Published in final edited form as:
Arch Phys Med Rehabil. 2013 September ; 94(9): 1679–1686. doi:10.1016/j.apmr.2013.03.012.

Scale Refinement and Initial Evaluation of a Behavioral Health Function Measurement Tool for Work Disability Evaluation

Elizabeth E. Marfeo, PhD, MPH[1], **Pengsheng Ni, MD, MPH**[1], **Stephen M. Haley, PT PhD**[1], **Kara Bogusz, BA**[1], **Mark Meterko, PhD**[1], **Christine M. McDonough, PT, PhD**[1,3], **Leighton Chan, MD, MPH**[2], **Elizabeth K. Rasch, PT, PhD**[2], **Diane E. Brandt, PT, MS, PhD**[2], and **Alan M. Jette, PT, PhD**[1]
Boston University School of Public Health, Health and Disability Research Institute

[1]Boston University School of Public Health; Health & Disability Research Institute 715 Albany St., T5W Boston, MA 02118-2526

[2]National Institutes of Health, Mark O. Hatfield Clinical Research Center; Rehabilitation Medicine Department 6100 Executive Boulevard, Suite 3C01, MSC 7515 Bethesda, MD 20892-7515

[3]The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Lebanon, NH 03766

Abstract

Objectives—To use item response theory (IRT) data simulations to construct and perform initial psychometric testing of a newly developed instrument, the Social Security Administration Behavioral Health Function (SSA-BH) instrument, that aims to assess behavioral health functioning relevant to the context of work.

Design—Cross-sectional survey followed by item response theory (IRT) calibration data simulations

Setting—Community

Participants—A sample of individuals applying for SSA disability benefits, claimants (N=1015), and a normative comparative sample of US adults (N=1000)

Interventions—None.

Main Outcome Measure—Social Security Administration Behavioral Health Function (SSA-BH) measurement instrument

Corresponding Author and Reprints: Elizabeth E. Marfeo, PhD, MPH, OTR/L, Boston University, School of Public Health, Health & Disability Research Institute, 715 Albany St., T5W, Boston, MA 02118-2526, T 617-638-1990 F 617-638-1999, emarfeo@bu.edu.

There are no conflicts of interest to disclose of the author and co-authors

Results—Item response theory analyses supported the unidimensionality of four SSA-BH scales: Mood and Emotions (35 items), Self-Efficacy (23 items), Social Interactions (6 items), and Behavioral Control (15 items). All SSA-BH scales demonstrated strong psychometric properties including reliability, accuracy, and breadth of coverage. High correlations of the simulated 5- or 10- item CATs with the full item bank indicated robust ability of the CAT approach to comprehensively characterize behavioral health function along four distinct dimensions.

Conclusions—Initial testing and evaluation of the SSA-BH instrument demonstrated good accuracy, reliability, and content coverage along all four scales. Behavioral function profiles of SSA claimants were generated and compared to age and sex matched norms along four scales: Mood and Emotions, Behavioral Control, Social Interactions, and Self-Efficacy. Utilizing the CAT based approach offers the ability to collect standardized, comprehensive functional information about claimants in an efficient way, which may prove useful in the context of the SSA's work disability programs.

Keywords

Behavioral health; Outcome assessment (healthcare); Work disability; SSA disability determination; Disability evaluation

INTRODUCTION

The Social Security Administration's (SSA) work disability insurance programs are the primary US federal programs providing financial support to over 9.8 million disabled workers and their families.[1] In 2011, mental health impairments represented one of the largest categories of disabling conditions for which individuals receive Social Security Administration's Disability Insurance (SSDI) benefits.[1] The latter fact highlights the need for disability evaluation processes to include systematic, efficient, and accurate assessment of mental and behavioral health functioning relevant to a person's ability to work.

Recent examinations of SSA's current disability determination procedures have identified significant conceptual and operational challenges to the current SSA work disability determination processes.[2-4] Conceptually, the current definition of disability used by SSA is limited to a medical perspective and does not encompass key aspects of the interaction between a person's underlying capabilities and the work environment. Under SSA, disability is assessed by focusing on specific conditions or symptoms rather than characterizing a person's overall functioning. This conceptual gap is especially problematic when applied to mental health conditions due to the episodic and context-specific nature of many mental health related disabilities.[5, 6]

SSA's current evaluation process for determining disability includes a five-stage sequential process that collects medical information data from various health care providers in an effort to characterize the extent to which a person's medical impairment may be limiting their capacity to work.[7, 8] Logistically, this process is time consuming and lacks standardized methods for collecting information about the claimant's symptoms, impairments, or functioning.[2, 9]

Advanced methodologies in measurement scale development have emerged that provide an opportunity to measure complex, multifactorial aspects of health and functioning related to physical or mental conditions using a standardized and efficient approach. Specifically, these methodologies utilize item response theory (IRT) to calibrate an item pool, which is then administered through computer adaptive testing (CAT).[10] The IRT methods create an instrument that can characterize a person's functional status along multiple dimensions of function or scales rather than being constrained to a single dimension. Further, IRT modeling techniques provide a method for evaluating a measurement tool at both the item and scale level, and CAT makes it possible to use that information to individualize survey content for each claimant.[11] These standardized, computer based approaches to collecting health status information could prove to be an efficient and accurate option for SSA, incorporating a more comprehensive assessment of behavioral health functioning into the disability determination processes.

To address some of the measurement challenges associated with the assessment of behavioral health functioning within the context of SSA disability evaluation,, we developed a new IRT/CAT based instrument—the Social Security Administration Behavioral Health Function (SSA-BH) Instrument. Previous work, reported in complementary articles, describes the initial stages of new measurement scale development.[12, 13] Findings from these studies support an underlying unidimensional structure of a set of items designed to assess behavioral health functioning along four key domains of behavioral health function relevant to work: Mood and Emotions, Self-Efficacy, Social Interactions, and Behavioral Control.[12, 13]

The primary objective of this article is to discuss the use of IRT/CAT methods to refine the SSA-BH measurement scales and conduct initial psychometric testing of the instrument. . Specifically, this article will (1) describe the process of applying IRT methods to order the items on a continuum indicative of low to high functional ability in a sample of SSA claimants in a way that is meaningful and useful for the purposes of SSA's disability determination processes, (2) present the development of interval level scales in each domain of behavioral health functioning, and (3) discuss the evaluation and testing of the initial psychometric properties of the SSA-BH Instrument. A secondary objective is to discuss a comparison of the SSA-BH score distribution of claimant response profiles versus response profiles of a normative sample of US adults for each of the functional scales.

METHODS

Participants

The study included a sample of SSA claimants applying for disability benefits, and a second comparative sample of U.S. adults to allow norm-based scoring of the SSA-BH instrument. Eligibility criteria for SSA claimants included: 21 years of age or older, able to speak, read, and understand English, and had to have filed the claim on his or her own behalf due to either a mental or both mental and physical condition. Exclusion criteria specified were related to certain mental conditions or symptoms of paranoia, psychosis, autism, intellectual disability, or Down's syndrome. SSA claimants who met eligibility criteria were stratified by both SSA region and urban/rural location then randomly selected for participation in the

study. For the comparative normative sample, data were collected on a sample of 1000 U.S. adults. The normative data sampling strategy was developed by the YouGov research organization which utilized a proximity sample matching method, drawing respondents from a large opt-in internet panel.[14] The normative sample was to be representative of a U.S. adult population matched on sex, racial/ethnic background, age, and education, weighted equally. A university institutional review board approved this study and participants in both samples provided informed consent.

Data Collection Procedures

The data collected from the SSA-BH instrument used identical methods for both the claimant and normative samples. Trained personnel from Westat research organization collected data from claimants by either phone or internet. Normative data were collected by YouGov research organization via the internet. In addition to responding to a set of 165 items used for developing the SSA-BH instrument, participants in both samples provided self-reported demographic data specifying their age, sex, marital status, race/ethnicity, and education. Details of development of the survey item and data collection procedures have been described in a previous article.[13]

Data Analytic Procedures

Descriptive statistics including frequency distributions were calculated for each item response category and for all demographic variables for both the claimant and normative samples. Where applicable, responses to items were reverse coded to numerically represent behavioral health functioning scores in ascending order from low functioning to high functioning. For the purposes of these analyses, we regard the items assigned to each factor reported by previous exploratory and confirmatory factor analysis results as an item bank.[13] Samejima's Graded Response model was used to calibrate and organize the data into discrete item banks representing four scales of behavioral function: Self-Efficacy, Mood and Emotions, Behavioral Control, and Social Interactions.[15] Three or four threshold parameters were tested based on pre-specified response categories (4-parameter frequency responses ranged from "Never" to "Always"; 3-parameter agreement scale responses ranged from "Strongly Disagree" to "Strongly Agree").

This iterative model building process was conducted in order to both characterize a claimant's level of ability for each dimension of behavioral function and to generate corresponding estimates of each item's difficulty as a measure of claimant's ability to function for each dimension. Item fit was tested using S-X^2 (Pearson's chi-square), comparing the expected and observed frequency distribution of the item summary scores at each score level, using weighted likelihood estimation; fit was also tested for each scale overall.[16-18] Due to multiple comparisons, we chose a cut-off value of 0.01 as indicating significant item misfit of any given item. Item fit analyses were conducted using IRTFIT.[19]

A second phase of instrument assessment included examining differential item functioning (DIF). DIF is a method of analysis to determine if subgroups of claimants at the same ability level demonstrated different probabilities of responses to a given item. For example, for the same item, DIF analysis tests if males and females respond differently as a function of sex

rather than the nature of the item difficulty. The subgroups of interest included examining DIF by age (less than 45 vs. 45 or older) and sex (Male vs. Female). DIF was assessed using two methods: (1) logistic regression based models and (2) IRT based analysis. The first method involved the estimation of ordinal logistic regression models in which the dependent variable was the item score and the independent variables were the background variables of interest (age, sex), ability level (claimant's score estimated from graded response model), and an interaction term of the background variable*ability level (i.e. age*ability level or sex*ability level). The criteria for determining DIF follows that if the background variable effect was significant but the interaction effect was not significant, then the item demonstrated uniform DIF; but, if the interaction term was significant then the item demonstrated non-uniform DIF. Model comparisons were based on the likelihood ratio test and R^2 change to assess effect size of both uniform and non-uniform DIF.[20] Bonferronni adjustment was used to adjust for potiential inflation of Type 1 error caused by significance testing on multiple items.

In the IRT based DIF analyses, hierarchical 2-group IRT models were estimated. The first model set all item parameters equal across the subgroups of interest. The second model excluded the test item, setting all other item parameters equal across the subgroups of interests. To examine DIF, the likelihood ratio test was used with the Benjamin-Hochberg procedure to control for multiple comparisons.[21, 22] Graphical assessment of the IRT models was also used to evaluate the magnitude of the effect size where significant DIF was detected.[23] Items identified as demonstrating DIF in both regression and IRT methods were removed from final items banks or calibrated as separate items according to subgroup categories.

For the final item banks, the breadth of item content coverage was evaluated graphically for each scale by mapping the expected values for the relevant items' response categories against the sample's score distribution on that scale. CAT algorithms were then created for each of the four scales using specialized software developed at Boston University. For each dimension, the CAT was designed to select and administer the first question from the middle of the item difficulty range. Weighted likelihood estimation analysis was then used to estimate the claimant's score and its standard error (SE), and using that updated score information the program selected the next questions for administration with the maximum item information value at the current score level.[24] This iterative process repeated until a preset maximum number of 10 items was reached. The final scores were transformed into a standardized scale with a mean ± SD of 50±10; lower scores indicate lower behavioral health functioning.

To assess the SSA-BH instrument's overall performance we compared scores produced by simulation of 5- and 10- item CATs to the full item bank scores in each domain. The CAT selected questions according to the algorithm described above. The CAT selects items to administer with the goal of maximizing precision of the instrument by using information from the claimant's response on previous questions to select successive items. Data were generated on the claimant's responses and were transmitted to the CAT, yielding a score and SE for each claimant for each scale based on 5 and then 10 items. Measurement accuracy was assessed using the Pearson correlation coefficient to evaluate degree of agreement

between the CAT-generated scale scores for the 5- and 10 item sets with those generated using full item banks. We examined the degree of the floor and ceiling effects by computing the percentage of respondents who selected, respectively, the lowest and highest response items for all items in a given scale.

Precision was evaluated by calculating the standard errors across the range of scores for each 5 or 10 item CAT. In the IRT models, we assumed a true variance equal to 1, so the observed variance could be defined as $1/(1+SE^2)$. Following the definition of reliability (true variance/observed variance), the conditional reliability estimated across the scales was $[1/(1+SE^2)]$.[25-29] Areas with reliabilities <0.70 were considered insufficient.[25]

Lastly, to evaluate and test normative sample scores the same sequence of analyses described above for the claimant sample were conducted. We examined DIF across the normative sample and claimant sample for items in each domain. For the normative sample, ordinal logistic regression models were used to assess DIF. Then we estimated the sample scores for the normative sample using weighted maximum likelihood estimation based on DIF-free items from the claimant sample calibrations. Scores for the claimant and normative sample were used to create two functional profiles using the SSA-BH instrument to characterize behavioral health functioning along the four scales.

RESULTS

Table 1 displays key background characteristics of the claimant (N=1015) and normative (N=1000) study samples. The SSA claimant sample was approximately 56% female, 61% white, and with an average age of 44±11 years. For the normative sample, approximately 52% were male, 77% white, with the average age of the normative sample being 49 ±15 years. There was little evidence of missing data (Claimants: average % of missing = 1.42%, SD 1.43%; Normative: average % of missing = 2.21%, SD 2.56%); therefore, no imputations were performed. “I don’t know” responses were considered missing values, but were not concerning due to low endorsement of this response option in both samples (Normative sample: average % missing =2.21%, SD 2.56%, range 0%-14.9%; Claimant sample: average % missing = 1.42%, SD 1.34%, range 0%-7.68%).

Results from the IRT analysis supported the unidimensionality of each of the four SSA-BH scales: Mood and Emotions, Self-Efficacy, Social Interactions, and Behavioral Control. All items in the scales met the criteria for item fit testing according the $p < 0.01$ of $S\text{-}X^2$ criteria. From DIF analysis, no items indicated a need for removal base on DIF by age, but one item in the Mood and Emotions scale yielded significant DIF by sex (“I don’t know why I cry so often”). Rather than deleting this item, the item was calibrated separately for males and females. The resulting final item bank consisted of 79 items across four scales: Mood and Emotions (k=35), Self-Efficacy (k=23), Social Interactions (k=6), and Behavioral Control (k=15).

Table 2 displays the accuracy of simulated 5- and 10- item CATs compared with the full SSA-BH item banks. Given that the Social Interactions scale had 6 items in the full item bank, only a 5-item CAT simulation was conducted. Correlations of all the 5-item scale

scores with those based on the total item banks for their respective dimensions reached or exceeded 0.91; correlations for the 10-item scale scores with the scores for the full item banks reached or exceeded 0.96. Table 3 compares the distribution of scores for the 10- item simulated CATs (5- item CAT for the Social Interactions dimension) and the full item bank CATs for each of the four dimensions. Among claimants, there were minimal floor or ceiling effects across all four scales; the maximum percent of respondents at the ceiling was under 1% and the maximum percent at the floor on any dimension was under 2%. Consistent with these findings, Figure 1 illustrates the distributions of SSA-BH response categories across each dimension of behavioral health function measured.

Figures 2a-2d present the distribution of scores from claimant and normative samples for each of the four scales including reliability values for the full item bank compared to the 5- and 10- item simulated CATs. These figures also illustrate a positive shift to the higher end of the distribution of behavioral health functioning for the normative sample relative to the claimant sample. Overlap in score distributions of the clamant and normative samples is also illustrated in these figures (Self-Efficacy 66.7%, Mood & Emotions 44.3%, Behavioral Control 71.7%, Social Interactions 46.4%).

The reliability curves indicate that there is some loss of reliability at the upper and lower bounds of each scale, but reliability across the middle range of the distribution for each scale is very good. In the Self-Efficacy scale, the score range of 20-70 revealed a reliability ≥0.9 for the full item bank which covers about 98% of the claimant population. The score range of 26-75, 24-59, 29-77 in Mood & Emotions scale and Behavioral Control scale achieved reliabilities of ≥0.9 for 98%, 80% of the claimant population, respectively. Because of the small item bank in Social Interactions scale, the score range of 35-68 achieved a reliability ≥0.8 using the full item bank and covered about 89% of the claimant population. Finally, comparison of the reliability curves between the 5- and 10- item CATs to the full item bank suggests some loss of reliability as the number of items declined, but only to a modest degree.

Figure 3a displays the functional profile of a 53-year-old female claimant with a history of panic attacks and lower back pain as compared to an individual matched by age and sex from the normative sample. The profile shows that while the claimant's functioning in the area of Social Interactions is similar to that of a matched norm, there are significant differences along the scales of Mood and Emotions, Self-Efficacy, and Behavioral Control. Most noticeably, this claimant is functioning at an especially lower level along the dimensions of Self-Efficacy, and Behavioral Control. In contrast, Figure 3b depicts a 42 year old male claimant with PTSD. In this profile, we can see that the most significant functional limitation, compared to age and sex matched norm scores, is in the Mood and Emotions sub-domain content scale.

DISCUSSION

The SSA-BH instrument testing and evaluation demonstrated strong psychometric properties in terms of accuracy, reliability, and content coverage. The final 79-item instrument demonstrated robust reliability across 5-10 item CATs as compared to using the full item

bank. Floor and ceiling effects were minimal among all four scales. The results of the IRT analysis also aligned conceptually with our hypothesized conceptual framework and previous factor analysis results. This confirmed the instrument's ability to characterize behavioral health functioning along four discrete dimensions relevant for work: Mood and Emotions, Self-Efficacy, Social Interactions, and Behavioral Control.[12, 13]

Preliminary validation of the instrument was demonstrated in the comparison of the distribution of responses of the SSA claimant versus normative samples. Consistently, across all four scales, the claimant sample was operating approximately 1 SD below the normative sample—supporting the notion that on average, claimants were reporting lower behavioral health functioning as compared to a normative US adult sample. Although these findings preliminarily support the validly of the SSA-BH instrument, additional future validation work should include comparison to other legacy instruments and within various claimant and/or patient samples to demonstrate the instrument's concurrent and discriminate validity.

Although the SSA-BH represents a novel approach to measuring aspects of behavioral health functioning relevant to the context of work, existing studies and ongoing development of patient reported health measurers implement similar methodologies for scale development.[10, 30-33] In particular, the work performed by the PROMIS initiative served as a model for development of the SSA-BH. The SSA-BH balances the inclusion of novel content targeting the needs of SSA with the potential for linking to legacy assessments, particularly the PROMIS domains of mental and social health. In the item development and scale refinement phases, careful attention was paid to preserving legacy items in the Mood & Emotions and Social Interactions scales.

The SSA-BH instrument was specifically developed and designed to be implemented in the context of evaluating SSDI/SSI claimant populations. However, many of the underlying constructs measured by the SSA-BH instrument may be relevant to other applications, such as vocational rehabilitation or return to work programs. Such examination and testing of the SSA-BH is beyond the scope of this particular study but offers a promising avenue for future research to examine the utility of the SSA-BH instrument in other relevant contexts assessing behavioral health functioning relevant to work.

By utilizing contemporary measurement development techniques such as IRT and CAT methodologies, we have developed the SSA-BH, an instrument that measures four key aspects of behavioral health function relevant to work: Mood and Emotions, Self-Efficacy, Social Interactions, and Behavioral Control. In addition to advancing the breadth of measurement capabilities, these CAT-based instruments offers promise in the areas of assessing behavioral health functioning in a more standardized, efficient manner. The CAT based approach allows assessments to be individualized, compared across domains using the same underlying metric, and reduces the probability of redundancy in the data collection process. The strength of CAT based instruments, such as the SSA-BH presented here, is its ability to develop functional profiles useful for characterizing behavioral health functioning and disability in a standardized, multifactorial, expedited fashion.

LIMITATIONS

The SSA-BH instrument offers several psychometric and conceptual advancements in measuring aspects of behavioral health in the context of work; however, a few limitations should be noted. Although the SSA-BH instrument allows for characterization of four distinct dimensions of behavioral function, other important aspects of a person's ability to work such as cognition, communication, and environmental factors should be taken into account when assessing the full spectrum of a person's potential ability to work.[5, 6] Several of the hypothesized items that were initially included in the Social Interactions item bank did not meet specified psychometric criteria, resulting in a finalized item bank including fewer items than required for CAT administration. To address this limitation, future item replenishment and additions should be performed. One of the strengths of utilizing the IRT based assessment development and CAT implementation is that these techniques allow for replenishment and updating of items in any given item bank with relatively little difficulty.[34] Additionally, although this study used a rigorous sampling strategy to approximate a representative sample of claimants, we were unable to test the degree to which our sample is representative of the larger SSA claimant population; therefore, limitations of generalizability of the findings should be considered.

CONCLUSION

The SSA-BH instrument represents important advancement in behavioral health assessment both conceptually and psychometrically. The rigorous attention to conceptual guidance from the framework development phase through psychometric testing resulted in an instrument that offers conceptual clarity to measuring behavioral health functioning in the context of work. Findings from this initial testing and evaluation indicate that the SSA-BH demonstrates comprehensive coverage in each content domain while demonstrating robust psychometric properties. The CAT simulation results illustrate that little information is lost when using a 5- or 10- item CAT as compared to a full item bank. Overall, the findings presented provide strong evidence for construct validity of the SSA-BH instrument. Utilizing a CAT based approach, such as the SSA-BH instrument, for measuring behavioral health within the SSA disability determination process creates an opportunity to collect data for characterizing claimant's functional profiles in a standardized, efficient way.

Acknowledgments

Funding for this project was provided through SSA-NIH Interagency Agreements under NIH Contract # HHSN269200900004C, NIH Contract # HHSN269201000011C, and NIH Contract # HHSN269201100009I and through the NIH intramural research program.

Abbreviations

SSA	Social Security Administration
SSDI	Social Security Disability Insurance
IRT	Item Response Theory

CAT	Computer Adaptive Testing
SSA-BH	Social Security Administration Behavioral Health Function instrument
DIF	Differential Item Functioning

REFERENCES

1. Social Security Administration. Annual Statistical Report on the Social Security Disability Insurance Program. Social Security Administration; Washington, DC: 2011. 2012;No. 13-11827
2. Social Security Advisory Board Disability Roundtable. A Disability System for the 21st Century. Available at: http://www.ssab.gov/documents/disability-system-21st.pdf
3. Committee on Improving the Disability Decision Process. Improving the Social Security Disability Decision Process. The National Academies Press; Washington, DC: 2007. SSA's Listing of Impairments and Agency Access to Medical Expertise. Summary.
4. Brandt DE, Houtenville AJ, Huynh M, Chan L, Rasch EK. Connecting Contemporary Paradigms to Social Security Administration's Disability Evaluation Process. Journal of Disability Policy Studies. 2011; 20:1–13.
5. Anthony WA, Rogers ES, Cohen M, Davies RR. Relationships between psychiatric symptomatology, work skills, and future vocational performance. Psychiatr Serv. 1995; 46:353–358. [PubMed: 7788456]
6. MacDonald-Wilson K, Rogers ES, Anthony WA. Unique issues in assessing work function among individuals with psychiatric disabilities. J Occup Rehabil. 2001; 11:217–232. [PubMed: 11822197]
7. Bilder S, Mechanic D. Navigating the disability process: persons with mental disorders applying for and receiving disability benefits. Milbank Q. 2003; 81:75–106. table of contents. [PubMed: 12669652]
8. Social Security Administration. Administrative review process for adjudicating initial disability claims. Final rule. Fed Regist. 2006; 71:16423–16462.
9. Committee on Improving the Disability Decision Process. Improving the Social Security Disability Decision Process. National Academies Press; Washington, DC: 2007. SSA's Listing of Impairments and Agency Access to Medical Expertise. 7 findings and recommendations.
10. Jette AM, Tulsky DS, Ni P, et al. Development and initial evaluation of the spinal cord injury-functional index. Arch Phys Med Rehabil. 2012; 93:1733–1750. [PubMed: 22609635]
11. Reeve, B. [Accessed 9/6/2012] Applications of Item Response Theory (IRT) Modeling for Building and Evaluating Questionnaires Measuring Patient-Reported Outcomes. 2004. Available from: http://outcomes.cancer.gov/conference/irt/reeve.pdf
12. Marfeo EE, Haley SM, Jette AM, et al. A Conceptual Foundation for Measures of Physical Function and Behavioral Health Function for Social Security Work Disability Evaluation. Archives of Physical Medicine and Rehabilitation. XXXX; XX:XX–XXX.
13. Marfeo EE, Pengsheng Ni, Jette AM, et al. Development of an Instrument to Measure Behavioral Health Function for Work Disability: Item Pool Construction and Factor Analysis. Archives of Physical Medicine and Rehabilitation. XXXX; X:XX–XX.
14. Rivers, D. A white paper on the advantages of the sample matching methodology. Sample Matching: Representative Sampling from Internet Panels. YouGovPolymetrix
15. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika. 1969 Monograph Supplement.
16. Orlando M, Thissen D. New item fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24:50–64.
17. 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.
18. Kang T, Chen T. Performance of the Generalized S-X2 Item Fit Index for Polytomous IRT Models. Journal of Educational Measurement. 2008; 45:391–406.

19. Bjorner, JB.; Smith, KJ.; Stone, C.; Sun, X. QualityMetric. QualityMetric; Lincoln, RI: 2007. IRTFIT: A macro for item fit and local dependence tests under IRT models.
20. Jodoin MG, Gierl MJ. Evaluating Type I error and power rates using an effect size measure with the Logistic Regression Procedure for DIF detection. Applied Measurement in Education. 2001; 14:329–349.
21. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995; 57:289–300.
22. Thissen D, Steinberg L, Kuang D. Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Mulitple Comparissons. Journal of Educational and Behavioral Statistics. 2002; 27:77–83.
23. Steinberg L, Thissen D. Using effect sizes for reserach reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods. 2006; 11:402–415. [PubMed: 17154754]
24. Warm TA. Weighted Likelihood Estimation of Ability in Item Response Theory. Psychometrika. 1989; 54:427–450.
25. Mâsse LC, Heesch KC, Eason KE, Wilson M. Evaluating the properties of a stage-specific self-efficacy scale for physical activity using classical test theory, confirmatory factor analysis and item response modeling. Health Educational Research. 2006; 21:i33–146.
26. McDonald, RP. Test Theory: A Unified Treatment. Lawrence Erlbaum Associates; Mahwah, NJ: 1999.
27. Raju NS, Price LH, Oshima TC, Nering ML. Standardized conditional SEM: A case for conditional reliability. Applied Psychological Measurement. 2007; 31:169–180.
28. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. Quality of Life Research. 2007:S5–S18.
29. Cheng Y, Yuan KH, Liu C. Comparison of Reliability Measures Under Factor Analysis and Item Response Theory. Educational and Psychological Measurement. 2012; 72:52–67.
30. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010; 63:1179–1194. [PubMed: 20685078]
31. Cella D, Young S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap Cooperative Group during its first two years. Medical Care. 2007; 45(5 Suppl):S3–S11. [PubMed: 17443116]
32. Gandek B, Sinclair SJ, Jette AM, Ware JE Jr. Development and initial psychometric evaluation of the participation measure for post-acute care (PM-PAC). Am J Phys Med Rehabil. 2007; 86:57–71. [PubMed: 17033591]
33. Jette AM, Haley SM, Tao W, et al. Prospective evaluation of the AM-PAC-CAT in outpatient rehabilitation settings. Phys Ther. 2007; 87:385–398. [PubMed: 17311888]
34. Haley SM, Ni P, Jette AM, et al. Replenishing a computerized adaptive test of patient-reported daily activity functioning. Qual Life Res. 2009; 18:461–471. [PubMed: 19288222]

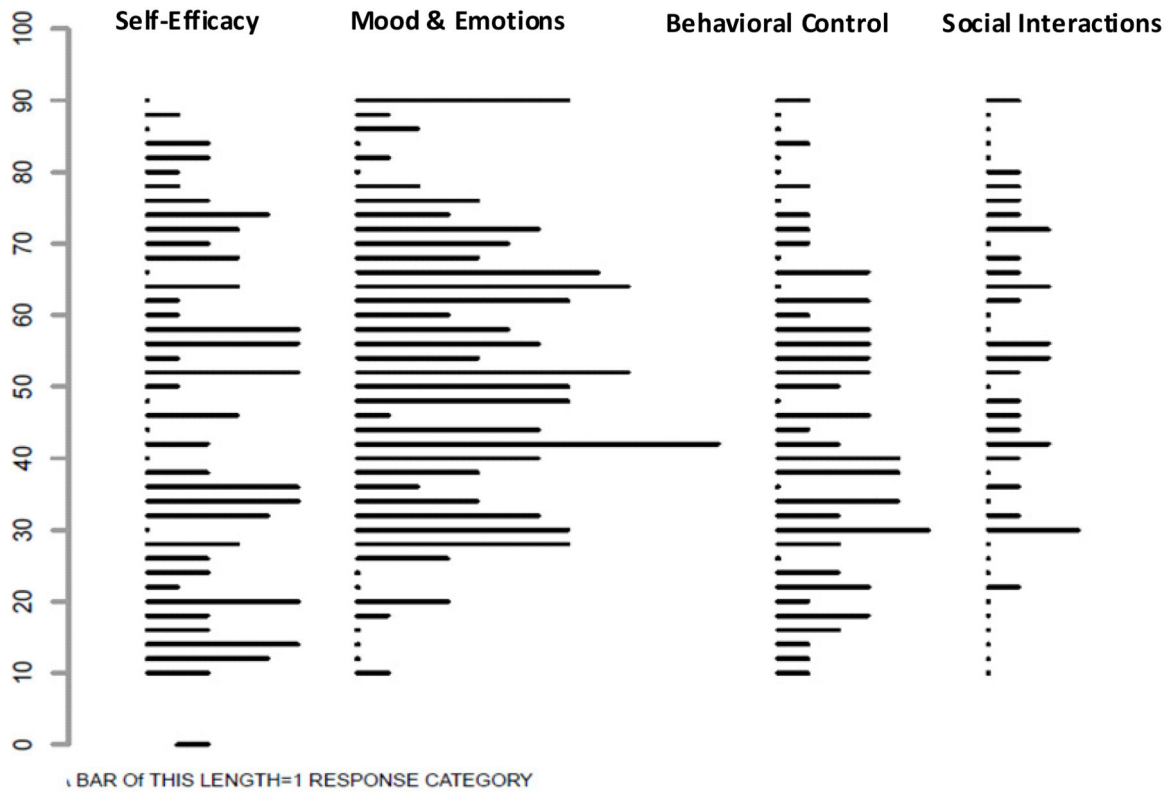
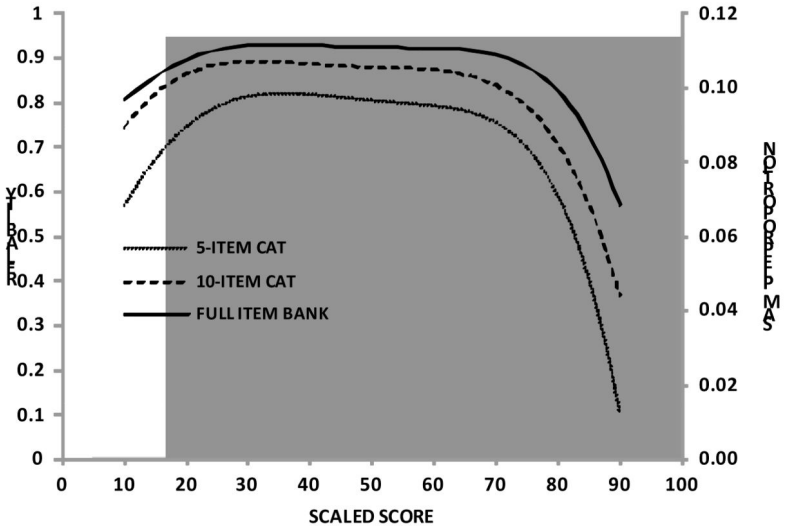
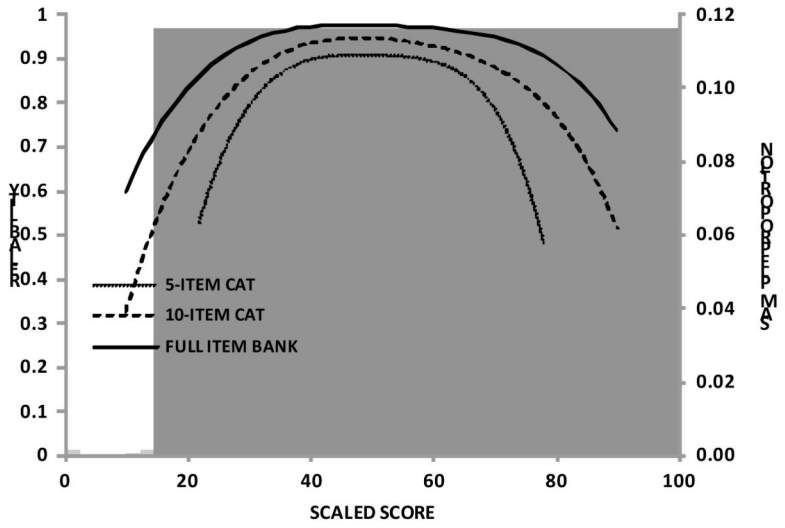


Figure 1. Distribution of SSA-BH Item/Categories for each Content Dimension





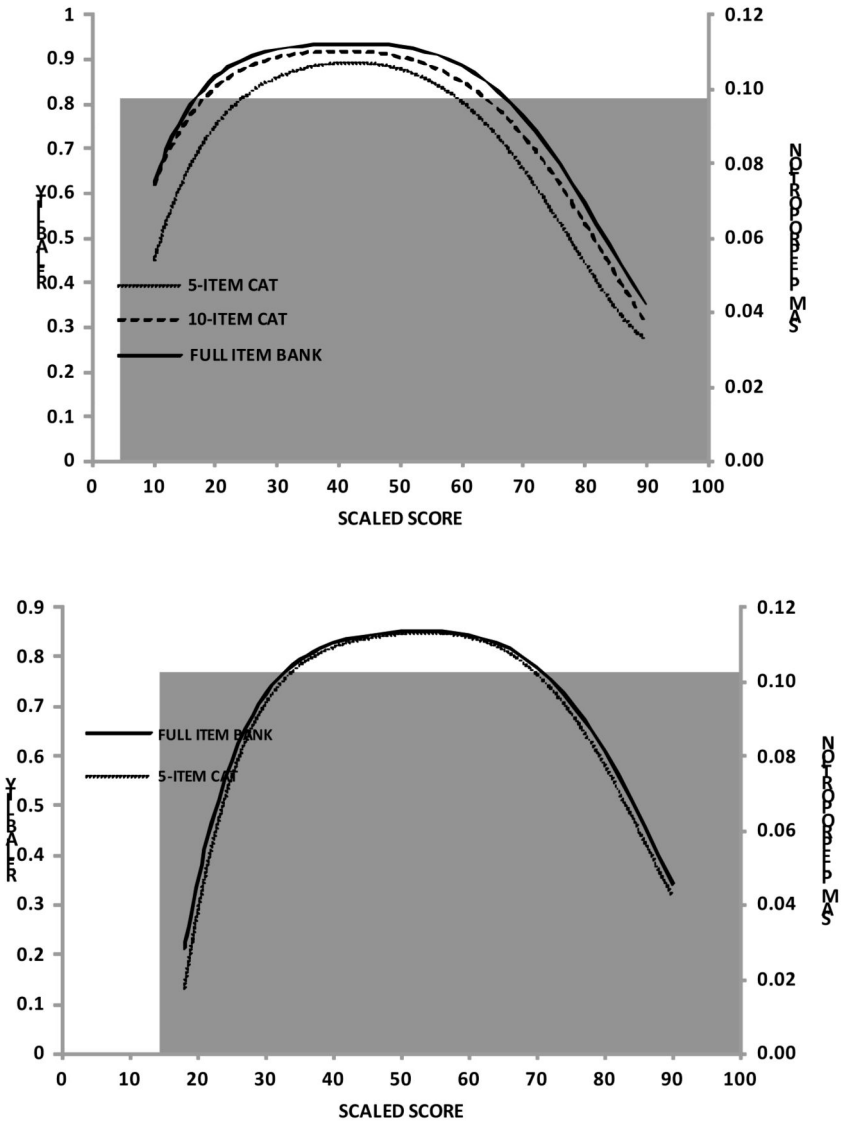


Figure 2a-2d. Distribution of SSA-BH Person Scores and Reliability of 5 item, 10 item, and Full Item Bank by Content Dimension for SSA Claimant (N=1015) and Normative Samples (N=1000)

a. Self-Efficacy Distribution of SSA-BH Person Scores and Reliability of 5 item, 10 item, and Full Item Bank for SSA Claimant (N=1015) and Normative Samples (N=1000)

Note: Claimant distribution in grey (on left), Normative in light grey (right), Overlap distribution (middle) dark grey

b. Mood & Emotions Distribution of SSA-BH Person Scores and Reliability of 5 item, 10 item, and Full Item Bank for SSA Claimant (N=1015) and Normative Samples (N=1000)

Note: Claimant distribution in grey (on left), Normative in light grey (right), Overlap distribution (middle) dark grey

c. Behavioral Control Distribution of SSA-BH Person Scores and Reliability of 5 item, 10 item, and Full Item Bank for SSA Claimant (N=1015) and Normative Samples (N=1000)

Note: Claimant distribution in grey (on left), Normative in light grey (right), Overlap distribution (middle) dark grey

d. Social Interactions Distribution of SSA-BH Person Scores and Reliability of 5 item and Full Item Bank for SSA Claimant (N=1015) and Normative Samples (N=1000)

Note: Claimant distribution in grey (on left), Normative in light grey (right), Overlap distribution (middle) dark grey

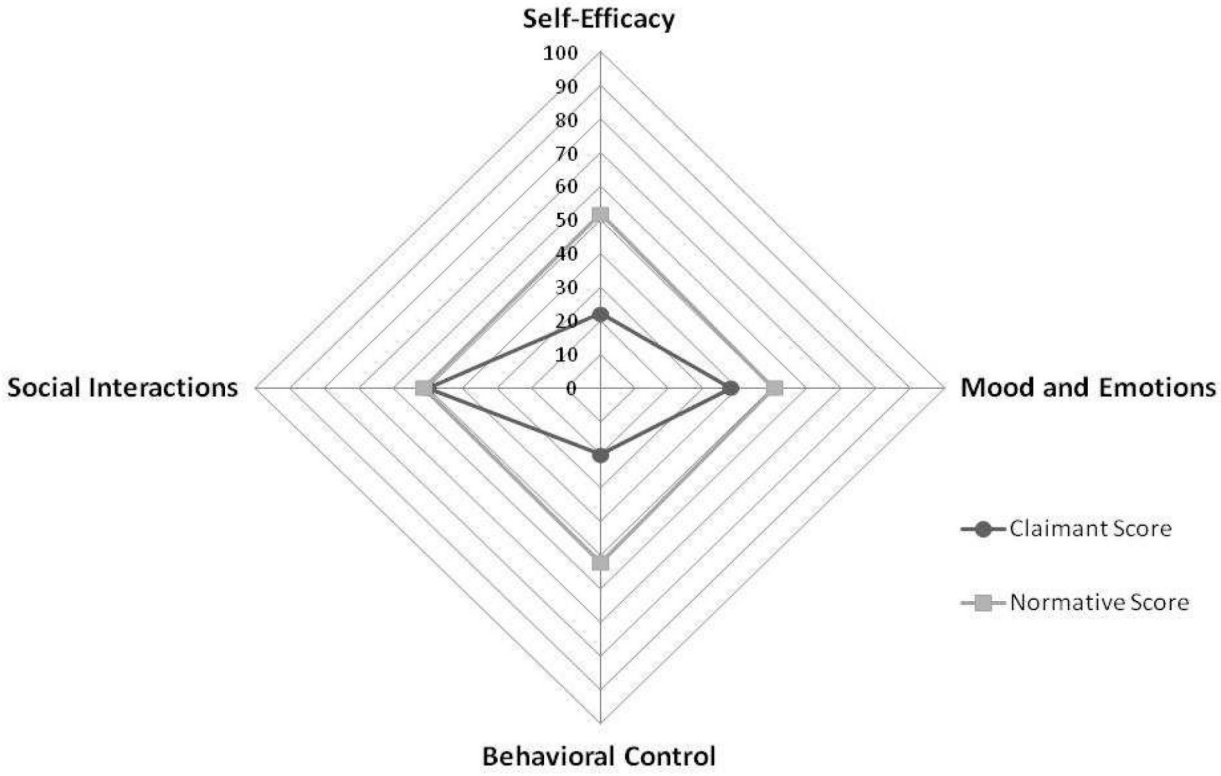


Figure 3a. SSA-BH Functional Profile: 51 year old Female reporting panic attacks, depression, and low back pain

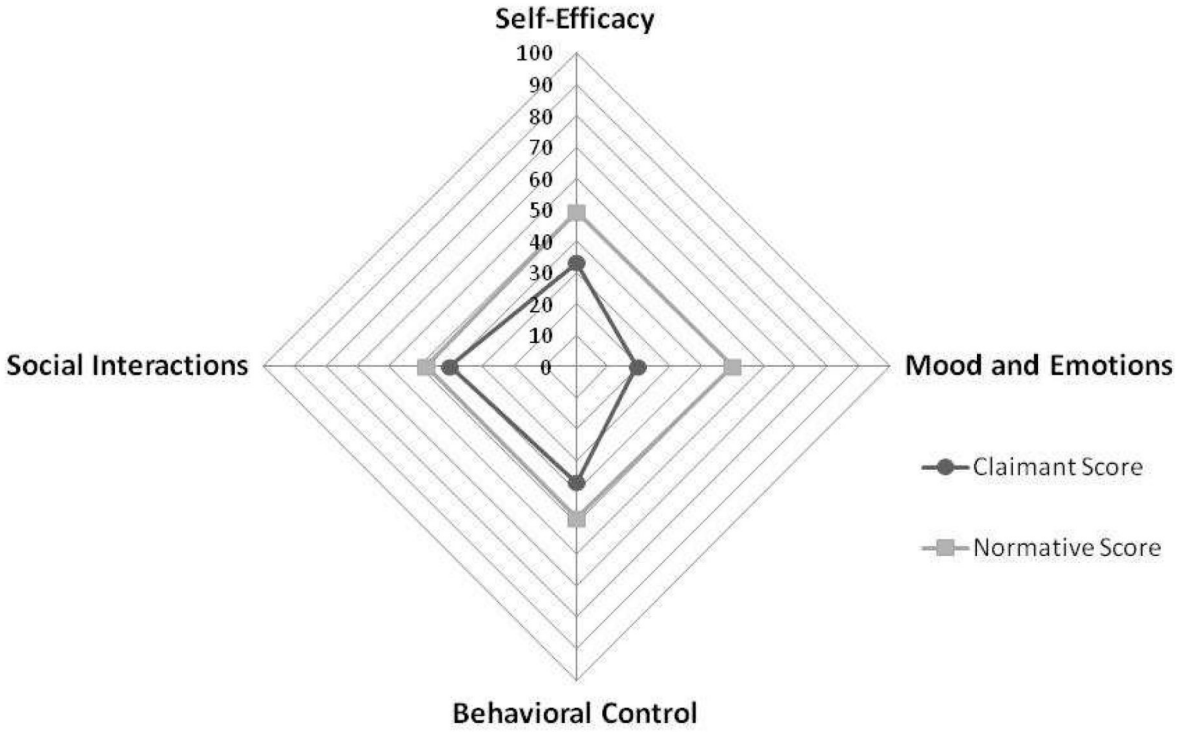


Figure 3b. SSA-BH Functional Profile: 42-year-old Male reporting PTSD and depression

Table 1

Background Characteristics of the Sample

Variable	Study Claimants (N= 1015) Mean ± SD or n (%)	Normative Sample (N=1000) Mean ± SD or n (%)	
Age *	43.76 ± 11.09	49.07± 15.48	T(1791)=−8.8, p<0.0001
Under 40	341 (33.63)	264 (26.64)	
40-55	499 (49.21)	314 (31.69)	
55+	174 (17.16)	413 (41.68)	$X^2(2)=149.96$, p<0.0001
Sex **			
Female	571 (56.26)	484 (48.50)	
Male	444 (43.74)	514 (51.50)	$X^2(1)=12.15$, p=0.0005
Race			
White	617 (60.79)	773 (77.30)	
Black/African	266 (26.21)	105 (10.50)	
American			
Other	111 (10.94)	104 (10.40)	$X^2(2)=87.53$, p<0.0001
missing	21 (2.07)	18 (1.80)	
Education			
Less than high school	238 (23.52)	44 (4.40)	
High School/GED	361 (35.67)	361 (36.10)	
Greater than high school	413 (40.81)	591 (59.10)	$X^2(2)=164.9$, p<0.0001
missing	3	4 (0.40)	

* Age Claimant sample (N=1014), Normative sample (N=991)

** Sex Claimant sample (N=998)

Table 2

Accuracy of 5 and 10-item CATs by Content Dimension (N=1015)

Content Dimension	5-item CAT	10-item CAT
Self-Efficacy	0.91(1014)	0.97(1013)
Mood and Emotions	0.91	0.96
Behavioral Control	0.94(1015)	0.99(1013)
Social Interactions *	0.99(1014)	n/a

*Social Interactions domain included 6 total items in the full item bank

Table 3

Breadth of Coverage for a Simulated 10-item CAT and Full Item Bank for each Content Dimension

Content Dimension	Mode	Claimant Scores				
		N	Mean(SD)	Range	%Ceiling	%Floor
Self-Efficacy	10-item CAT	1013	49.99(10.81)	14.43-92.18	0.1%	0.1%
	Full item bank	1014	50.02(10.47)	12.47-87.85	0%	0.1%
Mood and Emotions	10-item CAT	1015	48.2(10.56)	8.55-90.35	0%	0.89%
	Full item bank	1015	48.23(10.25)	2.18-96.33	0%	0.3%
Behavioral Control	10-item CAT	1013	49.99(10.69)	18.26-91.69	0.49%	0%
	Full item bank	1015	49.99(10.57)	16.47-91.63	0.3%	0%
Social Interactions*	5-item CAT	1014	50.1(11.23)	22.6-92.74	0.69%	1.77%
	Full item bank	1015	50.11(11.11)	21.74-92.49	0.59%	1.48%

*Social Interactions domain included 6 total items in the full item bank