J Rehabil Med 2015; 47: 394–402

ORIGINAL REPORT

INTERPRETING PHYSICAL AND BEHAVIORAL HEALTH SCORES FROM NEW WORK DISABILITY INSTRUMENTS

Elizabeth E. Marfeo, PhD, MPH, OTR/L[1], Pengsheng Ni, MD, MPH[1], Leighton Chan, MD, MPH[2], Elizabeth K. Rasch, PhD, PT[2], Christine M. McDonough, PhD[1], Diane E. Brandt, PT, MS, PhD[2], Kara Bogusz, BA[1] and Alan M. Jette, PhD, PT[1]

From the [1]Boston University School of Public Health, Health & Disability Research Institute, Boston, MA, and [2]National Institutes of Health, Rehabilitation Medicine Department, Mark O. Hatfield Clinical Research Center, Bethesda, MD, USA

Objective: **To develop a system to guide interpretation of scores generated from 2 new instruments measuring work-related physical and behavioral health functioning (Work Disability – Physical Function (WD-PF) and WD – Behavioral Function (WD-BH)).**

Design: **Cross-sectional, secondary data from 3 independent samples to develop and validate the functional levels for physical and behavioral health functioning.**

Subjects: **Physical group: 999 general adult subjects, 1,017 disability applicants and 497 work-disabled subjects. Behavioral health group: 1,000 general adult subjects, 1,015 disability applicants and 476 work-disabled subjects.**

Methods: **Three-phase analytic approach including item mapping, a modified-Delphi technique, and known-groups validation analysis were used to develop and validate cut-points for functional levels within each of the WD-PF and WD-BH instrument's scales.**

Results: **Four and 5 functional levels were developed for each of the scales in the WD-PF and WD-BH instruments. Distribution of the comparative samples was in the expected direction: the general adult samples consistently demonstrated scores at higher functional levels compared with the claimant and work-disabled samples.**

Conclusion: **Using an item-response theory-based methodology paired with a qualitative process appears to be a feasible and valid approach for translating the WD-BH and WD-PF scores into meaningful levels useful for interpreting a person's work-related physical and behavioral health functioning.**

Key words: outcome assessment (healthcare); disability evaluation; work.

J Rehabil Med 2015; 47: 394–402

Correspondence address: Elizabeth E. Marfeo, Boston University School of Public Health, Health & Disability Research Institute, 715 Albany Streeet, T5W Boston, MA 02118, USA. E-mail: emarfeo@bu.edu

Accepted Dec 10, 2014; Epub ahead of print Feb 27, 2015

INTRODUCTION

The primary federal provider of insurance and financial assistance for individuals with a work-related disability in the US is the Social Security Administration (SSA) (1–3). Support from the SSA represents a significant resource for disabled workers and their families. Previous literature has described both conceptual and programmatic challenges that SSA faces in efficiently and comprehensively characterizing a person's potential ability to work (4, 5). In its 2007 report, entitled, "Improving the Social Security Disability Decision Process," an Institute of Medicine (IOM) panel noted that, as medical treatment and assistive technologies advance, the diagnostic basis for the SSA's work disability adjudication process have become less useful as a marker of work disability (6). To address this gap in current disability assessment, we are currently developing a new instrument; the Work Disability Functional Assessment Battery (WD-FAB).

A unique feature of this new instrument is its conceptual foundation that uses principles outlined by the World Health Organization (WHO) International Classification of Functioning, Disability, and Health (ICF) (4). The ICF highlights the multifactorial nature of disability by focusing on biologic, personal, and social perspectives of disability (4, 7–10). Factors related to a person's ability to work are complex and extend beyond disease symptoms and impairments alone, but include factors such as functional activity limitations, psychological well-being and contextual factors (11, 12). In developing both the physical and behavioral health domains of the WB-FAB, the goal was to expand the scope of SSA's current work disability assessment by creating a measure of functional, activity-based aspects relevant to a person's potential ability to work. This research integrates a more functional approach into the paradigm of work disability assessment, focusing on activities or tasks that relate to a person's potential ability to participate in the workplace compared with SSA's current definition of disability, which focuses primarily on symptoms and impairments related to their disease (4, 13).

The IOM report also recommended the development of alternative approaches, including the creation of standardized functional assessment instruments to more accurately measure work disability in the context of a large federal disability program (2, 6). Clinician-administered and/or performance-based functional assessments are time-consuming, costly, and impractical to implement in large programs such as the SSA's

work disability adjudication process. Comprehensive functional assessment strategies exist, but they are time-consuming to administer and are impractical for widespread use (14). Item-response theory (IRT) and computer adaptive test (CAT)-based assessment of function may be a promising means to provide the SSA with standardized functional assessment tools that are feasible for widespread implementation, and that when combined with other medical evidence and workplace information, can provide valuable information on a person's ability to engage in substantial gainful employment activity (15–18). The goal of our research is to integrate a more functional approach into the process of work disability assessment among a large, national disability program, focusing on activities or tasks that relate to a person's potential ability to participate in the workplace. This approach lends itself well to utilizing the WHO ICF, as an underlying framework for assessing the multifactorial nature of disability (4, 7).

To be consistent with modern heath outcome assessment methodologies we developed 2 new assessments for measuring work-related physical and behavioral health functioning using IRT methodologies (15, 18). IRT measurement models, a class of statistical procedures used to develop measurement scales, examine the associations between individuals' response to a series of items designed to measure a specific outcome domain (e.g. physical functioning) (19). Data collected from samples of individuals are fit statistically to an underlying IRT model that best explains the covariance among item responses (20). When an instrument is developed using IRT methods, the items are calibrated to a common metric, which allows for a hierarchical ordering of ability level or performance. When assumptions of a particular IRT model are met, such as the graded response model, which was used in the WD-FAB development, estimates of a person's functional ability do not strictly depend on a particular fixed set of items (21). This scaling feature allows one to compare persons along a functional outcome dimension even if they have not completed the identical set of functional items. This is a significant advantage over many current instruments developed using classical test theory whereby a fixed set of items is required, limiting the breadth of potential content coverage and score precision possible compared with IRT techniques (21).

Several measures have been developed previously to assess a range of factors that are associated with a person's ability to work; however, many of these measures assess limited scope of the multifaceted construct of work ability, have less than optimal psychometric properties, or are disease-specific (22–24). Instruments such as the Workplace Activity Limitations Scale, Work Limitations Questionnaire, and Work Productivity and Activity Impairment Questionnaire are widely used in both clinical and research settings. A major limitation of these instruments is their limited breadth of coverage in characterizing a person's functional abilities related to work. Comprehensively characterizing a person's ability to work using these existing measures would require administration of many items, using several different instruments, and lead to an undue respondent burden as well as be impracticable for use in the context of a large governmental disability program such as the SSA. In addition, many of items used in these instruments fall slightly outside of the scope of item content relevant to the context of the SSA. Individuals applying for SSA disability benefits have been out of work for least 12 months; therefore items that refer to performance of specific work tasks would be outside the reasonable recall period to reliability assess. Lastly, given the heterogeneity of medical conditions for which individuals are apply for disability benefits in large, national disability programs, a more generic instrument provides advantages over administering disease condition-specific instruments that have been developed previously. For example, a major advantage is that the WD-FAB offers the opportunity to systematically and efficiently collect information about work-related functional abilities across all applicants using a common metric. Comprehensively characterizing a person's ability to work using these existing measures would require administration of many items, using several different instruments, and lead to an undue respondent burden as well as be impracticable for use in the context a large governmental disability program such as the SSA. Because the WD-FAB was developed using modern IRT/CAT methods it offers significant advantages over current measurement techniques in its ability to efficiently and comprehensively characterize work-related function across a wide range of work disabling conditions.

The scaling feature of instruments developed using IRT provides the basis for implementing assessments using CAT. CAT programs use a simple form of artificial intelligence that selects questions tailored to the test-taker, and thereby shortens or lengthens the test to achieve the level of precision desired by a user. The combination of IRT and CAT methods allows the WD-FAB to generate highly precise scores with relatively low respondent burden (25). One challenge with using the IRT/CAT-based instruments is translating the scores into meaningful information that can be used for clinical or policy decision-making or, as with our work, guide characterization of a person's ability to work (26, 27).

An approach to help facilitate the interpretation of IRT-based health outcomes assessment scores is called functional staging. Developing functional stages or levels allows for a brief, meaningful description of a patient's function in various domains of activity and facilitates interpretation of assessment scores. Approaches to developing these categorizations are relatively new in the area of health outcomes assessment (27, 28).

Results from previous work, provide evidence for initial psychometric and construct validity of the work-related physical and behavioral health functioning instruments, the Work Disability – Physical Function (WD-PF) and the WD – Behavioral Function (WD-BH) (4, 15–18). The WD-PF instrument measures the domain of physical function along 5 dimensions: Whole Body Mobility, Upper Body Function, Upper Extremity Fine Motor, Changing & Maintaining Body Position. The WD-BH scales include: Self-Efficacy, Mood & Emotions, Behavioral Control, and Social Interactions.

While IRT methods allow us to place individuals on a continuum, ranging from lowest functioning to highest functioning, the score itself does not indicate what a person can or cannot do

per se, but situates them along the continuum of ability level (20). Given this measurement property of these instruments, an important step to facilitate understanding a score is to provide a system to guide stakeholders' interpretation of a score, so that meaningful conclusions may be drawn. The objective of this study is to apply a novel methodology to develop and interpret functional levels for each of the WD-PF and WD-BH scales.

METHODS

Data for this study combines samples from 3 earlier studies aimed at development and validation of 2 new instruments that measure physical and behavioral health functioning relevant to work (the WD-PF and the WD-BH). Calibration data were collected from a group of individuals applying for SSA disability benefits: claimants. Additional data were collected in order to develop norm-based scores against which to compare the claimant scores. The third sample was a sample of work-disabled individuals used for initial validation of the newly developed instruments. Subjects contributed to either the development of the WD-PF or WD-BH functional levels; there was no overlap between these 2 groups.

Subject selection and setting

The claimant sample included a group of individuals who were applying for disability benefits through the US SSA's disability programs. Eligibility criteria required that the individual apply on his or her own behalf due to a condition that was physical, mental, or both physical and mental in nature. Additional criteria included: 21 years of age and being able to speak, read, and understand English. The general adult sample was drawn from a large internet opt-in survey pool, allowing for approximation of the sample to be representative of a US adult population matched on sex, racial/ethnic background, age, and education, weighted equally. These subjects had to be 21 years or older. Lastly, another independent sample was collected to allow for initial validation of the instruments with a sample of individuals self-reporting permanent work disability due to physical or mental conditions, the "work-disabled" sample. Subjects in all 3 samples had to provide informed consent prior to participating in any study activities. An institutional ethical review board approved all study procedures.

Data collection

Both the claimant and work-disabled samples completed either the WD-PF or WD-BH, depending on the nature of their self-reported disability (physical, mental or both). In addition, the work-disabled subjects completed legacy instruments to examine concurrent validation of the performance of each instrument. Individuals in the general adult sample completed either the WD-BH or WD-PF, selected randomly. Basic demographic information was collected for all study participants.

Instruments

Work Disability Functional Assessment Battery: Physical and Behavioral Health Measures – these instruments were developed for the purposes of characterizing a person's physical or behavioral health functioning across domains relevant to work. The WD-PF scales were Whole Body Mobility, Upper Body Function, Upper Extremity Fine Motor, Changing & Maintaining Body Position. All of the WD-BH scales were used for developing functional levels: Self-Efficacy, Mood & Emotions, Behavioral Control, and Social Interactions). Previous work confirmed the factor structure and construct validity of the WD-FAB scales for both domains (4, 15–18). All scales demonstrated good accuracy, reliability and content coverage for assessing work-related physical and behavioral health functioning. Details of the development and initial psychometric testing of these instruments have been discussed elsewhere (4, 15–18).

Sequential analytic approach

We used a 3-phase analytic strategy that incorporated both quantitative and qualitative procedures to develop functional levels describing categories of physical functioning and behavioral health functioning relevant to the context of work. The origins of this approach come from educational settings where experts use a data-driven consensus process for setting standards for academic performance (28–30). More recently, these methods have been modified and applied in the context of healthcare to develop a technique known as functional staging (27, 31). All quantitative data analysis procedures were performed using SAS computer software (32) and Microsoft Excel was used to generate the item maps.

- Phase 1 utilized the data collected from the WD-BH and WD-PF instruments to empirically derive item maps (27, 28). Item maps are tools that help facilitate the standard setting procedure by ordering the items by difficulty level sorted from easiest to most difficult item within each scale. Estimates of item difficulty were developed using item calibrations estimated from item response theory (IRT) analysis (15, 18). The item maps were based on the general adult sample responses because this is the sample that serves as the comparator of physical and behavioral health functioning when assessing a person's potential ability to work.
- The second phase was based on a modified-Delphi qualitative process that aims to reach consensus from experts and stakeholders in establishing the cut-points for the functional levels (33, 34, 35). These cut-points designated the threshold between one functional level and another. A panel of experts and stakeholders was convened for each of the 2 domains of interest. A total of 19 individuals participated in the modified-Delphi process (8 in the physical group, 11 in the behavioral health group). The panel included individuals with expertise in measurement development, work capacity evaluation, and vocational rehabilitation in both areas of physical and mental health. The expert's professional backgrounds included physical therapists, occupational therapists, a rehabilitation medicine physician, psychologists, and psychometricians. The modified-Delphi technique involved 3 rounds: independent cut-point designation, feedback and summary of the independent cut-points, then finalizing the cut-points with consensus. These levels provided a means by which to initially test construct validity of the WD-FAB's ability to differentiate between groups of individuals with various degrees of physical and behavioral health functioning.
- The last phase focused on validation of the cut-points using known-groups comparisons among 3 independent samples. We hypothesized that the general adult sample should demonstrate a greater proportion of subjects in the higher functional levels compared with the claimant and work-disabled samples; whereas, there would be a smaller proportion of individuals in the lower levels within the general adult sample compared with the claimant and work-disabled samples. Using χ^2 tests at the alpha 0.05 level, we tested statistical differences in distribution of percentages of the sample within each of physical and behavioral health functional levels across 3 comparative samples. The general adult sample served as the reference group. Post-hoc sensitivity analysis of the functional level distributions was conducted to test whether the distributions were affected by age and gender. In addition, within the known groups work-disabled sample, means scores were examined across each functional level for each scale within the WD-BH and WD-PF instrument, hypothesizing that each means score should be discriminative in a monotonic pattern increasing with each incremental functional level.

RESULTS

Table I describes basic demographic information for each of the 3 samples. The physical group included 999 subjects in the general adult sample, 1,017 in the SSA claimant and 497

Table I. *Demographics of subjects yielding physical and behavioral health functional profiles among 3 comparative samples: US general adult, SSA claimants, and a work-disabled sample*

	Physical Health Functional Profile				Behavioral Health Functional Profile			
	US general adult	Claimant	Work-disabled	Statistical test *p*-value	US general adult	Claimant	Work-disabled	Statistical test *p*-value
Sample size	999	1,017	497		1,000	1,015	476	
Age, years, mean (SD)	49.72 (16.12)	49.65 (9.85)	56.02 (8.52)	$F(2,2494)=51.01$, $p<0.0001$	49.07 (15.48)	43.46 (11.09)	51.2 (9.81)	$F(2,2478)=69.96$, $p<0.0001$
Sex, *n* (%)				$\chi^2(2)=0.52$, $p=0.77$				$\chi^2(2)=42.64$, $p<0.0001$
Male	516 (51.81)	543 (53.39)	260 (52.31)		514 (51.50)	444 (43.74)	160 (33.61)	
Female	480 (48.19)	474 (46.61)	237 (47.69)		484 (48.50)	571 (56.26)	316 (66.39)	
Race, *n* (%)				$\chi^2(4)=217.84$, $p<0.0001$				$\chi^2(4)=150.74$, $p<0.0001$
White	782 (78.28)	597 (58.7)	415 (83.5)		773 (77.30)	617 (60.79)	404 (84.87)	
Black/African American	110 (11.01)	323 (31.76)	32 (6.44)		105 (10.50)	266 (26.21)	28 (5.88)	
Other	105 (10.51)	63 (6.2)	42 (8.45)		104 (10.40)	111 (10.94)	38 (7.98)	
Missing	2 (0.20)	34 (3.34)	8 (1.61)		18 (1.80)	21 (2.07)	6 (1.26)	
Relationship status, *n* (%)				$\chi^2(4)=0.52$, $p<0.0001$				$\chi^2(2)=193.38$, $p<0.0001$
Never married	215 (21.52)	230 (22.62)	69 (13.88)		206 (20.60)	301 (29.66)	87 (18.28)	
Married or living with partner	581 (58.16)	424 (41.69	255 (51.30)		627 (62.70)	352 (35.68)	214 (44.95)	
Divorced, separated or widowed	192 (19.22)	359 (35.4)	172 (34.60)		159 (11.10)	357 (35.17)	173 (36.34)	
Refused	11 (1.1)	4 (0.39)	1 (0.2)		8 (0.80)	5 (0.49)	2 (0.42)	
Education, *n* (%)				$\chi^2(6)=278.63$, $p<0.0001$				$\chi^2(6)=311.29$, $p<0.0001$
Less than high school	40 (4.03)	199 (19.61)	18 (3.62)		44 (4.42)	238 (23.52)	17 (3.57)	
High school/GED	361 (36.39)	397 (39.11)	114 (22.94)		361 (36.24)	361 (35.67)	113 (23.74)	
Some college	331 (33.37)	316 (31.13)	243 (48.89)		319 (32.03)	307 (30.34)	219 (46.01)	
College graduate or more	260 (26.21)	103 (10.15)	122 (24.55)		272 (27.31)	106 (10.47)	127 (26.68)	
Nature of disability[a], *n* (%)								
Physical	na	982 (96.95)	497 (100)		na	0	0	
Mental	na	35 (3.44)	0		na	208 (30.49)	154 (32.35)	
Both physical & mental	na	0	0		na	807 (73.51)	322 (67.65)	

[a]Nature of disability question was not ascertained as part of the US General adult sample data collection process. SD: standard deviation; GED: general educational development; na: not applicable.

in the work-disabled sample. The behavioral health group comprised 1,000 general adult subjects, 1,015 SSA claimants and 476 work-disabled subjects. For each of the samples in the physical group, there were slightly more men than women, the samples were predominantly white, with the work-disabled sample mean age slightly older (56 years of age vs 49 years of age for the general adult and claimant samples). For individuals in the behavioral health group, there were slightly more males in the general adult sample (51.5%), and more females in the claimant and work-disabled samples (56.3% and 66.4%, respectively). Similar to the physical group all 3 samples in the behavioral group were predominantly white. On average, the behavioral health group's age was slightly younger than the physical group but was similar in that the work disabled group was slightly older than the general adult and claimant samples (51 years of age vs 49 years of age for the general adult and 43 years of age for the claimant).

Phase 1: Item maps

Results from the empirical analysis allowed graphical presentation of item difficulties for each item in the WD-BH and WD-PF scales. The item maps allowed integration of the item calibrations estimated from the IRT analyses, using a graded response model (35) to facilitate the phase-2 consensus process. The item maps served to simplify the cognitive task for the expert panel to evaluate the content and difficulty levels of each individual item in establishing cut-points for each functional level.

Phase 2: WD-BH and WD-PF functional level definitions

The initial functional levels developed for the WD-BH instrument included either 4 or 5 categories of behavioral health function ranging from poor to excellent. Results of the Behavioral Control scales. For the WD-PF instrument, 5 levels of physical function were established ranging from lowest to highest ability to perform tasks relevant for each scale (Whole Body Mobility, Upper Body Function, Upper Extremity Fine Motor, and Changing & Maintaining Body Position). See Table II for descriptions of the functional levels for the WD-BH and WD-PF instruments.

Table II. *Description of functional levels for the 2 work-related physical and behavioral health functioning (WD-PF and WD-BH) instruments*

WD-BH Functional Levels	
Level 1: Poor Behavioral Health Functioning	Persistent and significant difficulties with multiple aspects of interpersonal interactions and social behaviors within the context of their home, community, or work environments; may report serious difficulties with anger management, severe depression, frequent panic attacks or anxiety that limits their ability to perform daily tasks; demonstrate an inability to function in almost all areas of interpersonal interactions and social behaviors.
Level 2: Basic Behavioral Health Functioning	Some difficulties with aspects of their interpersonal interactions and social behaviors within the context of their home, community, or work environments. These functional limitations may include difficulties with tasks related to social relations with others, report poor judgment or emotional control.
Level 3: Average Behavioral Health Functioning[a]	Adequate interpersonal interaction and social behavior skills most of the time within their home, work, or community contexts; mild difficulties with their emotional states such as episodes of anger, depression, or anxiety; but overall demonstrate good functioning in their daily tasks and are generally satisfied with life.
Level 4: Above Average Behavioral Health Functioning[a]	These individuals may report no more than everyday problems or concerns (e.g. occasional argument with family member, temporarily falling behind on responsibilities) and are able to socialize with others effectively in multiple contexts.
Level 5: Excellent Behavioral Health Functioning	Above average, superior functioning in wide range of interpersonal interaction and social behavior skills, this person demonstrates an even tempered, confident, emotionally regulated state throughout the day and in multiple work, home, or community contexts.
WD-PF Functional Levels	
Level 1: Lowest Physical Functioning	Demonstrate inability or significant difficulty performing basic aspects of physical functioning (i.e. transfers, mobility, and upper extremity tasks) in the context of their home, community, or work environments.
Level 2: Low Physical Functioning	Demonstrate some to a lot of difficulty with aspects of physical functioning (i.e. basic transfers, mobility, and upper extremity tasks) in the context of their home, community, or work environments.
Level 3: Average Physical Functioning	May demonstrate mild difficulties with more advanced aspects of physical function (i.e. ambulation in challenging environments, manipulating small objects, physical tasks requiring repetition or extended duration of time to complete) but overall demonstrate the ability to perform basic tasks and activities.
Level 4: High Physical Functioning	May report a little to some difficulties in a few areas of physical functioning, but are likely to be able to perform high-level activities.
Level 5: Highest Physical Functioning	Report no difficulty with most physical activities. They may report mild difficulty with the most demanding tasks (i.e. highly repetitive activities, those that requiring complex coordination skills, tasks performed for a long duration of time), but report that they are able to do them.

[a]For scales with only 4 levels (WD-BH Social Interactions and Self-Efficacy scales) these levels are combined.

Phase 3: Cut point validation

Figs 1a–d illustrate the results of the Behavioral Health Functional Levels for each of the 3 samples: general adult, claimant, and work-disabled. Counts of claimants classified in each functional level, show a general pattern that the number of claimants at lower functional levels are greater in the claimant and work-disabled samples compared with the general adult sample. For most of the WD-BH scales, results supported the hypothesis that the general adult sample consistently included a higher percentage of individuals in the higher functional levels compared with both the claimant and work-disabled groups. This assumption did hold true for the general adult vs the work-disabled sample for the Self-Efficacy scale. All percentages of general adult vs claimant and general adult vs work disabled are statistically significantly different at the alpha 0.05 level with $p < 0.0001$. Similarly, our hypothesis was also supported for the Physical Functional Levels; the general adult sample consistently yielded significantly more

Table III. *Work-related physical and behavioral health functioning (WD-PF and WD-BH) mean scores across functional levels among a work-disabled sample*

Instrument Subscale	Level 1 Mean (SD)	Level 2 Mean (SD)	Level 3 Mean (SD)	Level 4 Mean (SD)	Level 5 Mean (SD)	Test of statistical significance
WD-BH Instrument						
Self-Efficacy[a]	5.65 (6.6)	26.65 (6.04)	43.74 (6.08)	67.88 (9.43)	NC	$F = 614.69$, $p < 0.001$
Mood & Emotions	13.58 (5.57)	29.41 (4.04)	41.19 (4.09)	54.82 (5.39)	80.44 (10.92)	$F = 496.59$, $p < 0.001$
Behavioral Control	9.22 (0)	23.64 (4.71)	35.45 (3.17)	47.4 (5.94)	71.34 (5.69)	$F = 386.22$, $p < 0.001$
Social Interactions[a]	17.57 (4.18)	34.72 (5.38)	49.64 (4.44)	77.07 (0)	NC	$F = 312.66$, $p < 0.001$
WD-PF Instrument						
Whole Body Mobility	14.92 (4.87)	27.5 (2.59)	36.8 (4.09)	47.35 (2.08)	56.87 (3.75)	$F = 323.86$, $p < 0.001$
Upper Body Function	17.65 (2.66)	26.6 (2.35)	36.14 (3.6)	47.22 (3.22)	59.61 (0)	$F = 551.40$, $p < 0.001$
Upper Extremity Fine Motor	NC[b]	26.53 (2.53)	35.34 (2.55)	45.08 (3.84)	55.96 (1.45)	$F = 743.37$, $p < 0.001$
Changing & Maintaining Body Position	9.15 (0.66)	22.4 (2.65)	33 (3.65)	44.62 (3.64)	58.61 (2.59)	$F = 411.74$, $p < 0.001$

[a]The WD-BH scales for Self-Efficacy and Social Interactions include only 4 functional levels; therefore, mean scores were not calculated.
[b]Within the work-disabled sample, no subjects were categorized in the lowest functional level for the Upper Extremity Fine Motor scale, therefore no mean score was calculated; otherwise all other levels demonstrated statistically significant mean sores in the expected incremental pattern.
SD: standard deviation; NC: not calculated.

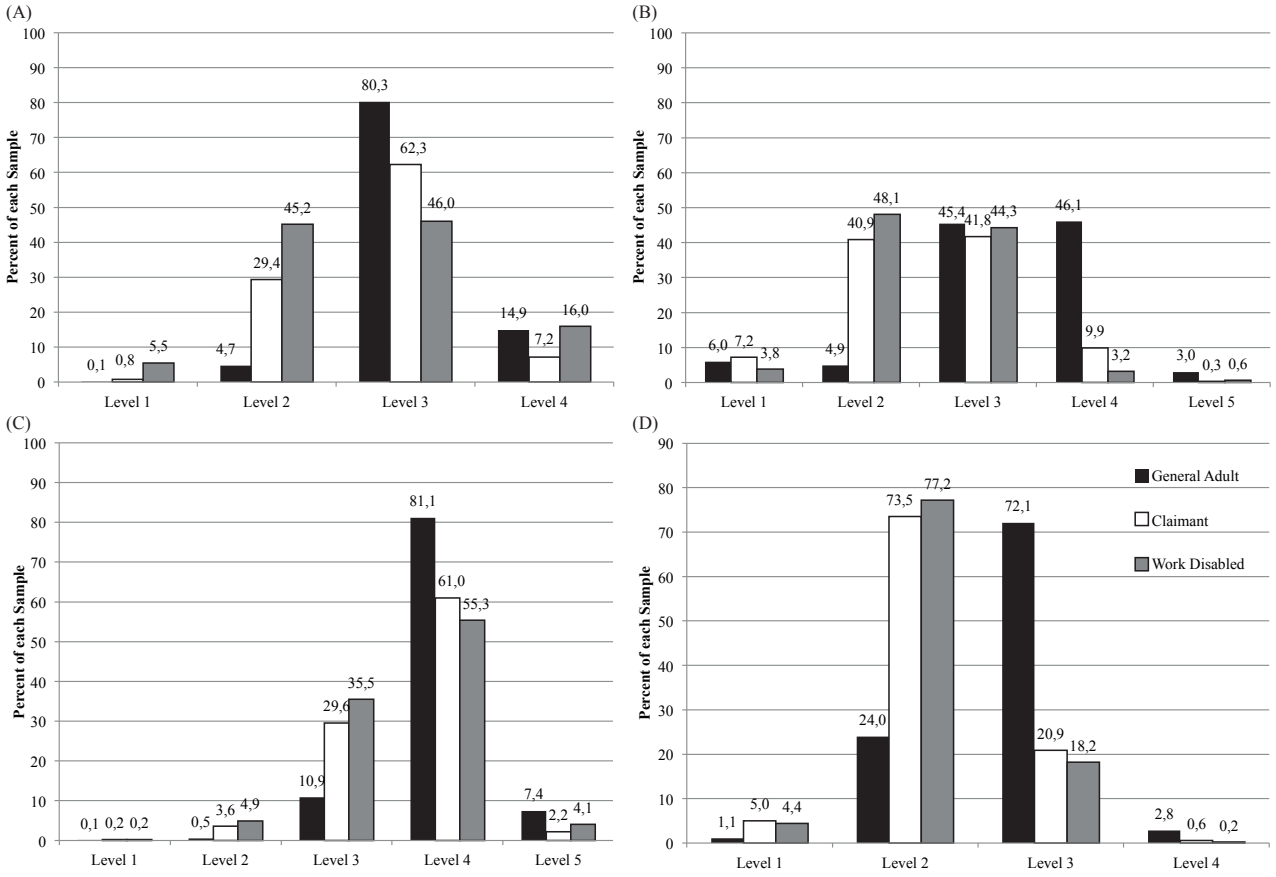

Fig. 1. Behavioral Health Functional levels across 3 samples: US general adult vs claimant, US general adult vs work-disabled. (A) Self-efficacy, (B) Mood and emotions, (C) Behavioral control, (D) Social interactions.

individuals in the higher functioning levels compared with the claimant and work-disabled samples (Figs 2a–d). Although our general hypothesis was supported, variation by scale was observed as to how the samples were distributed within each functional level. Lastly, results from comparing the means using the WD-FAB instruments demonstrated a monotonic relationship (progressively increasing with each increase in functional level) across the functional levels for each scales for both physical and behavioral health (Table III). Mean scores across each functional level were significantly different at the alpha 0.05 level.

DISCUSSION

Findings from this study provide initial validation of functional levels designed to assist in the interpretation of physical and behavioral health functional scores relevant to work. As hypothesized, the general adult sample demonstrated a greater proportion of individuals functioning at the higher levels compared with the claimant and work-disabled samples. Applying the functional level categorizations, the mean score significantly increased within each functional level across the current WD-FAB instrument scales. Results from this study provide initial evidence that using an IRT-based approach for creating functional levels describing both work-related physical and behavioral health is both feasible and psychometrically sound.

Overall, the results of this study suggested that the general adult sample reported higher functional levels for both the physical and behavioral health functional scales compared with the claimant and work-disabled samples. However, there was variation in how the samples were distributed across the levels for each domain: physical and behavioral health. The functional levels as applied using the WD-BH scales resulted in all 3 samples with the majority of individuals in the mid-range functional levels (2–4) with a smaller proportion of each sample being in the lowest and highest functional levels (1 and 5). This is in contrast to the functional levels applied to the WD-PF scales, where the general adult sample tended to have a greater proportion of individuals in the higher functional levels (4, 5). This variation highlights the differences in how to approach measuring physical and behavioral health functioning; physical domains lend themselves well to a hierarchical scoring system, where behavioral health assessments may not fit that structure as well.

When developing new health outcomes measures, it is important to develop tools that provide information that is meaningful and interpretable to relevant stakeholders (26, 36, 37). This study represents continued work in improving

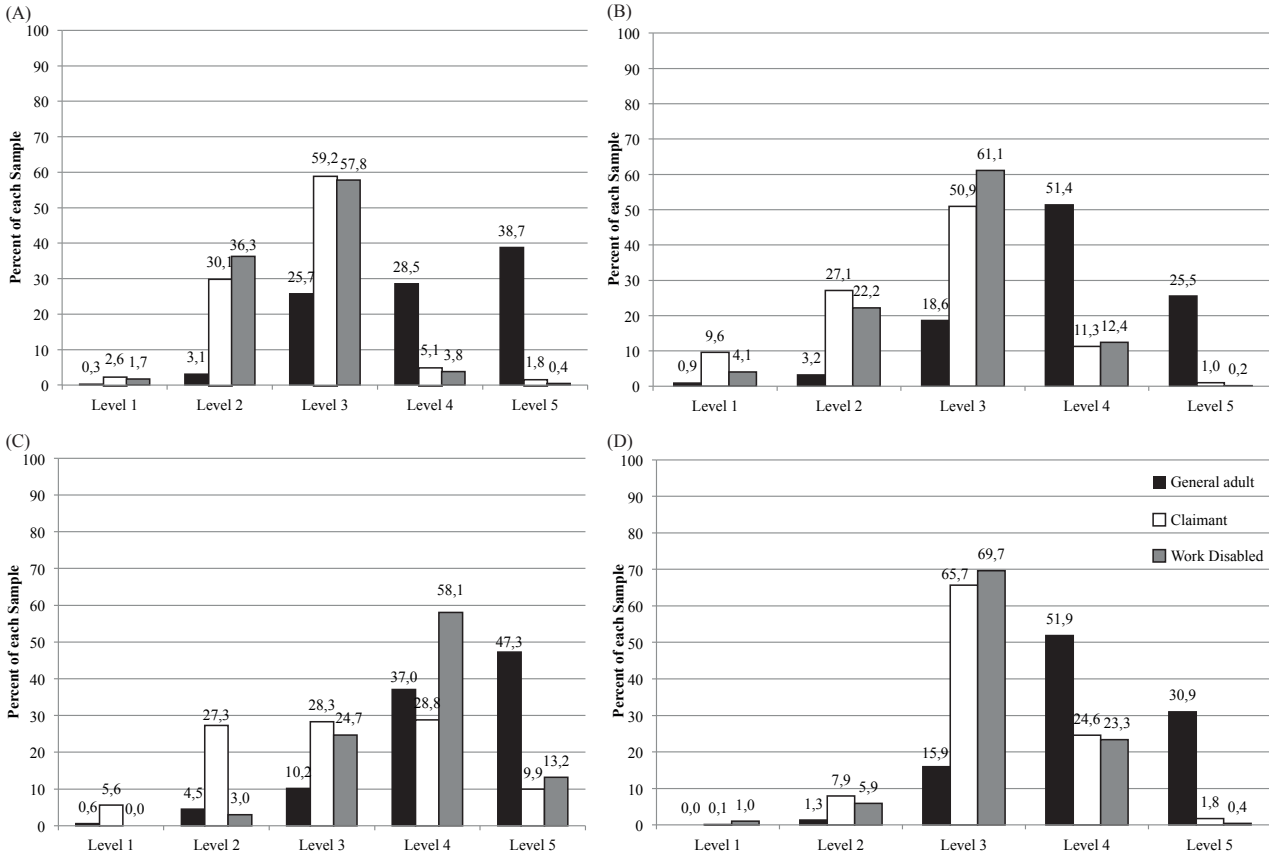

Fig. 2. Physical health functional levels across 3 samples: US general adult vs claimant, US general adult vs work-disabled. (A) Whole-body mobility, (B) Upper body function, (C) Upper extremity fine motor, (D) Changing and maintaining body position.

the way in which work-related physical and behavioral health functioning is measured. The goal of this study was to create a system for interpreting a person's physical and behavioral health functioning as an essential step in developing a new IRT-based measure. The WD-FAB offers advantages over current instruments, in that a highly precise score can be obtained with relatively low respondent burden. The challenge then becomes how to interpret the IRT-based scores in a clinically meaningful way. Findings from this study demonstrate the utility of the WD-BH and WD-PF beyond merely producing a score along a continuum of functioning, but outline a method to facilitate interpretation of those scores in the context of assessing work disability.

This study represents initial exploratory phases of translating the WD-BH and WD-PF scores into meaningful levels useful for interpreting a person's work-related physical and behavioral health functioning; however, a few limitations should be noted. The sequential analytic method used in this study is new in its application to health assessment, future replication and validation of this methodology should be completed to provide additional validation and demonstration of its feasibility in facilitating meaningful interpretation of IRT-based instruments. Within the work-disabled sample, there were no subjects in the Upper Extremity Fine motor lowest functional level category. Although this was not true for the general adult and SSA claimant samples, each of which having very few (0.60% and 5.6%, respectively), but some representation of individuals in this lowest functional level. Previous work demonstrated adequate psychometric performance of this scale, excluding floor effects (18). Unexpectedly, the overall pattern of higher functioning for the general adult sample compared with the claimant and work-disabled samples did not hold for the Self-Efficacy scale. The literature suggests that self-efficacy is an important, yet complex, factor related to work performance (38). From our previous work, we found that this was one of the weaker scales in terms of item fit (16). Through a process of item replenishment we plan to continue to refine and improve of the WD-BH Self-Efficacy scale, which will enable opportunities to better understand the relationship between work disability and self-efficacy (39–41). Lastly, an opt-in internet panel of respondents may possess unique characteristics as a sample population. Efforts were made to match these individuals to the US adult population on key demographic variables, but the permanently disabled sub-set may not be representative of individuals who are unable to work more generally. This study's findings suggest the need for future work in functional level refinement for this particular scale among a permanently disabled sample. Next steps to further validate the WD-FAB instruments and

their functional levels are currently underway. Such efforts include comparison of the WD-FAB with performance-based assessments of physical and mental work capacity.

This study was performed in the context of large US disability benefits program; however we believe the potential utility of these measures may extend beyond that specific context and may be relevant to a variety of researchers and policy-makers looking to assess a person's ability to function in the workplace. Currently, there is no universal standard for cut points to determine thresholds of work functioning that should indicate the allocation of benefits, yet many countries face challenges with rising number of individuals applying for sickness and disability benefits including the UK, the Netherlands and Scandinavian countries, and many other countries within the Organization for Economic Co-operation and Development (OECD) (42–45). Applying these WD-FAB instruments within varying populations would necessitate some level of re-examination of the scores and the relevant functional levels to be conducted as guided by the given program's needs and objectives. One limitation that still exists with many modern measurement development techniques is sometimes there is a need to re-test the psychometric properties when the measures are applied in new populations.

The process used in establishing these initial functional levels for the WD-BH and WD-PF builds upon work done in educational settings used for academic performance standard setting (28–30). More recently, researchers have begun to adapt these methods for application in other health status measures (26, 27). In the area of health assessment, little has been published as to the methodology of moving from score to interpretation. This work presents a novel approach that combines both quantitative and qualitative methods to arrive at both a psychometrically robust instrument that yields meaningful scores, meeting the needs of interested stakeholders. Systematically documenting such processes as described in the article may prove useful for applications beyond work disability assessment, but provide guidance for interpreting scores for IRT-based health outcomes assessments.

In conclusion, this study applied a novel approach utilizing IRT-based methodologies paired with a qualitative process to develop functional levels for 2 measures of work-related physical and behavioral health functioning. Four functional levels were developed for the WD-PF scales (Whole Body Mobility, Upper Body Function, Upper Extremity Fine Motor, and Changing & Maintaining Body Position). Four levels were developed for the WD-BH scales (Mood & Emotions, Behavioral Control, Social Interactions, and Self-Efficacy). Initial validation of the cut-points for each of the functional levels was supported through using 3 independent samples to test the distribution of the subjects in each sample across the functional levels. Results from this study provide further support for the newly developed WD-BH and WD-PF instruments in measuring work-related physical and behavioral health functioning.

ACKNOWLEDGMENTS

This research was supported by Social Security Administration and National Institutes of Health (SSA-NIH) Interagency Agreements (NIH contract nos. HHSN269200900004C, HHSN269201000011C, HHSN269201100009I) and by the NIH intramural research program. The authors and co-authors have no financial or other conflicts of interest to disclose that may bias the reporting of the results presented in this study.

REFERENCES

1. Social Security Administration. Annual Statistical Report on the Social Security Disability Insurance Program, 2011. No 13-11827 Baltimore, MD: Social Security Administration; 2012.
2. Committee on Improving the Disability Decision Process: SSA's Listing of Impairments and Agency Access to Medical Expertise. 7 Findings and Recommendations. Improving the Social Security Disability Decision Process Washington, DC: National Academies Press; 2007.
3. Social Security Advisory Board Disability Roundtable. A disability system for the 21st century. 2006. Available from: http://www.ssab.gov/documents/disability-system-21st.pdf.
4. Marfeo EE, Haley SM, Jette AM, Eisen SV, Ni P, Bogusz K, 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.
5. Brandt DE, Houtenville AJ, Minh MT, Chan L, Rasch EK. connecting contemporary paradigms to the social security administration's disability evaluation process. J Dis Policy Studies 2011; 22: 116.
6. Stobo JD, McGeary M, Barnes DK, editors. Improving the social security disability decision process. Washington, DC: National Academies Press; 2007.
7. Escorpizo R, Stucki G. Disability evaluation, social security, and the international classification of functioning, disability and health: the time is now. J Occup Environ Med 2013; 55: 644–651.
8. Escorpizo R, Reneman MF, Ekholm J, Fritz J, Krupa T, Marnetoft SU, et al. A conceptual definition of vocational rehabilitation based on the ICF: building a shared global model. J Occup Rehabil 2011; 21: 126–133.
9. Finger ME, Glassel A, Erhart P, Gradinger F, Klipstein A, Rivier G, et al. Identification of relevant ICF categories in vocational rehabilitation: a cross sectional study evaluating the clinical perspective. J Occup Rehabil 2011; 21: 156–166.
10. Brage S, Donceel P, Falez F, Working Group of the European Union of Medicine in Assurance and Social Security. Development of ICF core set for disability evaluation in social security. Disabil Rehabil 2008; 30: 1392–1396.
11. Dekkers-Sanchez PM, Wind H, Sluiter JK, Frings-Dresen MH. A qualitative study of perpetuating factors for long term sick leave and promoting factors for return to work: chronic work disabled patients in their own words. J Rehabil Med 2010; 42: 544–552.
12. Tengland P. The concept of work ability. J Occup Rehabil 2011; 21: 275–285.
13. Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. J Rehabil Med 2005; 37: 339–345.
14. Legge J. The evolving role of physiotherapists in pre-employment screening for workplace injury prevention: are functional capacity evaluations the answer? Phys Ther Rev 2013; 18: 350–357.
15. Marfeo EE, Ni P, Haley SM, Bogusz K, Meterko M, McDonough CM, et al. Scale refinement and initial evaluation of a behavioral health function measurement tool for work disability evaluation. Arch Phys Med Rehabil 2013; 94: 1679–1686.
16. Marfeo EE, Ni P, Haley SM, Jette AM, Bogusz K, Meterko M, et al. Development of an instrument to measure behavioral health function for work disability: item pool construction and factor analysis. Arch Phys Med Rehabil 2013; 94: 1670–1678.
17. McDonough CM, Jette AM, Ni P, Bogusz K, Marfeo EE, Brandt DE, 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: 1653–1660.
18. Ni P, McDonough CM, Jette AM, Bogusz K, Marfeo EE, Rasch

EK, et al. Development of a computer-adaptive physical function instrument for social security administration disability determination. Arch Phys Med Rehabil 2013; 94: 1661–1669.
19. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care 2007; 45 Suppl 1: S22–S31.
20. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Qual Life Res 2007; 16 Suppl 1: 5–18.
21. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000; 38 Suppl 9: II28.
22. Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. J Rehabil Med 2011; 43: 23–31.
23. Tang K, Beaton DE, Boonen A, Gignac MA, Bombardier C. Measures of work disability and productivity: Rheumatoid Arthritis Specific Work Productivity Survey (WPS-RA), Workplace Activity Limitations Scale (WALS), Work Instability Scale for Rheumatoid Arthritis (RA-WIS), Work Limitations Questionnaire (WLQ), and Work Productivity and Activity Impairment Questionnaire (WPAI). Arthritis Care Res (Hoboken) 2011; 63 Suppl 11: S337–S349.
24. Wasiak R, Young AE, Roessler RT, McPherson KM, van Poppel MN, Anema JR. Measuring return to work. J Occup Rehabil 2007; 17: 766–781.
25. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol 2006; 59: 1174–1182.
26. Jette AM, Tao W, Norweg A, Haley S. Interpreting rehabilitation outcome measurements. J Rehabil Med 2007; 39: 585–590.
27. Tao W, Haley SM, Coster WJ, Ni P, Jette AM. An exploratory analysis of functional staging using an item response theory approach. Arch Phys Med Rehabil 2008; 89: 1046.
28. Wang N. Use of the Rasch IRT model in standard setting: an item-mapping method. J Educ Meas 2003; 40: 231–253.
29. Reckase MD. A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. Educ Meas: Issues Pract 2006; 25: 4–18.
30. MacCann RG, Stanley G. The use of Rasch modeling to improve standard setting. Pract Assess Res Eval 2006; 11: 1.
31. Wang Y, Hart DL, Werneke M, Stratford PW, Mioduski JE. Clinical interpretation of outcome measures generated from a lumbar computerized adaptive test. Phys Ther 2010; 90: 1323–1335.
32. SAS Institute. SAS users guide, version 9.1. Cary, NC: SAS Institute, Inc.; 2004.
33. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ 1995; 311: 376.
34. de Meyrick J. The Delphi method and health research. Health Educ 2003; 103: 7–16.
35. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monograph No. 17. Richmond VA: Psyhometric Society. Available from: https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf.
36. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007; 6: 1094–1105.
37. Jette AM, Norweg A, Haley SM. Achieving meaningful measurements of ICF concepts. Disabil Rehabil 2008; 30: 963–969.
38. Stajkovic AD, Luthans F. Self-efficacy and work-related performance: \a meta-analysis. Psychol Bull 1998; 124: 240.
39. Hou WH, Tsauo JY, Lin CH, Liang HW, Du CL. Worker's compensation and return-to-work following orthopaedic injury to extremities. J Rehabil Med 2008; 40: 440–445.
40. Wahlin C, Ekberg K, Persson J, Bernfort L, Oberg B. Association between clinical and work-related interventions and return-to-work for patients with musculoskeletal or mental disorders. J Rehabil Med 2012; 44: 355–362.
41. Haley SM, Ni P, Jette AM, Tao W, Moed R, Meyers D, et al. Replenishing a computerized adaptive test of patient-reported daily activity functioning. Qual Life Res 2009; 18: 461–471.
42. Bound J, Burkhauser RV. Economic analysis of transfer programs targeted on people with disabilities. Handbook of Labor Economics 1999; 3: 3417–3528.
43. Marin B, Prinz C. Facts and figures on disability welfare. Eur Centre Social Welfare Policy Res 2003. Available from: http://www.euro.centre.org/detail.php?xml_id=469.
44. Prinz C. European disability pension policies: 11 country trends, 1970–2002. Farnham, UK: Ashgate; 2003.
45. McVicar D. Why have UK disability benefit rolls grown so much? J Econ Surv 2008; 22: 114–139.