Title

Evaluation of options for presenting health-states from PROMIS® item banks for valuation exercises

Journal

Quality of Life Research, 27(7)

ISSN

0962-9343

Authors

Hanmer, Janel
Cella, David
Feeny, David
et al.

Publication Date

2018-07-01

DOI

10.1007/s11136-018-1852-1

Peer reviewed

Quality of Life Research (2018) 27:1835–1843
https://doi.org/10.1007/s11136-018-1852-1

Evaluation of options for presenting health-states from PROMIS® item banks for valuation exercises

Janel Hanmer[1] · David Cella[2] · David Feeny[3,4] · Baruch Fischhoff[5] · Ron D. Hays[6] · Rachel Hess[7] · Paul A. Pilkonis[8] · Dennis Revicki[9] · Mark Roberts[10,11] · Joel Tsevat[12] · Lan Yu[10]

Accepted: 6 April 2018 / Published online: 12 April 2018

Abstract

Purpose Health status descriptive systems based on item response theory (IRT), such as the Patient-Reported Outcomes Measurement Information System (PROMIS®), have item banks to measure domains of health. We developed a method to present such banks for health-state valuation.

Methods We evaluated four different presentation approaches: a single item (1S), 2 items presented separately (2S), 2 items presented together (2T), or 5 items presented together (5T). We evaluated these four approaches in three PROMIS item banks (depression, physical function, and sleep disturbance). Adult community members valued health-state descriptions using the visual analog scale and standard gamble methods. We compared the approaches by the range of item bank theta scores captured, participants' assessments of difficulty (1 = very easy to 7 = very hard), and exit interviews.

Results Participants ($n = 118$) ranged in age from 18 to 71; 63% were female and 54% were white. The 1S approach captured the smallest range of theta scores. A monotonic relationship between theta score and mean standard gamble estimate was found with all approaches except 2S. Across all 3 item banks, mean difficulty assessments were 2.35 (1S), 2.69 (2T), 2.78 (5T), and 2.80 (2S). In exit interviews, participants generally found all four approaches similarly meaningful and realistic.

Conclusions Creating health descriptions by presenting 2 items maximized the range of theta while minimizing difficulty and maintaining a monotonic relationship with utility estimates. We recommend this approach for valuation of IRT-based descriptive systems such as PROMIS.

Keywords Health-state descriptions · Preference-based scores · Valuation of health-states · Utilities

Introduction

Preference-based summary scores of health are useful for tracking population health, comparing groups, and performing cost-effectiveness analyses [1]. Preference-based scoring functions are estimated by having individuals value a set of health-state descriptions. Valuation procedures include standard gamble, time trade-off, and visual analog scales [2–4].

The health-state descriptive space used for preference-based measures has consisted of a fixed set of health domains and levels for each domain [5–10]. For example, the EuroQol-5D-3L has five health domains: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, with three levels on each (no problems, some problems, extreme problems) [5].

In the last decade, there have been significant advancements in health-state description systems. The Patient-Reported Outcomes Measurement Information System (PROMIS®) is a major effort sponsored by the National Institutes of Health to advance measuring health through state-of-the-science qualitative and quantitative methods [11, 12]. In particular, PROMIS utilizes item response theory (IRT) [13] to create unidimensional item banks for health domains (e.g., pain, physical function) calibrated on a common metric. Any set of items selected from the item bank can be used to estimate an individual's score ("theta") for a domain. The PROMIS domains are usually reported as a T-scores which are constructed with a mean of 50 and

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-018-1852-1) contains supplementary material, which is available to authorized users.

✉ Janel Hanmer
hanmerjz@upmc.edu

Extended author information available on the last page of the article

standard deviation of 10 relative to a target population (e.g., the U.S. general population).

Leveraging the improvements in health descriptive systems has the potential to improve preference-based measurement. However, because item banks comprise a large number of items, one can imagine many permutations and combinations of items that could be combined into a health-state description for valuation. Bookmarking methods and scale judgement methods, which build descriptive vignettes from 5 items, have recently been used to create PROMIS item bank health-state descriptions for minimally important difference studies and establishing clinically relevant classifications [14, 15]. Here, we describe and evaluate four different methods for presenting health-state descriptions from the PROMIS item banks for use in preference valuation. These methods preserve the advantages of IRT by linking the descriptions to the underlying unidimensional construct.

Methods

Creation of health-state descriptions

We selected three PROMIS item banks which cover disparate aspect of health: depression [16], physical function [17], and sleep disturbance [18]. Each item in these banks has five possible response options (i.e., never, rarely, sometimes, often, or always). We created health-state descriptions from those items through a combination of item calibration data and qualitative analysis by domain experts as described below. We evaluated four different approaches: a single item (1S), 2 items presented separately (2S), 2 items presented together (2T), and 5 items presented together (5T).

Selection of the items used for the 1S, 2S, and 2T sets started by examining the theta estimate for each response category for each item in each domain, such as would be produced if only those responses were used to estimate theta. For example, the depression item bank has the item, "I felt sad… never, rarely, sometimes, often, or always." The associated theta estimates for the five possible responses are −0.96, −0.03, 0.67, 1.38, and 2.08.

The depression item bank was Emotional Distress–Depression v1.0. Parameter estimates were from the PROMIS Wave one sample. The physical function item bank was Physical Function v1.2. Parameter estimates were from PROMIS 1 Wave 1 with Extension sample. The sleep disturbance item bank was Sleep Disturbance v1.0. The parameter estimates were from PROMIS Sleep Wave 1. [see http://www.healthmeasures.net/index.php and https://www.assessmentcenter.net].

To select a single item for the 1S method, we presented the 5 items with the largest range to domain experts who then selected the most representative item. For the 2S and 2T methods, we wanted to capture a wide range of domain scores. We selected a set of 5 items which, based on the IRT parameters, best covered the highest range of the target concept (i.e., physical function) and selected 5 items which, based on the IRT parameters, best covered the lowest range of the concept. These 10 items were then presented to domain experts who picked 1 item from the highest set and 1 item from the lowest set which they felt captured the important aspects of the domain. For depression, those aspects were mood and anhedonia. For physical function, they were mobility and dexterity. For sleep disturbance, they were sleep restfulness and duration.

Domain experts were asked to avoid items which shared content with other domains. For example, one of the items with the highest theta estimates in the sleep disturbance item bank is "I felt sad at bedtime…" This item was not considered because it may have content overlap with depression.

All descriptions used in the valuations are included in Online Appendix A. In the 1S method, each response to the selected item was presented separately for a total of five descriptions. In the 2S method, each item response was presented separately; because each item has five response options, there was a total of ten descriptions. In the 2T method, the 2 items were presented together. Item parameter information was used to create the most likely response combinations for a total of nine descriptions. Creation of the 5T descriptions followed the procedure outlined by Cook et al. [14] where a theta score is selected and then a representative set of 5 items and their responses at that theta score are selected. We created 5T descriptions at increments of 0.5 on the theta score. There were eight depression descriptions, nine physical function descriptions, and nine sleep disturbance descriptions.

Community sample

We recruited 118 adults from metropolitan Pittsburgh, Pennsylvania using the Clinical and Translational Science Institute's Research Participant Registry [https://researchrecruitment.pitt.edu/ctsi/home/about] to participate in a video-recorded, in-person interview at a research office. Participants responded to an advertisement on the registry's website. Inclusion criteria were of age 18 years or older and to be comfortable communicating in English. There were no other inclusion or exclusion criteria. Participants were paid $35.

Evaluation procedure

We assigned participants to evaluate a single health domain from the three domains. The first 40 participants completed depression, the next 40 completed sleep disturbance, and the last 38 completed physical function. Participants first completed the 8-item short form of the domain to familiarize

themselves with the concept. The health descriptions for each of the four methods were printed on individual cards (e.g., four "card sets"). Then they evaluated each of the four card sets in random order. They first ranked the cards from best to worst. Second, the best and worst cards were used as anchors on a 0–100 visual analog scale and respondents were asked to place the other cards on the scale. Third, respondents evaluated the cards using the standard gamble on a chance board. In the standard gamble, their best and worst cards were used as the best and worst outcomes, with the others as intermediate outcomes [6, 7]. Respondents assessed each card set's difficulty by using a 7-category response scale (1 = very easy to 7 = very hard).

After completing all four card sets, respondents completed a self-administered questionnaire with demographic information, the subjective numeracy scale [19], the Iron-Wood religiosity scale [20], self-rated health, number of physician visits per year, number of times hospitalized, and experience with conditions that limit a person's ability to take care of him or herself. Finally, respondents were asked questions about the meaningfulness and realism of the cards in semi-structured exit interviews.

Analyses

We compared the results from the four methods in several ways. First, we examined the range of item bank theta scores captured. Second, we compared the monotonicity of mean valuations across the methods; monotonicity was considered violated if the rank order of health-states by valuation was different at least twice from the rank order of health-states by theta. Third, we compared the methods by participant assessments of difficulty on the 7-point response scale. Fourth, we evaluated participant reports from the semi-structured exit interviews.

Results

We recruited 118 participants. The mean age of the sample was 37 (SD = 16, range 18–71); 63% were female; and 54% were white and 34% were African-American (Table 1). The sample included a range of educational backgrounds, experience with the domain to which they were assigned, and experience with other health problems that limit a person's ability to take care of him or herself. Thirty percent of participants reported their health as excellent, whereas 48% reported their health as very good, 19% as good, 3% as fair, and 0% as poor. The in-person interview took an average of 44 min (Table 2).

As seen in Fig. 1, the 1S method had the narrowest range of theta scores for each domain. The 5T method always had the widest range and the 2T method had the second widest range. The figure includes dashed lines to indicate the 5th and 95th percentile scores for the item bank's calibration sample [21–23]. Although these samples are not perfectly representative of the U.S. general population [24], they provide an indicator of the distribution of scores in the general population. A monotonic relationship between item bank theta estimate and mean standard gamble estimate was found for the 1S, 2T, and 5T methods in all three domains (Fig. 2a–c). With the 2S method, mean standard gamble estimates trended with theta scores, but monotonicity was violated several times. Figure 2a illustrates the results for the depression item bank, Fig. 2b for physical function, and Fig. 2c for sleep disturbance.

Across all 3 item banks, 74% of participants found 1S to be easiest and 71% found 5T to be most difficult. Mean difficulty assessments for the combined sample on the 7-point response scale were 2.25 (1S), 3.04 (2T), 3.25 (5T), and 3.34 (2S). The rank order of difficulty was the same in each domain.

In exit interviews, participants generally reported all four methods to be similarly meaningful and realistic. Most participants reported that the 5T method provided too much information; a notable exception was a participant who had personal experience with the item bank she was evaluating (depression) and found the rich descriptions helpful. Participants reported that the 1S method was easiest, but that the 2T method was still manageable. Many participants found the 2S method frustrating, as they had difficulty comparing single responses from different items.

Overall, participants were generally engaged in the task and expressed thoughtful reasoning about their responses. For example, a small subset of participants found the best (by theta) depression health-states to be "unnatural" saying that rarely feeling sad was preferable to never feeling sad.

Discussion

The construction of health-states for valuation studies requires a careful balance between descriptive richness in content and respondent burden. Historically, the health-states used in valuation studies have either been created de novo by instrument developers (HUI, EQ-5D) or taken from an existing static health descriptive system (SF-6D, FACT) [5–10, 25]. Item response theory has modernized health descriptive systems by calibrating items on underlying constructs; from such item banks, a small set of informative items can be used to measure the construct. We have developed a method to present an item bank for valuation. This method uses the advantages of item response theory, particularly knowledge about an item's location on the underlying construct, to improve the descriptive system for a preference-based scoring system [26].

Table 1 Sample demographics

	Depression	Physical function	Sleep disturbance	Combined sample
N	40	40	38	118
Age mean	42.5	32.0	35.4	36.7
Age range	20–71	18–69	19–68	18–71
Female	60.0	60.0	68.4	62.7
Domain limitation experience, personal (%)	40.0	35.0	21.0	32.2
Domain limitation experience, caregiver (%)	7.5	5.0	2.6	5.1
Domain limitation experience, family (%)	50.0	72.5	47.4	56.8
Domain limitation experience, work (%)	10.0	15.0	13.1	12.7
General limitation experience, personal (%)	22.5	22.5	13.1	19.5
General limitation experience, caregiver (%)	25	20.0	13.1	19.5
General limitation experience, family (%)	62.5	80.0	50.0	64.4
General limitation experience, work (%)	17.5	25.0	10.5	17.8
Doctor visits per year (range)	0–20	0–15	1–12	0–20
Ever hospitalized (%)	57.5	52.5	50.0	53.4
Race				
White (%)	50.0	47.5	63.2	53.5
Black (%)	40.0	35.0	26.3	33.9
Asian (%)	0	7.5	2.6	3.4
Other (%)	7.5	7.5	5.3	6.9
Hispanic (%)	5.0	2.5	2.6	3.4
Education				
High school (%)	20.0	7.5	10.5	12.7
Some college (%)	30.0	60.0	31.6	40.7
College (%)	15.0	20.0	21.1	18.6
Some post-graduate (%)	15.0	7.5	10.5	11.0
Post-graduate (%)	20.0	5.0	26.3	16.7
Self-rated health				
Excellent (%)	27.5	27.5	34.2	30.0
Very good (%)	40.0	50.0	55.3	48.3
Good (%)	27.5	17.5	10.5	18.6
Fair (%)	5.0	5.0	0	3.4
Poor (%)	0	0	0	0

Table 2 Difficulty assessments on a 7-point Likert scale (1 = very easy to 7 = very hard) for each presentation method

	Depression $n = 40$, mean (SD)	Physical function $n = 40$, mean (SD)	Sleep disturbance $n = 38$, mean (SD)	Combined sample $n = 118$, mean (SD)
1S	2.38 (1.35)	2.03 (1.10)	2.33 (1.46)	2.25 (1.31)
2S	3.25 (1.37)	3.53 (1.71)	2.97 (1.35)	3.25 (1.49)
2T	3.05 (1.52)	3.12 (1.64)	2.94 (1.65)	3.04 (1.60)
5T	3.23 (1.44)	3.45 (1.65)	3.34 (1.51)	3.34 (1.53)

IS is a single item, 2S is two items presented separately, 2T is two items presented together, and 5T is five items presented together

Based on this study, we recommend approach 2T: select 2 representative items from an item bank and present them together. We compared 2T with three other methods: using a single item (1S), using 2 items separately (2S), and using 5 items together (5T). While evaluating a single item (1T) was easiest for participants, this method captured significantly less of the item bank's range than evaluating 2 items together. As a result, using a single item would be more likely to produce ceiling and floor effects. Presenting 2 items separately covers a wider range of the item bank, but was

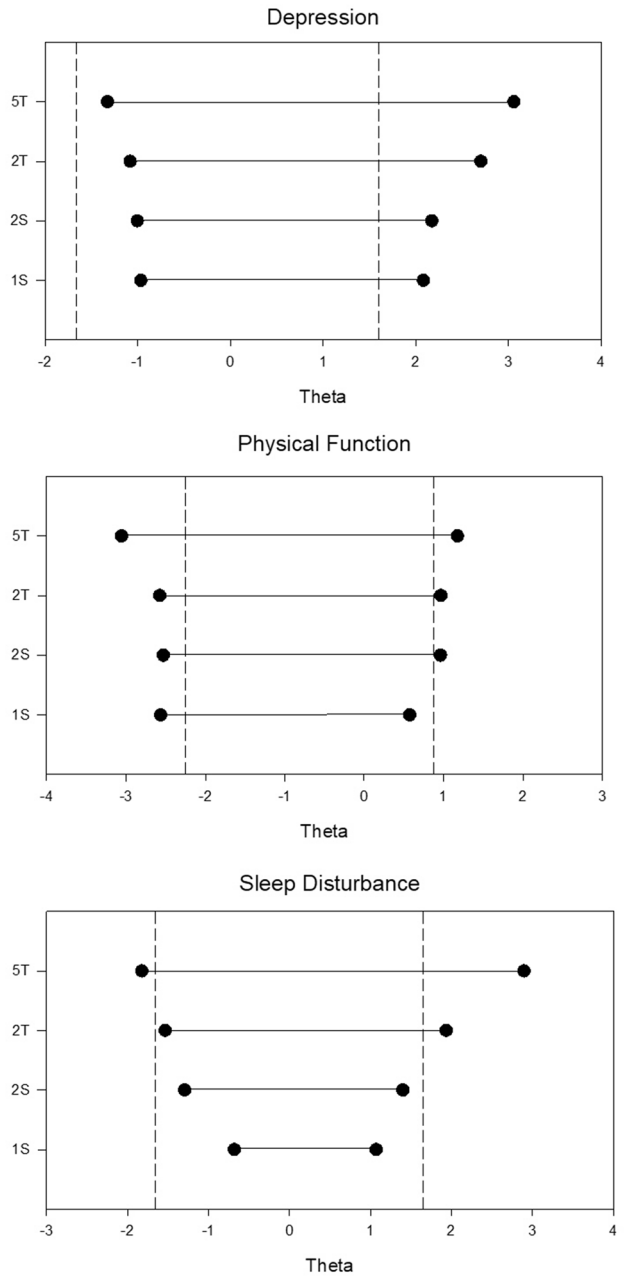


Fig. 1 Range of item bank theta scores captured by each method. IS is a single item, 2S is two items presented separately, 2T is two items presented together, and 5T is five items presented together. Vertical dashed lines represent the area of theta that captures 90% of the item bank calibration sample

more difficult for our participants, who report finding it hard to compare responses to the 2 items, perhaps because we had purposefully chosen the ones that capture different aspects of the health domain. Using the same 2 items to create a single health-state was easier for the participants, captured a wide range of the theta distribution, and had a monotonic relationship to standard gamble valuations. Presenting 5 items together also captured a wide range of the item bank and produced monotonic functions. However, participants found it very complex, cognitively burdensome, and unnecessarily detailed.

To ensure that the widest range of the construct is captured, thereby reducing ceiling and floor effects, we recommend selecting 2 items with values at each end of the theta distribution. We also recommend not relying solely on measurement properties, but also having experts review the content of the items to ensure that they capture key aspects of the health domain (i.e., for depression, selecting items to capture both mood and anhedonia). Once the items are selected, the item information can be used to produce likely combinations of responses for valuation tasks. These valuation tasks can be completed by either the general population or subgroups of interest, such as patients with a specific health condition.

This study has several limitations. First, it used a convenience community sample which is not fully representative of the United States; it had 2.5 times as many Black respondents, 1.6 times as many respondents with educational attainment higher than a bachelor's degree, and one-fifth as many Hispanic respondents as the general US population in the 2010 census. The sample had varying age, race, educational, and health backgrounds from a single city in the United States and we do not believe that the particular geographic area from which participants were sampled would have a significant effect on our findings. Second, we tested four distinct methods to present item banks for valuation, but there are certainly other possible methods which we did not consider. Future work may find that an intermediate approach, such as using 3 or 4 items, may be preferable to using 2. We would recommend testing any other approach using the same criteria as this study: the range of item bank scores captured, monotonicity in mean valuations, participant assessments of difficulty, and semi-structured exit interviews with participants. Third, we assume but do not test that the specific items selected from an item bank will have no impact on the valuations; future work should directly test this assumption. It should be noted that the valuations obtained in this study are not intended for applied research; rather, they were meant to test different methodologies. Fourth, we tested the methods with three different item banks but have not tested the methods in item banks measuring domains like cognition, pain, or social function.

In conclusion, we have developed an acceptable method to present health-state descriptions for IRT-calibrated item banks. This method uses two carefully selected items and presents them in combination. It captures a wide area of the underlying construct, is readily understood by community members, and produces monotonic valuations over the underlying construct. Our recommendation is strengthened by consistent findings in three distinct item banks: depression, physical functioning, and sleep disturbance. While the

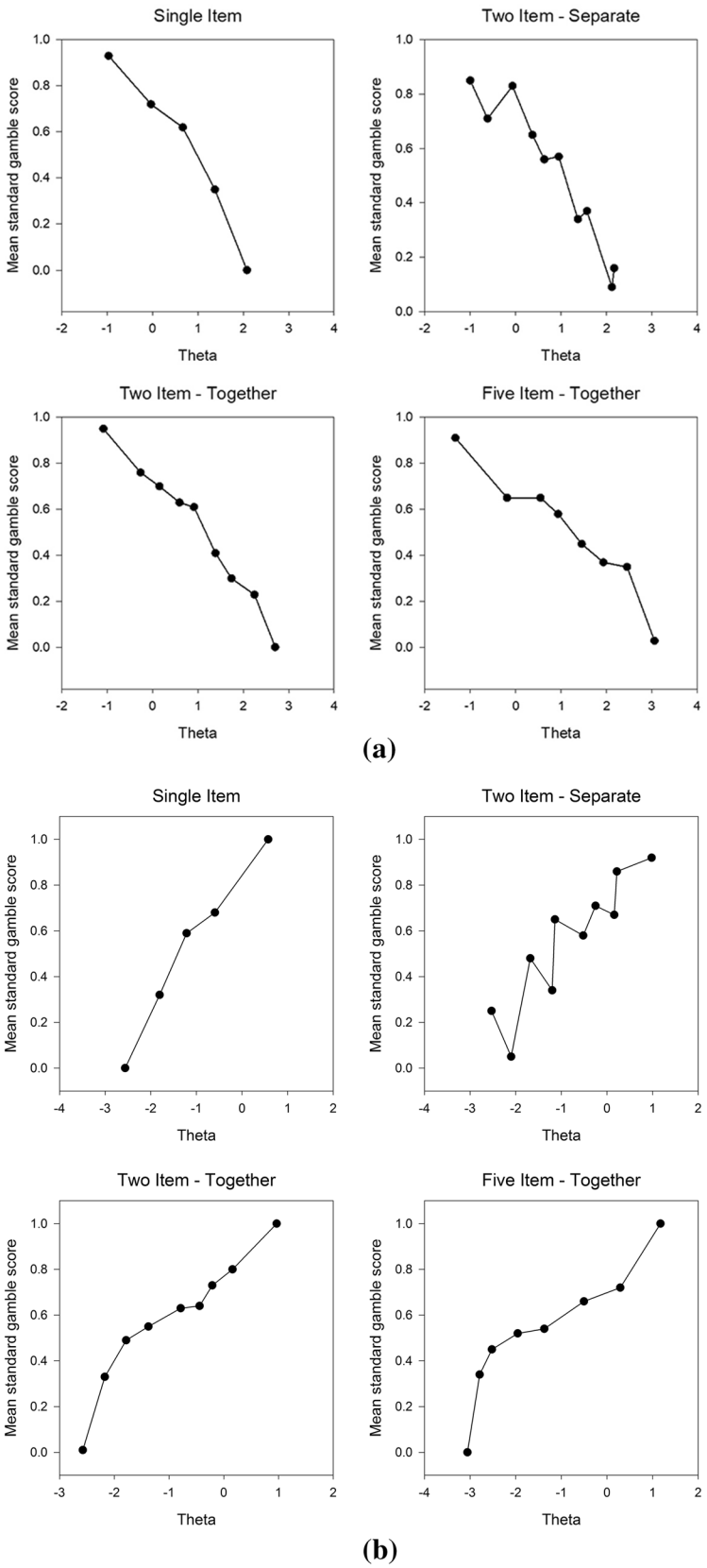
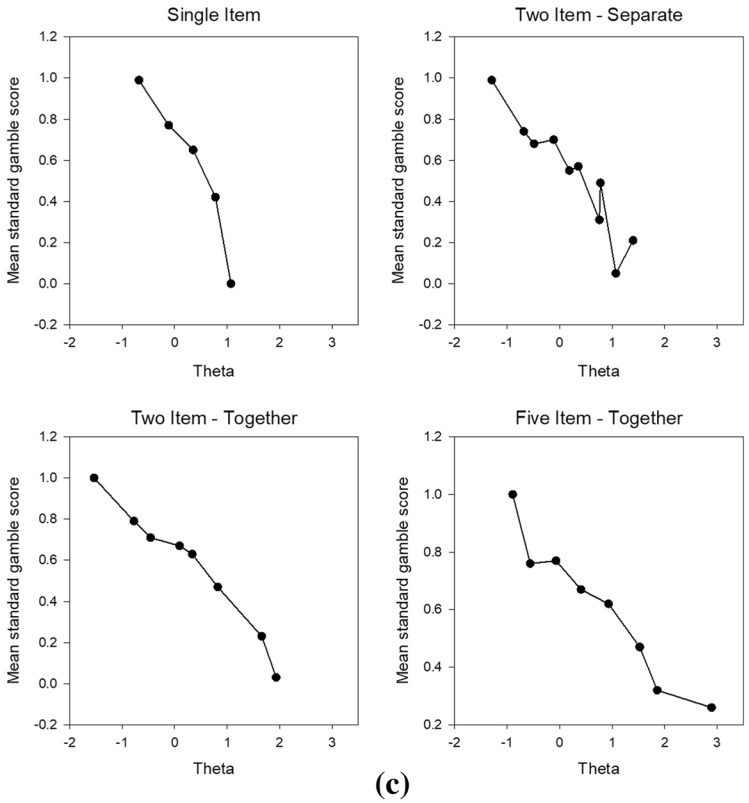


Fig. 2 **a** Mean standard gamble scores for each presentation method in the Depression item bank. **b** Mean standard gamble scores for each presentation method in the Physical Function item bank. **c** Mean standard gamble scores for each presentation method in the Sleep Disturbance item bank

Fig. 2 (continued)



present study used PROMIS as an exemplar, the method can be applied to any descriptive system developed using IRT.

Acknowledgements Janel Hanmer was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number KL2TR001856. The project was supported by the National Institutes of Health through Grants Numbers UL1TR000005 and UL1TR001857, and a supplement to the PROMIS statistical center Grant 3U54AR057951-04S4. The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Compliance with ethical standards

Conflict of interest The authors have no conflicts of interest to report.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. IRB approval for the project was obtained from the University of Pittsburgh (PRO14110193).

References

1. McHorney, C. A. (1999). Health status assessment methods for adults: Past accomplishments and future challenges. *Annual Review of Public Health, 20*, 309–335.
2. Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal: a review. *Journal of health economics*, *5*(1), 1–30.
3. Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford: Oxford University Press.
4. Neumann, P. J., Sanders, G. D., Russell, L. B., Siegel, J. E., & Ganiats, T. G. (Eds.). (2016). *Cost-effectiveness in health and medicine* (2nd ed.). Oxford: Oxford University Press.
5. Brooks, R., Rabin, R., & de Charro, F. (2003). *The measurement and valuation of health status using EQ-5D: A European perspective*. Dordrecht: Kluwer.
6. Feeny, D., Furlong, W., Torrance, G. W., et al. (2002). Multiattribute and singleattribute utility functions for the health utilities index mark 3 system. *Medical Care, 40*, 113–128.
7. Feeny, D., Torrance, G., & Furlong, W. (1996). Health utilities index. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials*. Philadelphia: Lippincott-Raven Press.
8. Kaplan, R. M., & Anderson, J. P. (1988). A general health policy model: Update and applications. *Health Services Research, 23*, 203–234.
9. Brazier, J. E., & Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical Care, 42*, 851–859.
10. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics, 21*, 271–292.
11. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5), S3-11.

12. Cella, D., Riley, W., Reeve, B., Stone, A., Young, S., Rothrock, N., et al. (2010). Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194.
13. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New York: Psychology Press.
14. Cook, K. F., Victorson, D. E., Cella, D., Schalet, B. D., & Miller, D. (2015). Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. *Quality of Life Research, 24*(3), 575–589.
15. Thissen, D., Liu, Y., Magnus, B., Quinn, H., Gipson, D. S., Dampier, C., … Gross, H. E. (2016). Estimating minimally important difference (MID) in PROMIS pediatric measures using the scale-judgment method. *Quality of Life Research, 25*(1), 13–23.
16. Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): Depression, anxiety, and anger. *Assessment, 18*(3), 263–283.
17. Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology, 67*(5), 516–526.
18. Buysse, D. J., Yu, L., Moul, D. E., Germain, A., Stover, A., Dodds, N. E., … Pilkonis, P. A. (2010). Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. *Sleep, 33*(6), 781–792.
19. McNaughton, C. D., Cavanaugh, K. L., Kripalani, S., Rothman, R. L., & Wallston, K. A. (2015). Validation of a short, 3-item version of the Subjective Numeracy Scale. *Medical Decision Making, 35*(8), 932–936.
20. Ironson, G., Solomon, G. F., Balbin, E. G., O'Cleirigh, C., George, A., Kumar, M., … Woods, T. E. (2002). The Ironson-Woods Spirituality/Religiousness Index is associated with long survival, health behaviors, less distress, and low cortisol in people with HIV/AIDS. *Annals of Behavioral Medicine, 24*(1), 34–48.
21. PROMIS Depression Scoring Manual. (2015). Accessed August, 2017, from https://www.assessmentcenter.net/documents/PROMIS%20Depression%20Scoring%20Manual.pdf.
22. PROMIS Physical Function Scoring Manual. (2015). Accessed August, 2017, from https://www.assessmentcenter.net/documents/PROMIS%20Physical%20Function%20Scoring%20Manual.pdf.
23. PROMIS Sleep Disturbance Scoring Manual. (2015). Accessed August, 2017, from http://www.healthmeasures.net/images/promis/manuals/PROMIS_Sleep_Disturbance_Scoring_Manual.pdf.
24. Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., & Hays, R. D. (2010). Representativeness of the Patient-Reported Outcomes Measurement Information System internet panel. *Journal of Clinical Epidemiology, 63*(11), 1169–1178.
25. Dobrez, D., Cella, D., Pickard, A. S., Lai, J. S., & Nickolov, A. (2007). Estimation of patient preference-based utility weights from the functional assessment of cancer therapy—General. *Value in Health, 10*(4), 266–272.
26. Hanmer, J., Feeny, D., Fischhoff, B., Hays, R. D., Hess, R., Pilkonis, P. A., … Yu, L. (2015). The PROMIS of QALYs. *Health and Quality of Life Outcomes, 13*(1), 122.

Affiliations

Janel Hanmer[1] · David Cella[2] · David Feeny[3,4] · Baruch Fischhoff[5] · Ron D. Hays[6] · Rachel Hess[7] · Paul A. Pilkonis[8] · Dennis Revicki[9] · Mark Roberts[10,11] · Joel Tsevat[12] · Lan Yu[10]

David Cella
D-cella@northwestern.edu

David Feeny
feeny@mcmaster.ca

Baruch Fischhoff
baruch@cmu.edu

Ron D. Hays
drhays@ucla.edu

Rachel Hess
rachel.hess@hsc.utah.edu

Paul A. Pilkonis
pilkonispa@upmc.edu

Dennis Revicki
Dennis.Revicki@evidera.com

Mark Roberts
mroberts@pitt.edu

Joel Tsevat
joel.tsevat@uc.edu

Lan Yu
Yul2@upmc.edu

[1] Department of General Internal Medicine, University of Pittsburgh Medical Center, 230 McKee Place, Suite 600, Pittsburgh, PA 15213, USA

[2] Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[3] Department of Economics, McMaster University, Hamilton, ON, Canada

[4] Health Utilities Incorporated, Dundas, ON, Canada

[5] Department of Engineering and Public Policy and Institute for Politics and Strategy, Carnegie Mellon University, Pittsburgh, PA, USA

[6] Division of General Internal Medicine & Health Services Research, UCLA, Los Angeles, CA, USA

[7] Division of Health System Innovation and Research, University of Utah Schools of the Health Sciences, Salt Lake City, UT, USA

[8] Department of Psychiatry, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

[9] Outcomes Research, Evidera, Bethesda, MD, USA

[10] Department of General Internal Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

[11] Department of Health Policy and Management, University of Pittsburgh, Pittsburgh, PA, USA

[12] Division of General Internal Medicine, University of Cincinnati College of Medicine and Cincinnati VA Medical Center, Cincinnati, OH, USA