Volume 12 • Number 4 • 2009
VALUE IN HEALTH

Assessing Items on the SF-8 Japanese Version for Health-Related Quality of Life: A Psychometric Analysis Based on the Nominal Categories Model of Item Response Theory

Yasuharu Tokuda, MD, MPH, Tomoya Okubo, PhD, Sachiko Ohde, EdM, Joshua Jacobs, MD, Osamu Takahashi, MD, MPH, Fumio Omata, MD, MPH, Haruo Yanai, PhD, Shigeaki Hinohara, MD, Tsuguya Fukui, MD, MPH

St. Luke's International Hospital, Tokyo, Japan

ABSTRACT

Objectives: The Short Form-8 (SF-8) questionnaire is a commonly used 8-item instrument of health-related quality of life (QOL) and provides a health profile of eight subdimensions. Our aim was to examine the psychometric properties of the Japanese version of the SF-8 instrument using methodology based on nominal categories model.

Methods: Using data from an adjusted random sample from a nationally representative panel, the nominal categories modeling was applied to SF-8 items to characterize coverage of the latent trait (θ). Probabilities for response choices were described as functions on the latent trait. Information functions were generated based on the estimated item parameters.

Results: A total of 3344 participants (53%, women; median age, 35 years) provided responses. One factor was retained (eigenvalue, 4.65; variance proportion of 0.58) and used as θ. All item response category characteristic curves satisfied the monotonicity assumption in accurate order with corresponding ordinal responses. Four items (general health, bodily pain, vitality, and mental health) cover most of the spectrum of θ, while the other four items (physical function, role physical [role limitations because of physical health], social functioning, and role emotional [role limitations because of emotional problems]) cover most of the negative range of θ. Information function for all items combined peaked at –0.7 of θ (information = 18.5) and decreased with increasing θ.

Conclusion: The SF-8 instrument performs well among those with poor QOL across the continuum of the latent trait and thus can recognize more effectively persons with relatively poorer QOL than those with relatively better QOL.

Keywords: health-related quality of life, information function, item response theory, nominal categories model, psychometric analysis.

Introduction

The Short Form-8 (SF-8) questionnaire is an 8-item instrument that measures general aspect of health-related quality of life (QOL) [1]. The original instrument was developed in the English language (see Appendix) and subsequently translated into Japanese [1,2]. Each administration of the SF-8 generates a health profile of eight dimensions, including general health, physical function, role physical (role limitations because of physical health), bodily pain, vitality, social functioning, mental health, and role emotional (role limitations because of emotional problems). Using the SF-8 to assess QOL has become popular in part because of its ease of administration [3–5].

After the reliability and validity of the Japanese version of the SF-8 was measured for the general Japanese population [2], multiple Japanese researchers have used it as an outcome measure for health-related QOL [6–8]. The QOL scores were assessed using the norm-based scoring (NBS) method outlined in the manual of the original version of the SF-8 [1]. The NBS method can be used to make results comparable across SF-8 and SF-36 (the widely-used instrument for health-related QOL) instruments. The means, variances, and regression weights used to score the Japanese version of the SF-8 were obtained from data in the general Japanese population [2]. Assigning the scale values in the NBS table to the participants' responses generate the person's scores, which can be comparable with those in the Japanese general public (the mean value of 50). A higher response indicates better health. Normative tables for the summary measures by age group and gender are provided in the manual. After contacting the distributors and conducting a literature review, other sources for the information on the reliability, validity, and normative tables could not be located.

Little is known, however, about the relative contribution of the SF-8 constituent items in capturing the full spectrum of health-related QOL. Since the classical test theory [9], which was exclusively used for the previous evaluation of the SF-8 [2], does not address issues related to in-depth characterization of individual items of the SF-8, additional psychometric evaluations are needed. Methodology based on item response theory (IRT) is particularly appropriate for item analysis, such as effects of item location, item discriminating power, and item information functions [10–13]. The IRT was established as a method of applying new test theories designed to overcome some difficulties embedded in classical test theory. The IRT was developed mainly in connection with educational measurements for achievement abilities of students [14]. The IRT is rich in potential applications. In addition to measuring achievement abilities, application of the IRT to psychometric evaluations of surveys such as personality scales has been proposed [15].

Thus, in the current study, we aimed to extend examination of the properties of individual items of the SF-8 instrument to better characterize its ability to map the continuum of health-related QOL. We used data of responses to the SF-8 from a nationally representative panel of the general Japanese population. The items of the SF-8 were composed of two or more choices, so we employed the nominal categories model based on IRT. This modeling technique is an extended model of IRT to analyze nominally scored items [16], and it is more effective in

Address correspondence to: Yasuharu Tokuda, Center for Clinical Epidemiology, St. Luke's Life Science Institute, St. Luke's International Hospital, 9-1 Akashi-cho, Chuo City, Tokyo 104-8560, Japan. E-mail: tokuyasu@orange.ocn.ne.jp

10.1111/j.1524-4733.2008.00449.x

examining the full spectrum of contributions of each item and possible responses in the survey instrument than using a graded categories model or binary model [17,18]. We used a large sample size (n = 3344), enough to meet the requirements of nominal categories modeling.

Methods

Participants

The methodology of the health diary study from which the data for the current analysis were abstracted is described in detail elsewhere [19]. This study was conducted from October 1 to 31, 2003. Briefly, from a nationally representative panel comprising 210,000 households (panel administered by the Japan Statistics & Research Co. Ltd., Tokyo, Japan), we selected a population-weighted random sample of households by controlling for the size of cities, towns, and villages. Using the method of probability proportionate to size, the wards for sampling were selected randomly from a list of total wards for the given area. The higher population wards were given a higher probability of being selected, so that all households had an equal probability of being selected, ensuring that the sample was geographically representative of the area. Each household was sent a soliciting letter with a return envelope for voluntary participation. We obtained prior approval from the Research Ethics Committee of Kyoto University Graduate School of Medicine and informed consent from each participant before the study.

Data Collection

Each member of participant households provided information on demographics, including age, gender, and the area of residence, as well as responses to the SF-8 instrument. Completed questionnaires were submitted by prepaid return envelopes. The area of residence was recorded in one of four categories: 1) large city with a population of more than 1 million; 2) city with a population between 100,000 and 1 million; 3) city or town with a population of less than 100,000; and 4) rural area or village. Surrogate members (typically a parent) responded to the questionnaire on behalf of children under the age of 12 years.

Statistical Analysis

The SF-8 is comprised of eight questions, each with an ordinal response format. Analysis proceeded as follows. First, each item of the SF-8 was scored by assigning a weight derived from NBS estimated from standardized scores of the Japanese general public [2]. Descriptive statistics were also calculated. Confirmatory factor analysis was performed using principal component method to generate factor loadings of 8 items and to confirm the satisfaction of a one-factor solution. We assumed that the unidimensionality assumption would be satisfied if the factor loadings of all items exceeded 0.6 and no other factors would exceed unity. In other words, we may assume that the 8 items all rely on a common latent trait and so may be considered in the single dimension of health-related QOL.

Second, we applied the nominal categories model, and, for the purpose of examining characteristics of item responses, the responses in each item were scored in order from "1" to "5" or to "6," depending on the number of responses possible for each item. The nominal categories model was proposed by Bock [16]. This model was an extended model of the IRT that helped in analyzing nominally scored items. In the nominal categories model, the response probability p_{ijk} that respondent i with a latent trait θ_i response to category k ($k = 1, 2, \ldots, K_j$) of item j is described as follows [20]:

$$p_{ijk} = \frac{\exp(\alpha_{jk}\theta_i + \gamma_{jk})}{\sum_{k'=1}^{K_j} \exp(\alpha'_{jk}\theta_i + \gamma_{jk'})} \quad (1)$$

where K_j denotes the number of the category of item j.
In order to estimate item parameters, Bock imposed arbitrary linear restrictions as follows [16].

$$\sum_{k=}^{K_j} \alpha_{jk} = \sum_{k=1}^{K_j} \gamma_{jk} = 0 \quad \text{for any } j. \quad (2)$$

Moreover, in Okubo [21], the restriction is imposed as follows:

$$\alpha_{j1} = \gamma_{j1} = 0 \quad (3)$$

The response probabilities to the categories in the latent scale θ_i then turn out to be the same between these two methods of restrictions. We cannot interpret the parameters of the categories independently in the nominal categories model because the equation defined for a response probability to the category contains other parameters. The role of the α parameter is that of a slope in the linear function. The larger slope α implies that the item discriminates the latent trait θ_i sharply, while the smaller slope α implies low discrimination ability of the item. On the other hand, the role of the γ parameter is that of an intercept. The larger intercept γ implies that the item is difficult to solve, while the smaller intercept γ implies it is easy to solve.

As mentioned above, however, the Item Response Category Characteristic Curve (IRCCC) is determined by a relative relation among the parameters; thus, each parameter cannot be interpreted alone. The usual method to analyze the characteristics of items is to draw the IRCCC by using the estimated parameters. IRCCC is a multinomial logistic regression curve whose independent variable is a factor and, in this case, the factor is QOL.

We used response data of all available family members in this study. The IRT analysis constructs the likelihood function based on individual response patterns and estimates parameters by using the Marginal Maximum Likelihood Estimation via expectation-maximization (EM) Algorithm. Because correlation matrices between individual data are not used in the process of estimating parameters of the IRT and data with large sample size are used as in this study, intrafamilial correlation is not likely to lead to a significant bias. For instance, if we had used a data set collected from 100 people, each from a big family of 20 members each (five families total), this might have led to a substantial bias.

Finally, item information functions were generated for individual items. Analyzing all items combined generated the instrument information function. Statistical analyses were performed using R version 2.6.6 (R Foundation for Statistical Computing, Vienna, Austria) and graphics were generated using Mathematica version 6.0 (Wolfram Research, Illinois). The codes used in this study for R programming are available from the authors upon request.

Results

Of the 3344 participants who completed the questionnaire, there were 1565 (46.8%) men and 1779 (53.2%) women and the median age was 35 years (range, 0 to 96; 25 percentile, 14; 75 percentile, 52). For the area of residence, 17% of participants lived in large cities, 24% in medium-sized cities, 38% in small-

Table 1 Descriptive statistics of the SF-8 scores and its factor structure from 3344 Japanese respondents

Item	Mean	SD	Minimum	Maximum	Factor loadings*
General health	51.2	7.3	30.4	61.5	0.771
Physical function	50.4	5.4	13.5	53.6	0.772
Role physical	50.1	5.9	15.7	53.9	0.826
Bodily pain	51.1	8.5	28.1	60.2	0.689
Vitality	51.8	6.4	28.3	59.6	0.801
Social functioning	49.4	7.4	20.5	54.7	0.792
Mental health	50.0	6.9	28.8	57.5	0.668
Role emotional	50.3	5.9	13.5	54.3	0.770

*One factor was retained with eigenvalue of 4.65 and variance proportion of 0.58.
SD, standard deviation.

sized cities, and 21% in rural areas. The area of residence and age distribution were consistent with the general Japanese population.

Table 1 presents the descriptive statistics of NBS of SF-8 along with its factor structure. The mean weighted scores of all eight items exhibited about 50 (the standardized mean of the general Japanese population). Based on confirmatory factor analysis, one factor was retained with an eigenvalue of 4.65 and variance proportion of 0.58, and no other factors exceeded unity. Thus, we performed subsequent analyses assuming that there is a single latent trait for the QOL score. Although the SF-8 subscales have two components (physical and mental components), this result indicating unidimensionality of the SF-8 may not be surprising because the physical and mental components are likely to correlate with each other as characteristics of health-related QOL scales.

Figure 1 presents IRCCC for each response to individual items. Positive numbers of the latent trait in the horizontal axis indicate better health-related QOL. The nominal categories model was fit for each response among individual items. All IRCCCs satisfied the monotonicity assumption by ranking with accurate order with corresponding ordinal responses. In items 1 (general health), 4 (bodily pain), 5 (vitality), and 7 (mental health), the IRCCCs convey the coverage of almost the entire spectrum of the latent trait. The curves of items 2 (physical function), 3 (role physical), 6 (social functioning), and 8 (role emotional) cover most of the negative spectrum of the latent trait score.

Figure 2 shows item information function for individual items. Two patterns for information function spectrums are demonstrated. Similarly, the curves of items 1, 4, 5, and 7 generate the information of almost the entire spectrum of the latent trait and are fairly even spread. On the other hand, the curves of items 2, 3, 6, and 8 show information functions mostly in the negative spectrum.

Lastly, Figure 3 presents the overall instrument information function curve from all information of eight items combined. The area of information function curve of this instrument is larger in negative scores of the latent trait (also referred to as θ) than in positive scores. Information peaked at -0.7 of θ (information = 18.5) and decreased with increasing θ. Thus, the instrument has discriminating power among people with lower scores of the latent trait than among those with higher scores.

Discussion

The current study characterizes items of the SF-8 using the nominal categories model based on IRT. All responses of all items satisfy the monotonicity assumption and thus, these responses are considered as correctly ranked in order. Although several items convey the coverage of almost the entire spectrum of the latent trait, some of them preferentially cover the negative range of the latent trait score and, overall, persons with relatively poor QOL can be more effectively recognized than those with relatively good QOL. Because the mean weighted scores of all eight items in our study participants exhibited about 50, which is the standardized mean of the general Japanese population, our study participants well represent the general Japanese population and thus, the item characteristics of our study results may be generalizable to the Japanese public.

Based on the findings of the current study, we support use of the SF-8 for the measure of health-related QOL. The instrument performs well in assessing QOL among those with moderately poor QOL. Public health policies should be implemented for improving QOL among people with very poor QOL as well as those with moderately poor QOL [22]. Thus, from a public health perspective, it is important to characterize more effectively those with very poor and moderately poor QOL than those with relatively better QOL. We can safely say that the SF-8 well deserves to do just that.

We extended the evaluation of the SF-8 by using item responses coded into nominal categories. We successfully elucidated the individual discriminating power and item position effect in the SF-8 by using the nominal categories model. These findings may help identify the response categories with good psychometric characteristics and those without. For instance, among six response categories given to the item 4 (bodily pain), the sixth response category may be eliminated because of its marginal location and gentle slope, without compromising the item characteristics and information function. Furthermore, our findings may also help identify a single general item with psychometric properties covering the entire spectrum of the latent trait. For instance, item 1 (general health) may be useful as a single item QOL measure in cases where only a single item is feasible.

Because the results of our study were based on a different cultural environment (Japan) other than that of the instrument's original development (USA), a limit to the generalizability of the current findings must be considered. Although development of the Japanese version of the SF-8 included a careful evaluation based on strict procedural guidelines [2,23], comparable studies are needed using IRT on the original version of SF-8 in English as well as in other languages.

In summary, the current study is the first to explore all items of the SF-8 Japanese version using a methodology of IRT. Our in-depth analysis characterizes psychometric properties of this QOL instrument by showing that the instrument performs well among those with relatively poor QOL in terms of discriminating power and position across the continuum of the latent trait. Because it is important to characterize more effectively those with relatively poor QOL, our results support use of the SF-8 Japanese version as a convenient measure of QOL.

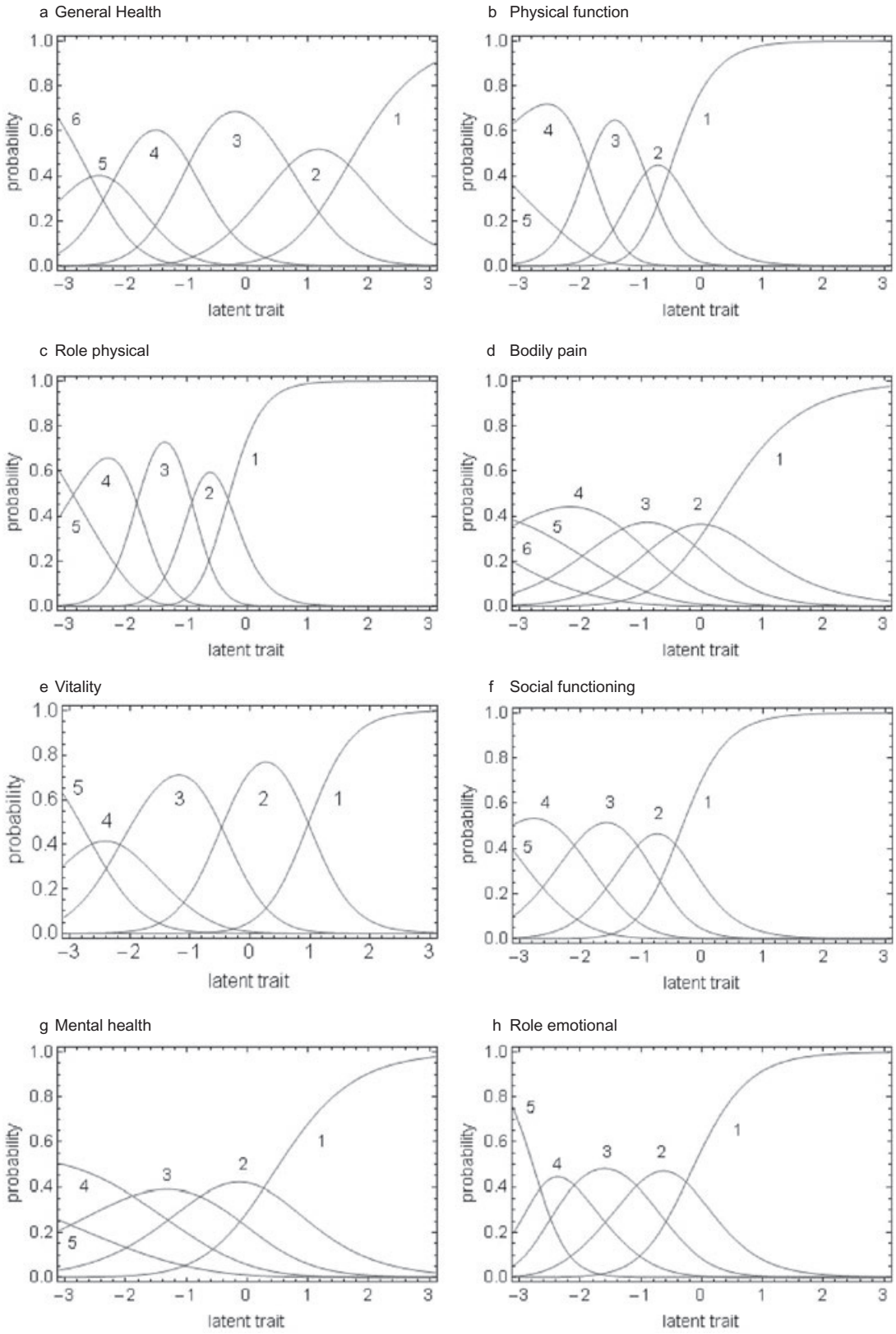

Figure 1 Item response category characteristic curves for each item of the SF-8 generated by the analysis of data from 3344 Japanese respondents.

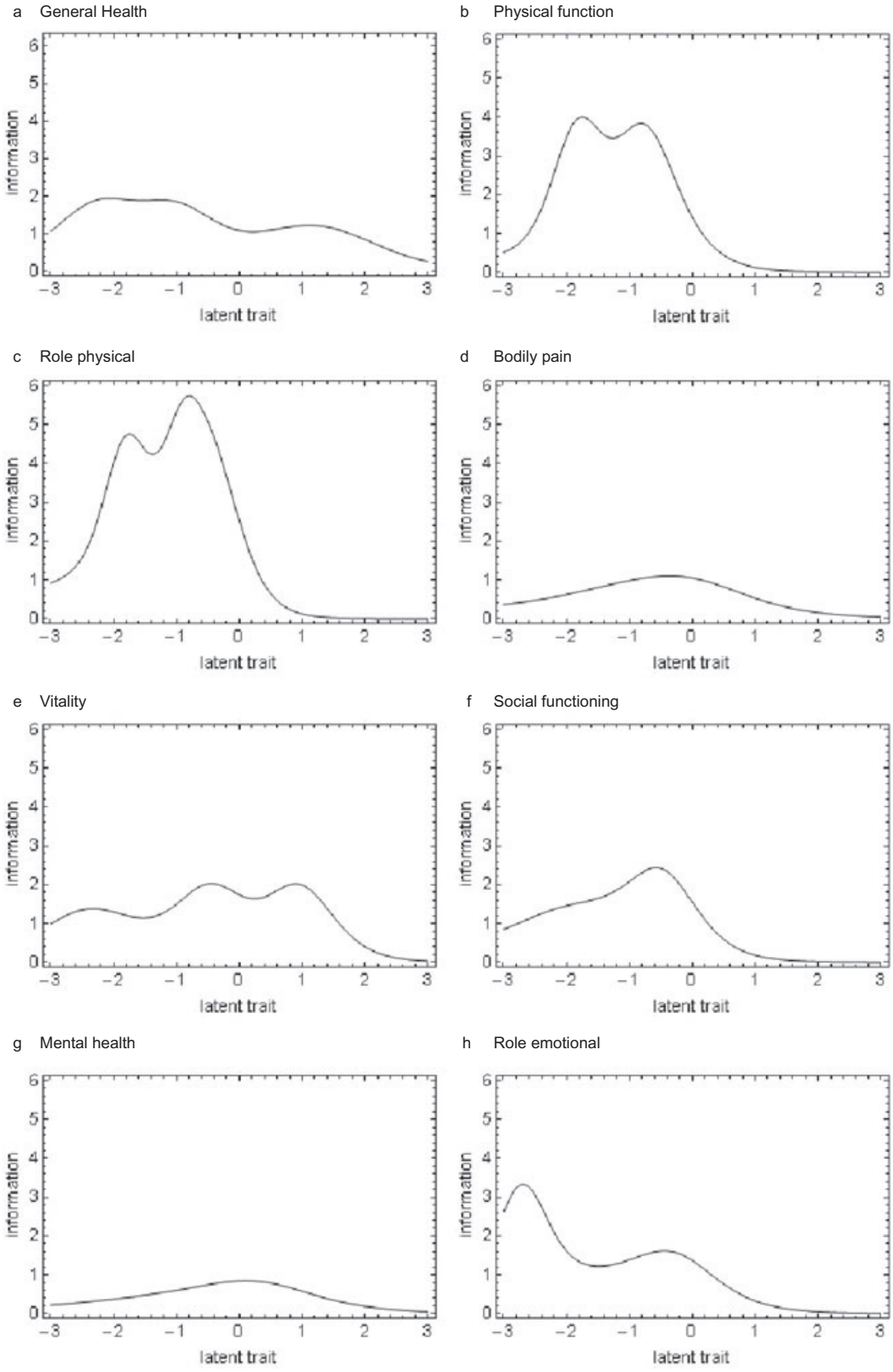

Figure 2 Item information function of each item of the SF-8 generated by the analysis of data from 3344 Japanese respondents.

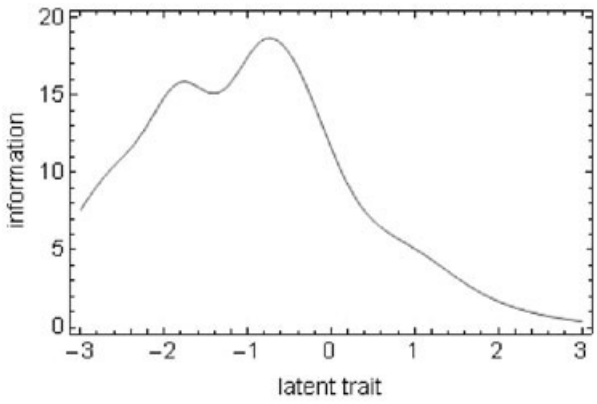

Figure 3 Overall item information function of the SF-8 instrument generated by the analysis of data from 3344 Japanese respondents.

We wish to thank Kenji Sakurai, MD, FACS, for his support of our research. We also wish to thank Mikio Kumagai, Masato Ichikawa, Yumiko Yotsumoto, Yuko Iwasawa, Sayoko Yamauchi, and Kayo Ichikawa for their excellent secretarial assistance. We also thank Shunichi Fukuhara, MD, MSc, Department of Epidemiology and Healthcare Research, Kyoto University Graduate School of Medicine and Public Health, for kindly allowing us to use the SF-8 Japanese version in our study.

Source of financial support: Intramural research grant of St. Luke's Life Science Institute.

Supporting information for this article can be found at: http://www.ispor.org/publications/value/ViHsupplementary.asp

References

1 Ware JE, Kosinski M, Dewey JE, Gandek B. How to Score and Interpret Single-Item Health Status Measures: A Manual for Users of the SF-8 Health Survey. Lincoln, RI: Quality Metric Inc., 2001.
2 Fukuhara S, Suzukamo Y. Manual of the SF-8 Japanese Version, (in Japanese). Kyoto: Institute for Health Outcomes and Process Evaluation Research, 2004.
3 Wu C, Volk RJ, Steinbauer JR, et al. A wireless mobile health-related quality of life assessment. AMIA Annu Symp Proc 2003;1:1054.
4 Shanafelt TD, West C, Zhao X, et al. Relationship between increased personal well-being and enhanced empathy among internal medicine residents. J Gen Intern Med 2005;20:559–64.
5 Lefante JJ Jr, Harmon GN, Ashby KM, et al. Use of the SF-8 to assess health-related quality of life for a chronically ill, low-income population participating in the Central Louisiana Medication Access Program (CMAP). Qual Life Res 2005;14:665–73.
6 Shiozaki M, Hirai K, Dohke R, et al. Measuring the regret of bereaved family members regarding the decision to admit cancer patients to palliative care units. Psychooncology 2007;16:1142–7.
7 Shibata A, Oka K, Nakamura Y, Muraoka I. Recommended level of physical activity and health-related quality of life among Japanese adults. Health Qual Life Outcomes 2007;5:64–71.
8 Uramoto H, Kagami S, Iwashige A, Tsukada J. Evaluation of the quality of life between inpatients and outpatients receiving cancer chemotherapy in Japan. Anticancer Res 2007;27:1127–32.
9 Gulliksen H. Theory of Mental Tests. Hillsdale, NJ: L. Erlbaum Associates, 1987.
10 Haley SM, McHorney CA, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47:671–84.
11 Noerholm V, Groenvold M, Watt T, et al. Quality of life in the Danish general population—normative data and validity of WHOQOL-BREF using Rasch and item response theory models. Qual Life Res 2004;13:531–40.
12 Sijtsma K, Emons WH, Bouwmeester S, et al. Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). Qual Life Res 2008;17:275–90.
13 Samejima F. Estimation of latent trait ability using a response pattern of graded scores. Psychometrika Monograph 1969.
14 Lord FM. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1980.
15 Kelly CW. Commitment to health scale. J Nurs Meas 2005;13:219–29.
16 Bock R. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika 1972;39:29–51.
17 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38(Suppl. 9):SII28–42.
18 Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof 2005;28:264–82.
19 Fukui T, Rhaman M, Takahashi O, et al. The ecology of medical care in Japan. JMAJ 2005;48:163–7.
20 Thissen D, Steinberg L. A response model for multiple-choice items. In: Linden W, Hambleton R, eds. Handbook of Modern Item Response Theory. New York: Springer, 1997.
21 Okubo T. An item parameter estimation programme for nominal categories model using R. DNC Research note. 2007;RN-07-18.
22 Hays RD, Liu H, Spritzer K, Cella D. Item response theory analyses of physical functioning items in the medical outcomes study. Med Care 2007;45(Suppl. 1):S32–8.
23 Shim EJ, Mehnert A, Koyama A, et al. Health-related quality of life in breast cancer: a cross-cultural survey of German, Japanese, and South Korean patients. Breast Cancer Res Treat 2006;99:341–50.