE	B1	B2
1	1.41586	-3.64186	-2.96763
2	2.94184	-2.27532	-1.36466
3	3.2332	-2.70606	-1.72574
4	3.96116	-2.88246	-2.01754
5	2.1034	-2.38854	-1.45034

Community Mobility Public Transportation

Item	SLOPE	B1	B2	B3	B4
1	1.26799	-3.39656	-2.20991	-1.30148	1.06968
2	3.09302	-2.09553	-1.36414	-0.60617	0.14678
3	2.14687	-1.89949	-0.9949	-0.13357	0.6907
4	3.37969	-2.49808	-1.56445	-0.80958	-0.16306
5	2.40906	-2.72233	-1.75852	-0.88794	-0.22973
6	1.8949	-3.09564	-2.08561	-1.15934	-0.44957

Appendix B. The number of items at each analytic stage for each WD-FAB scales

Claimant Calibration				
	Basic Mobility	Upper Body Function	Upper Extremity Fine Motor	Community Mobility
Items Fielded	45	33	36	20
Factor analysis				
Items removed	14	11	1	9
Items retained	31	22	35	11
IRT analysis for FAB 2.0 data				
Items removed	1 (DIF between FAB 1.0 and FAB 2.0 samples)	0	4 (misfit)	0
Items Retained	30	22	31	11
Linking analysis (between FAB1.0 and FAB2.0)				-
Anchor items	16	9	14	
Items Removed	1 ^a (Age DIF)	1	3 ^b (Gender DIF)	-
Number of Anchor Items Retained	15	8	11	-
Final FAB 2.0				
New items	14	13	17	-
Old items	27	12	17	-
Total	56	34	45	11

^a Removed as an anchor item from linking analysis and retained in the item bank as “Old” item with separate calibrations by gender.

^b Removed as anchor items in linking analysis and retained in the item banks as “Old” items with no gender DIF, based on FAB 1.0 scale.

Working Age Adult Calibration				
Analytic Stage	Basic Mobility	Upper Body Function	Upper Extremity Fine Motor	Community Mobility
Items Fielded	45	33	36	20
Factor Analysis				
Items Removed	14	0	0	5
Items Retained	31	33	36	15
IRT analysis				
Items Removed	0	5 (misfit)	1 (misfit)	0
Items Retained	31	28	35	15
Linking analysis (between Claimant sample and working age sample)				
Number of anchor items	28	16	24	11
Items Removed	2 (DIF between claimant sample and general working age sample)	0	2 (DIF between claimant sample and general working age sample)	0
Items retained	26	16	22	11
Final FAB 2.0 (Working Age Adult)				
New items	14	18	21	15
Old items	17	10	14	0
Total	31	28	35	15

Figure 1.
Flow of WD-FAB Item Wining by Subdomain

Figure 2. Figures 2a, b, c, d, e. Score distributions of claimant and general US working age adult samples

2a Basic Mobility

2b. Upper Body Function

2c Fine Motor Function

2d. Community Mobility (Using Public Transportation)

2e. Community Mobility (Driving)

Table 1

Characteristics of the General Working Age and Claimant Samples for Replenishment of the WD-FAB Physical Function Scales.

	SSA Claimants (n=3,532)		General Working Age Sample (n=2025)	
Characteristic	N	%	N	%
<i>Age mean $\pm$ SD</i>	48.29(11.18)		42.19(12.76)	
Under 40	823	23.30	913	45.09
40–55	1426	40.37	654	32.30
>55	1276	36.13	458	22.62
missing	7	0.20	0	0
<i>Sex</i>				
Female	1927	54.56	1113	54.96
missing	8	0.23	0	0
<i>Race</i>				
White	2241	63.45	1083	53.48
Black/African American	900	25.48	412	20.35
Other	355	10.05	483	23.85
missing	36	1	47	2.32
<i>Hispanic Ethnicity</i>				
Yes	263	7.45	447	22.07
Refuse	37	1.04	17	0.84
<i>Education</i>				
Less than high school	538	15.23	81	4
High School/GED	1123	31.80	636	31.41
Greater than high school	1863	52.75	1302	64.29
missing	8	0.23	6	0.3
<i>Conditions warranting consulting provider</i>				
Physical Impairments	2025	53.39		
Physical & Mental Impairments	1507	39.73		

Table 2

Confirmatory Factor Analysis Fit Statistics

Subdomain	Claimant Sample			Working-Age Adult Sample		
	CFI	TLI	RMSEA	CFI	TLI	RMSEA
Basic Mobility	0.95	0.95	0.08	0.99	0.99	0.05
Upper Body Function	0.98	0.98	0.07	0.99	0.98	0.05
Fine Motor Function	0.96	0.95	0.07	0.99	0.99	0.03
Community Mobility	0.95	0.95	0.05	0.94	0.94	0.08

Table 3

Descriptive characteristics for the WD-FAB physical function scales by domain for the claimant sample.

	Mode	N	Mean (SD)	Reliability	% Ceiling Mean (SD)%	% Floor Mean (SD)%
Basic Mobility	5–10 item CAT	2979	37.03(5.16)	0.92	40(1.34%)	5(0.17%)
	Full item bank	2979	37.02(5.20)	0.96	39(1.31%)	4(0.13%)
Upper Body Function	5–10 item CAT	2975	35.90(5.56)	0.93	81(2.72%)	7(0.24%)
	Full item bank	2975	35.86(5.52)	0.96	80(2.69%)	7(0.24%)
Fine Motor Function	5–10 item CAT	2979	38.64(6.47)	0.86	360(12.08%)	4(0.13%)
	Full item bank	2979	38.48(6.60)	0.90	348(11.68%)	4(0.13%)
Community Mobility(driving)	Full item bank	2205	33.84(1.70)	N/A ^a	1710(77.55%)	2(0.09%)
Community Mobility (public transportation)	Full item bank	653	41.23(5.96)	0.76	70(10.72%)	4(0.61%)

^aReliability for the driving scale of Community Mobility was not calculated due to limited variance in the data.