Published in final edited form as:
J Clin Epidemiol. 2014 July ; 67(7): 781–784. doi:10.1016/j.jclinepi.2014.02.005.

Optimizing Psychometrics in Measuring Behavioral Health Functioning: Combining Agreement and Frequency Rating Scales

Elizabeth E. Marfeo[1], **Pengsheng Ni**[1], **Leighton Chan**[2], **Elizabeth K. Rasch**[2], and **Alan M. Jette**[1]

[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

Abstract

Objective—To investigate optimal functioning of using frequency versus agreement rating scales in two subdomains of the newly developed Work Disability Functional Assessment Battery (WD-FAB): the Mood & Emotions and Behavioral Control scales.

Study Design and Setting—A psychometric study comparing rating scale performance embedded in a cross-sectional survey used for developing a new instrument to measure behavioral health functioning among adults applying for disability benefits in the United States.

Results—Within the sample of 1017 respondents, the range of response category endorsement was similar for both frequency and agreement item types for both scales. There were fewer missing values in the frequency items than the agreement items. Both frequency and agreement items showed acceptable reliability. The frequency items demonstrated optimal effectiveness around the mean ± 1–2 SD score range; the agreement items performed better at the extreme score ranges.

Conclusion—Findings suggest an optimal response format requires a mix of both agreement-based and frequency-based items. Frequency items perform better in the normal range of responses, capturing specific behaviors, reactions, or situations that may elicit a specific response. Agreement items do better for those who scores are more extreme and capture subjective content related to general attitudes, behaviors, or feelings of work-related behavioral health functioning.

Corresponding Author and Reprint Requests: Elizabeth E. Marfeo, PhD, MPH, 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

Work was performed at Boston University School of Public Health, Health and Disability Research Institute

Keywords

Scale Usage; Response Styles; Measurement development; Patient Reported Outcomes; Work; Disability evaluation

1. Introduction

Developing Patient-Reported Outcome (PRO) assessments involves several steps that will ultimately determine the tests' overall performance. One of the critical early steps to consider is item structure.[1] Aspects of item structure include the timeframe, content, and rating scale options. The choice of rating scales has been extensively studied in psychological and educational testing.[1, 2] Despite recent advances in measurement development, debate remains as to the optimal item format. Within a given sample, a rating scale may be interpreted by respondents in different ways, resulting in variation in the response elicited. [1] This is especially true in the context of measuring health status attributes, which are often subjective.

This study compared the performance of the Mood & Emotions and Behavioral Control scales of the Work Disability Functional Assessment Battery (WD-FAB), which was developed for use by the Social Security Administration (SSA) to assess work related functioning in its 3 million yearly disability applicants.[3–5] The WD-FAB instrument measures work-related behavioral health functioning in four scales: Mood & Emotions, Self-Efficacy, Social Interactions, and Behavioral Control. A detailed report of the initial psychometric properties of these scales has been reported elsewhere.[3–5] The Mood & Emotions and Behavioral Control scales are composed of items that include either frequency or agreement response categories. The agreement items use a 4-point rating scale ranging from Strongly Disagree to Strongly Agree and typically reflect a general tendency of behavior, attitude, or feeling.[6] Since severity of mental conditions often fluctuates over time, the agreement rating scale allows respondents to reflect on their typical functioning, rather than referencing a specific time frame. The frequency items utilize a 5-point rating scale ranging from Never to Always. The items included in the frequency scale originate from the Patient Reported Outcomes Measurement Instrument System (PROMIS) and Quality of Life Outcomes in Neurological Disorders (Neuro-Qol).[7–10]

The goal of this paper was to examine the extent to which the agreement-based rating scales differ in the information about respondents' Mood & Emotions and Behavioral Control they elicit, when compared with the frequency-based rating scale for the same item content in order to optimize the effectiveness of the WD-BH instrument.

2. Participants and methods

The study included a sample of SSA claimants applying for disability benefits (SSDI/SSI) who were 21 years of age or older, able to speak, read, and understand English, and had recently filed for disability benefits due to a mental health related condition. SSA claimants were stratified by both SSA region and urban/rural location, and then randomly selected for participation. Data were collected on a 165-item instrument by either phone or the internet.

Details of the development of the WD-FAB behavioral health instrument, which in previous studies met assumptions of local dependence and unidimensionalty for all scales, using the standards for item fit testing according the $p < 0.01$ of $S\text{-}X^2$ criteria. A detailed report of the development and psychometrics of the WD-FAB have been described elsewhere.[3–5] Ethics approval was obtained from the university institutional review board.

In the administration of the the Mood & Emotions and Behavioral Control scales, both frequency and agreement response format items for the same item content were administered to respondents. For example, the following two items were fielded to all subjects: (1) “In the past 7 days, many situations made me worry [Never, Rarely, Sometimes, Often, Always]” and (2) “Please specify your level of agreement: Many situations make me worry [Strongly agree, Agree, Strongly disagree, Disagree, I don’t know].”

IRT measurement models can be used examine the associations between individuals’ response to a series of items designed to measure a specific outcome domain. Using IRT modeling techniques, data analysis focused on determining if using the agreement rating scale rather than the frequency rating scale affected the psychometrics of the scales. To perform this comparison, we first looked at the item category response frequencies and percentage missing for each response category. Missing responses included true missing for the frequency response items (item skipped/not answered) and for the agreement response items also included endorsement of “I don’t know” responses.” To assess reliability we applied a graded response model (GRM) to the frequency and agreement items within each scale. We also estimated the group reliability for “Frequency” and “Agree” items, which defined as $1/[1 + E_{\Theta}(1/\text{information})]$. The $E_{\Theta}(\,)$ means the expected value was calculated based on the assumption that the latent trait followed a standard normal distribution.[12] The item fit was examined by $s\text{-}x^2$ using IRTFIT [11] with p-value less than 0.01 indicating a misfit. Item parameter estimates were calculated using IRTPRO.[13]

Based on the estimated item parameters, we calculated the information function curves of “Frequency” items and “Agree” items for each scale. Plots of item information were created to examine how much information a specific item contributes to discriminating various ability levels along the scale’s score range.[14] Item information curves are typically bell-shaped with highly discriminating items reflecting high information functions typically over a narrow range with poorly discriminating items providing less information often covering a wider score range.[15] The test information function was constructed by summing the information functions across all of the items used in the scale.[14, 16, 17]

3. Results

The study sample included 1,015 claimants: 56% female, 61% white, average age of 44 +/− 11 years, see Table 1 for details. There were 5 items in Behavioral Control scale with corresponding “Agree” and “Frequency” item and 10 items in the Mood & Emotions scale with corresponding “Agree” and “Frequency” structures. The range of the item response category endorsement was similar across the item formats for both scales. For missing values in the Behavioral Control scale, the “Frequency” item average percentage of missing

(0.04%, range: 0%~0.1%) was smaller than that in “Agree” item (2.27%, range 1.38%~3.15%). Similarly in the Mood & Emotions scale, the “Frequency” items’ average percentage of missing (0%, range: 0%~0.3%) was smaller than that in “Agree” item (1.78%, range: 0.79%~4.04%).

Results for the GRM had no p-value less than 0.01 based on S-x^2 statistics, indicating all items fit the model. The reliability of the “Frequency” items in the Behavioral Control scale was 0.76 and 0.73 for the “Agree” items. The reliability of “Frequency” items in the Mood & Emotions scale was 0.91 and 0.84 for “Agree” items. The test information function plots illustrate differences in the item performance in both scales. The Behavioral Control scale “Frequency” items provided more information compared to the “Agree” items at the range between 1 standard deviation (SD) above the mean and 3 SDs below the mean, whereas the “Agree” items performed better at 1 SD above the mean and higher (Fig. 1). In the Mood & Emotions scale (Fig. 2), the “Frequency” items performed better than “Agree” items at range between 2 SDs above and below the mean, but the “Agree” items worked better at 2 SDs above the mean and higher.

4. Discussion

This study offered a unique opportunity to explicitly examine potential implications of utilizing a frequency versus an agreement item format for eliciting information about behavioral health functioning. To optimize scale performance in the Behavioral Control and Mood & Emotions subdomains of the WD-FAB instrument, we examined rating scale psychometrics using existing frequency-based response options versus an agreement scale version of the same items. Results revealed that the reliability of the frequency items was marginally higher than that of the agreement items, while the test information function plots illustrated systematic differences in how the frequency and agreement items performed at various places along the possible score distribution for both the scales.

The test function based on the frequency items tended to demonstrate optimal effectiveness around the mean +/− 1–2 SD for both scales. In contrast, the agreement item test function performed better +2 SD above the mean for the Behavioral Control scale and +3 SD above the mean for the Mood & Emotions scale. Assuming the majority of the population will be within plus or minus 2 SD of the mean, our results indicated that the frequency items will typically perform better for the majority of the population. However, this finding also means that for the unique individuals who are functioning at the extreme lower and higher ends of the distribution, the agreement options may be the optimal choice. Similar methodological work has been published in the area of psychological testing and educational settings, but such analysis is sparse in the area of self-reported health outcomes assessment.[1, 2, 6]

Some limitations should be noted. Although the item response scale effectiveness could be isolated and analyzed in this study, isolating the potential differences due to the timeframe variation of the items was not performed. This type of evaluation would require a third item that differed based on timeframe alone, with all else being equal. Such a study is beyond the scope of this project.

5. Conclusion

This study supported the notion that differences in item structure may have important effects on a test psychometric performance. When developing a new assessment tool, balancing item content coverage with item structure becomes an important factor to consider. Future work examining the optimal mix of frequency and agreement items should be conducted. The final WD-FAB behavioral health scales included both frequency-based and agreement-based items in an effort to balance a goal of optimizing the scale performance, and achieve breadth of content coverage.

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.

References

1. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas. 2002; 3(1):85–106. [PubMed: 11997586]
2. Brown GT. Measusring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. Psychol Rep. 2004; 94(3):1015–24. [PubMed: 15217065]
3. 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 Mar 30.
4. 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 Mar 28.
5. 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 Mar 30.
6. Demorest ME, Erdman SA. Development of the communication profile for the hearing impaired. J Speech Hear Disord. 1987; 52(2):129. [PubMed: 3573744]
7. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, 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 Nov; 63(11):1179–94. [PubMed: 20685078]
8. Cella D, Young S, Rothrock N, Gershon R, Cook K, Reeve B, 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]
9. 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 the patient-reported outcomes measurement information system (PROMIS). Med Care. 2007 May; 45(5 Suppl 1):S22–31. [PubMed: 17443115]
10. Cella D. Quality of life outcomes in neurological disorders (NeuroQOL). 2006
11. Bjorner, J.; Smith, K.; Stone, C.; Sun, X. IRTFIT: A macro for item fit and local dependence tests under IRT models. Lincoln, RI: QualityMetric Incorporated; 2007.
12. Cheng Y, Yuan K, Liu C. Comparison of reliability measures under factor analysis and item response theory. Educational and Psychological Measurement. 2012; 72(1):52–67.
13. Cai, L.; du Toit, S.; Thissen, D. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling. Chicago, IL: Scientific Software International; 2011.
14. Hambleton, RK. Fundamentals of item response theory. Sage Publications; 1991. Incorporated

15. Lord FM. Practical applications of item characteristic curve theory. Journal of Educational Measurement. 2005; 14(2):117–38.

16. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. J Pers Assess. 2005; 84(3):228–38. [PubMed: 15907159]

17. Reise SP, Waller NG. Item response theory and clinical measurement. Annual review of clinical psychology. 2009; 5:27–48.

What is new?

- Assessing rating scale performance has a long history in traditional educational test theory. However, research is still required in the public health setting. This study was unique in that we assessed rating scale performance, comparing agreement and frequency scales, in a critical new self-report measure of work related behavioral health functioning to be used by the Social Security Administration.
- This study supported the notion that choosing an optimal response format requires a mix of both agreement and frequency based items. Frequency based items performed better in the normal range of responses and captured information about specific behaviors, reactions, or situations that may elicit a specific response. The agreement items did better for those whose scores were more extreme and captured more subjective content related to general attitudes, behaviors, or feelings of work-related behavioral health functioning.
- These findings have implications for researchers who are interested in developing or improving existing self-report measures and give insight into ways to optimize rating scale effectiveness in measuring health-related outcomes such as work related behavioral health.

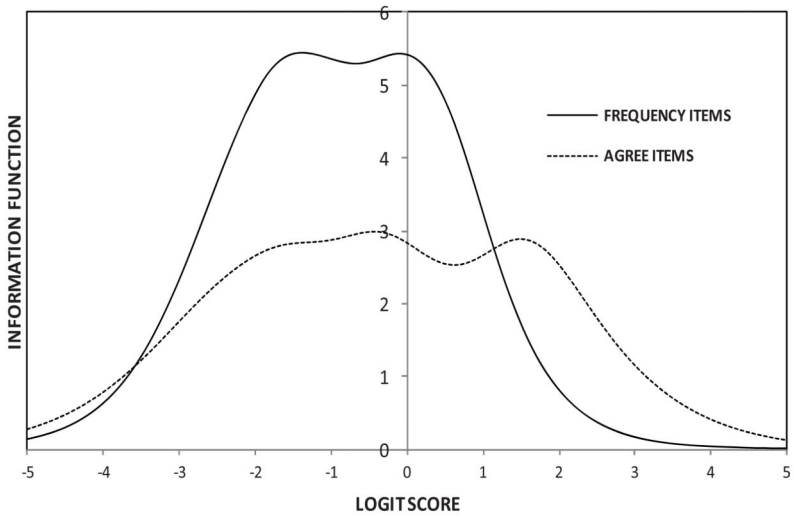

Fig. 1.

Test Information function comparison between “FREQUENCY” and corresponding “AGREE” items in Behavioral Control scale

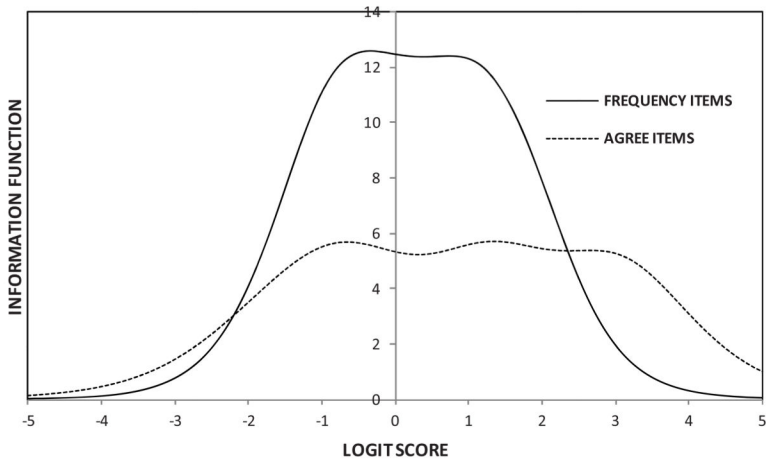

Fig. 2.

Test Information function comparison between “FREQUENCY” and corresponding “AGREE” items in Mood & Emotions Scale

Table 1

Background Characteristics of the Sample (N=1015)

Variable	Mean ± SD or n (%)
Age*	43.76 ± 11.09
Gender**	
Female	571 (56.26)
Male	444 (43.74)
Race	
White	617 (60.79)
Black/African American	266 (26.21)
Other	111 (10.94)
missing	21 (2.07)

Age* (N=991)

Gender (N=998)