Qual Life Res (2009) 18:605–618
DOI 10.1007/s11136-009-9471-5

Psychometric evaluation of the WHOQOL-BREF in community-dwelling older people in Taiwan using Rasch analysis

Wen-Miin Liang · Chih-Hung Chang · Yi-Chun Yeh · Haw-Yaw Shy · Hung-Wei Chen · Mau-Roung Lin

Accepted: 10 March 2009 / Published online: 2 April 2009

Abstract
Objective To examine the psychometric characteristics of the brief version of the World Health Organization Quality of Life (WHOQOL-BREF) questionnaire in rural-community-dwelling older people in Taiwan using Rasch analysis.
Methods This is a cross-sectional study. A total of 1200 subjects aged ≥65 years were recruited to complete the 26-item WHOQOL-BREF. Scale dimensionality, item difficulty, scale reliability and separation, item targeting, item-person map, and differential item functioning (DIF) were examined.
Results The four WHOQOL-BREF scales (physical capacity, psychological well-being, social relationships, and environment) were found to be unidimensional and reliable. The item–person map for each domain indicated that the spread of the item thresholds sufficiently covered the latent trait continuum being measured. However, gaps in content coverage were identified in the social domain. Analyses of the DIF revealed that one psychological item (body image) exhibited DIF across the two age groups (old–old vs. young–old) and that two social items (sexual activity and friends' support) displayed DIF across genders and the two age groups.
Conclusions Rasch analysis is a comprehensive method of psychometric evaluation of the WHOQOL-BREF and identifies areas for improvements. Three items displaying age-related DIF (body image, sexual activity, and friends' support) may potentially cause biased health-related QOL assessments, and their impacts on scores should be carefully examined.

Keywords Health-related quality of life · Item response theory · Older people · Rasch model · WHOQOL-BREF

Abbreviations

CTT	Classical test theory
DIF	Differential item functioning
HRQL	Health-related quality of life
IRT	Item response theory
WHOQOL-BREF	World Health Organization Quality of Life-BREF

W.-M. Liang
Graduate Institute of Biostatistics, Institute of Environmental Health, Biostatistics Center, China Medical University, Taichung, Taiwan, ROC

C.-H. Chang
Buehler Center on Aging, Health & Society, Division of General Internal Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

C.-H. Chang
Graduate Institute of Biostatistics, China Medical University, Taichung, Taiwan, ROC

Y.-C. Yeh
Biostatistics Center, China Medical University, Taichung, Taiwan, ROC

H.-Y. Shy
Department of Mathematics, National Changhua University of Education, Changhua, Taiwan, ROC

H.-W. Chen
Institute of Environmental Health, China Medical University, Taichung, Taiwan, ROC

M.-R. Lin (✉)
Institute of Injury Prevention and Control, Taipei Medical University, 250 Wu-Hsing Street, Taipei, Taiwan, ROC
e-mail: mrlin@tmu.edu.tw

Introduction

With the increasing acceptability of the concept of health-related quality of life (HRQL) as a measure of healthcare outcomes over the past two decades, many HRQL instruments have been developed and refined [1, 2]. In 1990s, the World Health Organization (WHO) initiated a worldwide 15-country project to develop a brief version of the WHO's quality of life instruments (WHOQOL-BREF) [3, 4]. The WHOQOL-BREF, a generic HRQL instrument designed for healthy people and patients, has been rapidly introduced and utilized as an outcome measure in many countries and cultures [5–10]. It has shown excellent psychometric and clinimetric properties, such as test–retest reliability, internal consistency, discriminant validity, construct validity, and responsiveness to community-dwelling older people, primarily based on the principles of classical test theory (CTT) [11].

Item response theory (IRT), originally developed in the fields of education and psychology, has proven to be a powerful tool for health outcome assessment [12, 13]. This tool comprises a set of statistical models suitable for analyzing a scale or survey instrument with multiple items that measure the same construct (e.g., physical functioning). An IRT model specifies how both person–trait level and item characteristics are related to a person's item responses. This is different from the CTT approach in which items and the person latent trait being measured are considered separately and, therefore, cannot be meaningfully and systematically compared [14, 15]. Unlike the manner in which the trait level is estimated in the context of an IRT model, in the CTT method, true score estimates are typically obtained by summing responses across items. In CTT, the two assumptions made to support the summed scores (each item within the same construct is valued equally ,and the score difference between the two adjacent response scales is uniform) do not hold in most situations [16]. These limitations can be solved rationally using IRT modeling even though IRT also makes assumptions, such as unidimensionality and local independence. Many useful statistics, such as the reliability and separation index, can be calculated directly from the IRT model, and differential item functioning (DIF) or item bias can be examined for measurement invariance [14, 15].

The WHOQOL-BREF is a generic measure, and this enables it to be used in different cultures and disease groups. It can be an effective tool for assessing HRQL in the context of a cost-effectiveness evaluation and health intervention in a large-scale study, such as insurance policies for different diseases and welfare policies for different age groups among older people [17, 18]. However, two aspects of the measurement need to be considered when applying a generic HRQL measure to older people. First, the items and domains in a HRQL measure should take into account the aspects of life (access to health services and environmental safety) identified as important by older people [19]. Although the HRQL aspects considered by older people be relevant and important to health may be similar to those of young people, the relative importance and definition of each domain or facet (e.g., role functioning may be defined in terms of roles other than work) may differ substantially [20, 21]. Second, extreme low or high scores on HRQL measures may appear to be common among older people if these measures are directly used without proper adaptation. Floor effects can potentially reduce the ability to detect distinguishing features, such as those between ill and very ill people, for example, and may also reduce the ability to detect changes in HRQL scores over time or after a health intervention [22]. In addition, it is important for physicians and healthcare decision-makers to be aware of potential biases in the instruments being used. For example, some response instability may have to do with the aging process when the study group comprises older subjects [23]. An accurate and precise scoring of the WHOQOL-BREF can lead to a more efficient allocation of healthcare resources or to more efficient screening of patients for specialized geriatric care. Since findings of previous WHOQOL-BREF studies were subject to the limitations of CTT methods, we have applied the Rasch model to examine psychometric properties of the WHOQOL-BREF among community-dwelling older people in rural Taiwan. This is a population subgroup that has to date not received as much attention as urban-dwelling elders with regard to HRQL.

Methods

Study subjects and procedures

This was a cross-sectional study. Elderly subjects aged 65 and older were recruited from six of 15 villages of the Shin-Sher Township of Taichung County in west–central Taiwan. These villages were selected according to the order of the magnitude of populations of the Shin-Sher Township and were found to be representative of rural community-dwelling older people in Taiwan [11]. A postcard describing the aim of the study and the schedule of interviews was mailed to all eligible residents in the six villages. During a 2-week interview period, 1200 of the 2072 eligible subjects signed informed consent forms to participate and completed the WHOQOL-BREF. Of the 872 eligible elderly who did not participate, 24 had died, 59 were hospitalized or bed-ridden, 252 had moved out of the area, 323 were not at home during the assessment period, and 214 declined to be interviewed. The respondents had

similar distributions in gender and educational level, but tended to be younger ($P = 0.073$) when compared with non-respondents [11].

Each interviewer completed a 4-h training session to ensure that they followed standardized procedures in conducting the face-to-face interviews. Each enrolled participant was interviewed by a trained interviewer using the structured questions from the WHOQOL-BREF, which included questions related to demographic characteristics, life habits, and medical history.

The WHOQOL-BREF

The standard WHOQOL-BREF contains 26 items: two items from the overall QOL and general health facet and one item from each of the remaining 24 health-related facets [24]. The Taiwanese version of the WHOQOL-BREF (see Appendix) was developed in compliance with the WHO guidelines on translation procedures as well as the design and selection of appropriate items [25]. The 28 items of the Taiwanese WHOQOL-BREF include two general items (i.e., G1: overall QOL; G4: general health), 24 items universally adopted for WHOQOL-BREF to cover four domains (namely, physical, psychological, social, and environment), plus two national items that were more specific for the culture of people of Taiwan (i.e., being respected/accepted among people, and eating what one loves to eat). This translated version showed good reliabilities (including internal consistencies of 0.70–0.77 and test–retest reliabilities of 0.76–0.80) and validities (including content, criterion, discriminant, predictive, and construct validities) [26–28]. Each domain score was calculated by multiplying the mean of all facet scores in the same domain by a factor of four, with a higher scoring indicating a better QOL (range 4–20). We used abbreviations, with a single letter representing each of the four domains to denote items for ease of reference in this paper: P for physical, Y for psychological, S for social, and E for environmental. The number attached to the letter indicated the item number in the questionnaire. For example, S21 (satisfaction with sexual life) indicates item 21, which belongs to the social domain.

Rasch analysis

Partial credit model

We used the partial credit model, an extension of the dichotomous Rasch model, as it is suitable for an ordered polytomous response scale used in the WHOQOL-BREF [29]. As modeled, an item with five response categories would have four threshold parameters. At each threshold, a person has a 50/50 chance of choosing one category over another [16]. For example, for a 5-point Likert scale (1: strongly dissatisfied, 2: dissatisfied, 3: moderately satisfied, 4: satisfied, 5: strongly satisfied), the first threshold is modeled as the value at which the probability of choosing a response of 2 (dissatisfied) over a response of 1 (strongly dissatisfied) is equal to 50%. The item difficulty estimates within each domain were standardized with a mean of 0 and a standard deviation of 1 (in logit units) [30].

Unidimensionality

In the Rasch model, item difficulty and a person's latent trait are modeled to share the same scale with a single unit of logit, i.e., the unidimensionality of the scale [31, 32]. Confirmatory factor analysis was used to assess the property of unidimensionality using LISREL 8.72 (Scientific Software International, Lincolnwood, IL) [33]. Unidimensionality, the measurement of one underlying construct, was determined by the magnitude of factor loadings, with a value >0.3 indicative of importance, and three model fit indices—goodness-of-fit index (GFI), comparative fit index (CFI), Bentler–Bonett normed fit index (NFI)—with an index >0.9 indicative of good fit [34, 35]. We also applied a two-index presentation strategy suggested by Hu and Bentler [36], using the indices standardized root mean square residual (SRMR) ≤ 0.08 and CFI >0.95 to confirm the model fit. After unidimensionality was established by confirmatory factor analysis (CFA), the Rasch model analysis was employed and the item infit statistic was used to further evaluate item–level model fit. An item with an infit statistic >1.4 or <0.6 was used to indicate the lack of fit in unidimensionality [37]. Specifically, an infit value >1.4 indicates that the item does not contribute to the same underlying construct or differs from other items in the same scale in its ability to discriminate among persons. An infit value <0.6 indicates that the item is muted or often has interdependence with other items in the same scale and may occur when several items are similar or highly correlated or when one item is dependent on another [38].

Reliability and separation

Person reliability, another fit statistic of the Rasch model, was used to test the internal consistency among items at the domain level. Similar to Cronbach's α, a value close to 1 indicates high internal consistency and a value <0.7 indicates model misfitting [16, 39]. The person separation is a measure of the ability of items to discriminate subjects and was used to assess the ability of items to spread the elderly subjects along the HRQL continuum being measured in this study. Person separation statistics ranging between 1.5 and 2.0 were considered to be acceptable, 2.0–3.0 to be good, and >3.0 to be excellent [39].

Other aspects regarding item difficulty and personal ability

Targeting: Perfect targeting is defined as the equivalence of the average personal trait and the mean item difficulty. A targeting index, the average of personal ability, was used to examine whether the level of difficulty of each WHOQOL-BREF domain was appropriate for our sample. A targeting index >0 indicates that the subjects tend to give 'positive' responses (e.g., 'satisfied'), and a value <0 indicates that the subjects tend to give 'negative' responses (e.g., 'dissatisfied'). Values of between 0.5 and 1.0 or between −1.0 and −0.5 are considered to be slight mis-targeting, and those >1 and <−1 are considered to be substantially mis-targeting [39, 40].

Range and gap: The range between the highest and lowest threshold values of all items within a domain is considered to be good when it covers at least 95% of the levels of personal ability [41]. A gap is defined as the difference in two adjacent item difficulties which are ≥1 logit [42], and it implies that the item calibrations for a particular domain of the WHOQOL-BREF are not evenly spread or the number of items within the domain is not sufficient.

Ceiling and floor effects: The percentages of ceiling (highest) or floor (lowest) values among all subjects for each WHOQOL-BREF domain were used to assess the extent to which the latent HRQL trait of a subject is not reliably discriminated at both extremes. The ceiling effect for a particular WHOQOL-BREF domain is defined as the number of people with a level of personal latent trait greater than the highest (i.e., the fourth) threshold, and the floor effect as the number of people with a level of a personal latent trait less than the lowest (i.e., the first) threshold [41].

Item-person map: The relationship between item difficulties and personal latent traits for each domain was also examined by plotting the item difficulties and person measures together along the same line, referred to as an item-person map in the Rasch model analysis. The distribution of item difficulties allowed us to identify regions along the latent continuum that may be lacking items for reliable assessment.

Differential item functioning (DIF)

Differential item functioning refers to an item lacking measurement equivalence in different groups or settings [16]. In this study, sets of item difficulties were compared between genders (males vs. females) and between two age groups (young–old: 65–74 years vs. old–old: ≥75 years) to detect DIF. Due to our large sample size and the different facets in each domain of the WHOQOL-BREF, the results may be more sensitive to statistical testing to show a significant difference in score comparisons. We therefore employed the same methodological approach as those used in other similar studies [7]. A criterion of 0.5 logits between item difficulties in different groups was applied to determine whether an item exhibited DIF [7, 43]. All Rasch analyses were performed using Winsteps software ver. 3.47 [30].

Results

Table 1 shows the demographic characteristics of the study sample. This study cohort ($n = 1200$) consisted of elderly subjects aged between 65 and 103 years (mean 73.4 years); 59% ($n = 709$) were male, 86% ($n = 1032$) had an elementary school of education or less, 66% ($n = 792$) lived with their spouse, 68% ($n = 821$) had been diagnosed with

Table 1 Demographic characteristics of the study sample ($n = 1200$ community-dwelling older people)

Variable	*n*	Percentage
Age	73.4 (6.04)[a]	65–103[b]
Gender		
Male	709	59.08
Female	491	40.92
Education		
Elementary school or below	1032	86.00
Higher than elementary school	168	14.00
Living with spouse		
Yes	792	66.00
No	408	34.00
Number of physician-diagnosed chronic conditions		
1	353	43
2	225	27
>3	243	30
Depression		
No	1135	94.74
Yes (GDS score > 10)	63	5.26
Cognitive impairment		
No	980	81.67
Yes (MMSE score > 18)	220	18.33
WHOQOL-BREF domain score		
Physical	13.31 (2.25)[a]	0.77[c]
Psychological	12.76 (2.14)[a]	0.79[c]
Social	13.16 (1.92)[a]	0.73[c]
Environmental	13.08 (1.99)[a]	0.79[c]

GDS Geriatric depression scale, *MMSE* mini-mental state examination, *WHOQOL-BREF* World Health Organization Quality of Life-BREF

[a] Mean, with the standard deviation given in parenthesis

[b] Range

[c] Cronbach's α

at least one chronic disease (353 with one; 225 with two; 243 with three or more), and 5% ($n = 63$) had suffered from depression. All of the four domains of the WHOQOL-BREF showed good reliability, with all Cronbach's α values >0.7. The mean QOL scores were 13.3 for the physical, 12.8 for the psychological, 13.2 for the social, and 13.1 for the environmental domain.

Unidimensionality

Domain-specific CFA results showed that the three indices for the two domains, psychological (GFI 0.98, CFI 0.97, NFI 0.96) and social (GFI 1.00, CFI 1.00, NFI 1.00), were >0.95. The SRMR were 0.03 and 0.01 for each of these two domains, respectively. The three indices for the other two domains, physical (GFI 0.98, CFI 0.97, NFI 0.97) and environmental (GFI 0.97, CFI 0.95, NFI 0.93), increased to ≥0.95 when one or two pairs of error covariance were added separately for each of the two domains; the SRMR were 0.04 and 0.04, respectively. The standardized factor loading for each item in its respective domain, namely physical (0.28–0.83), psychological (0.44–0.76), social (0.47–0.80), and environmental (0.45–0.65), was >0.4 with one exception. Item 'P4 dependence on medication or treatment' had the lowest factor loading of 0.28. When the item P4 was excluded in the analysis, only negligible changes in GFI, CFI, and NFI statistics were found, indicating that the assumption of unidimensionality of each WHOQOL-BREF domain was confirmed.

Infit statistics and item difficulties

Table 2 shows the model fit index, item difficulty estimates, and their standard errors as well as the person reliability and separation indices for each domain. All of the infit statistics fell in the 0.7–1.4 ranges, thereby

Table 2 Results of the Rasch analysis of the four WHOQOL-BREF domains

Domain/Item	Infit index	Item difficulty (SE)	Reliability (separation)
Physical domain			0.76 (1.79)
P3 Pain and discomfort	1.05	−1.1 (0.04)	
P4 Dependence on medication or treatment	1.35	−0.6 (0.04)	
P16 Sleep and rest	1.20	0.16 (0.05)	
P17 Activities of daily living	0.72	0.21 (0.05)	
P15 Mobility	0.90	0.42 (0.04)	
P18 Working capacity	0.83	0.42 (0.05)	
P10 Energy and fatigue	0.93	0.48 (0.05)	
Psychological domain			0.77 (1.83)
Y26 Negative feeling	1.38	−1.19 (0.05)	
Y19 Self-satisfaction	0.97	−0.61 (0.06)	
Y11 Body image and appearance	0.95	−0.52 (0.06)	
Y6 Spirituality, religion and personal beliefs	0.82	0.37 (0.06)	
Y7 Thinking, learning, memory, and concentration	0.95	0.67 (0.05)	
Y5 Enjoyment of life	0.87	1.29 (0.05)	
Social domain			0.68 (1.45)
S22 Friends' support	0.81	−0.77 (0.07)	
S20 Personal relationship	0.84	−0.71 (0.07)	
S27 Esteem and respect	1.24	0.31 (0.07)	
S21 Sexual activity	1.04	1.17 (0.08)	
Environmental domain			0.78 (1.86)
E28 Eating food	1.17	−1.08 (0.05)	
E9 Physical environment	1.14	−0.63 (0.05)	
E23 Home environment	0.80	−0.56 (0.05)	
E25 Transportation	0.96	−0.03 (0.05)	
E24 Health and social care: availability and quality	0.96	0.01 (0.05)	
E8 Physical safety and security	0.99	0.1 (0.04)	
E13 Opportunities	0.86	0.29 (0.05)	
E12 Financial environment	1.00	0.94 (0.04)	
E14 Participation and support of leisure activities	1.15	0.96 (0.04)	

P physica, *Y* psychologica, *S* social, *E* environmental, *SE* standard error

satisfying the unidimensionality assumption of the IRT model. The items in each domain are listed in Table 2 in ascending order of their difficulty estimates. For example, the item 'P10 energy and fatigue' was the most difficult item in the physical domain. The ranges of the average difficulty estimates of the physical, psychological, social, and environmental categories were −1.10–0.48, −1.19–1.29, −0.77–0.17, and −1.08–0.96, respectively. Four items in the social domain had relatively large standard errors, with item 'S21 satisfaction with sex life' having the largest standard error of 0.08.

Reliability and separation

As shown in Table 2, the person reliability for each of the WHOQOL-BREF domain was acceptable, with reliability coefficients ranging from 0.68 (social domain) to 0.78 (environmental domain). Separation indices for the physical, psychological, social, and environmental domains were 1.79, 1.83, 1.45, and 1.86, respectively, suggesting that all but the social domains had acceptable separation properties.

Item-person map

Figures 1, 2, 3, and 4 show the four item–person maps depicting the person latent trait estimates (upper panel) and sets of threshold parameter estimates (represented by the numbers 1–4 for each item in the lower panel) jointly located along the same 'logit' scale. For example, the fourth threshold of the item 'P3 pain and discomfort' shown in Fig. 1 had a lower logit value (1.23) than the fourth threshold of the item 'P10 energy and fatigue' with a higher logit value (5.30). This indicates that a subject with a latent trait level of between 1.23 and 5.3 logits is more likely to choose the fifth response category for item 'P3 pain and discomfort' and the fourth response category for item 'P10 energy and fatigue'. Furthermore, the sequences of the four threshold parameters for each item were all in order from less (category 1) to more (category 4). When the data do not fit the model, the thresholds may be out of order (e.g., 1, 2, 4, and 3). It should be noted that the mean of the threshold parameters for each item is the item difficulty estimate listed in Table 2.

The average values of person traits (i.e., the targeting indices) were 0.57, 0.44, 1.02, and 0.36 for the physical, psychological, social, and environmental domains, respectively, indicating that the study population had a higher tendency to select more 'positive' response categories in each domain. However, very few subjects had a HRQL score greater than the highest threshold (0.5, 0.33, 0.17, and 0.5% for the physical, psychological, social, and environmental domains, respectively) or less than the lowest threshold (0.33, 0.00, 0.08, and 0.25 for the physical, psychological, social, and environmental domains, respectively), suggesting that there was no significant ceiling or floor effect for any of the four domains. In addition, the ranges of the thresholds in each domain (−3.9–5.3 for physical, −6.7–6.8 for psychological, −8.4–9.7 for social, and −3.7–4.8 for environmental) covered at least 95% of subjects' traits (which ranged from −1.5 to 3.3, from −2.4 to 3.3, from −3.3 to 5.2, and from −1.6 to 2.4 for the physical, psychological, social, and environmental domains, respectively), indicating that the WHOQOL-BREF provided a satisfactory estimate for most subjects in this study.

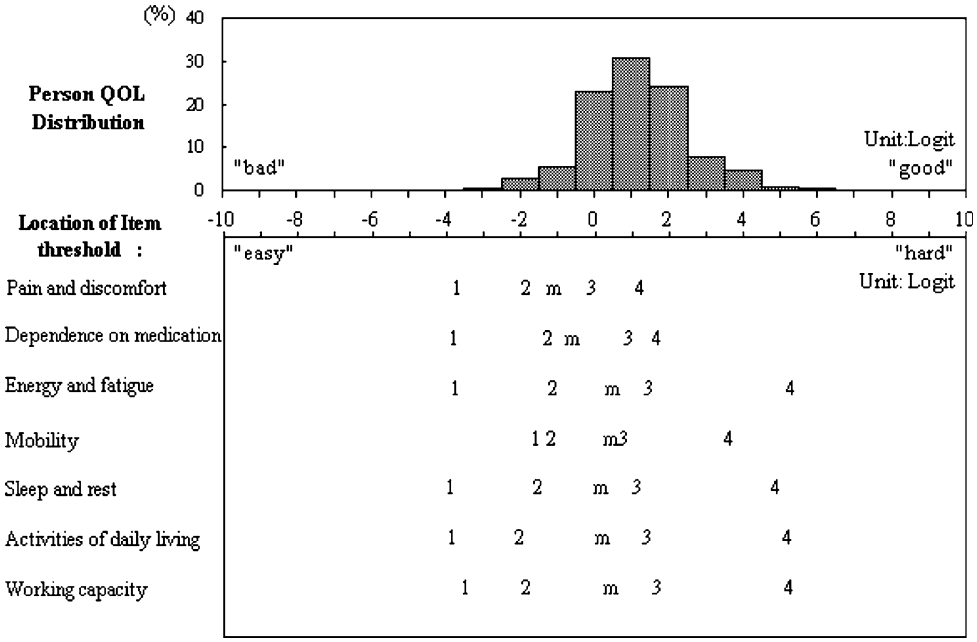

Fig. 1 Item–person map on the logit scale for the physical domain of the World Health Organization Quality of Life-BREF (WHOQOL-BREF). Logit values for the four thresholds (*1*, *2*, *3*, *4*, respectively) and the mean item difficulty (*m*) are given

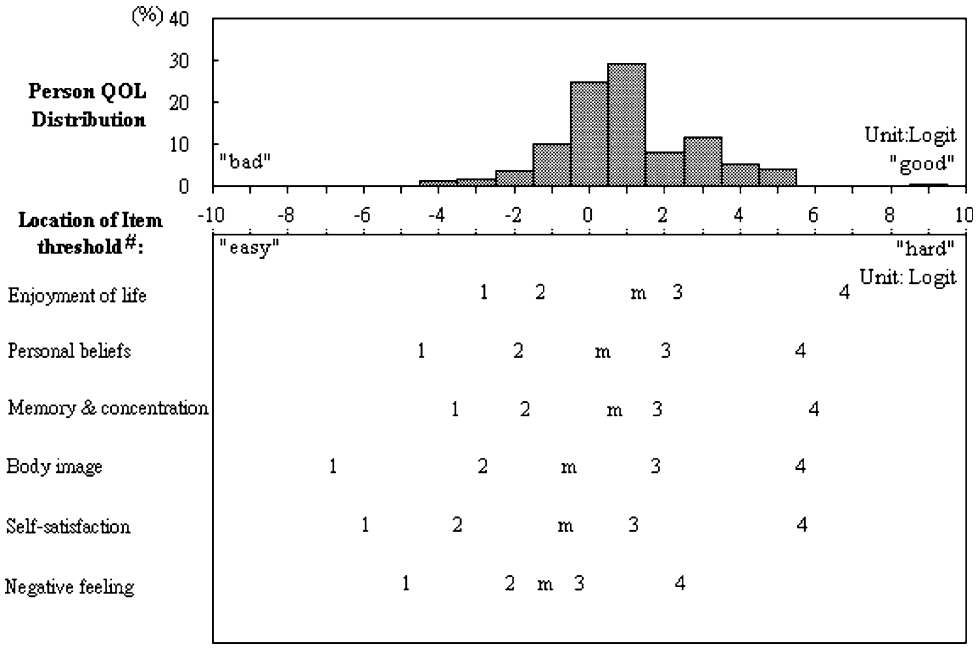

Fig. 2 Item–person map on the logit scale for the psychological domain of the WHOQOL-BREF. Logit values for the four thresholds (*1*, *2*, *3*, *4*, respectively) and the mean item difficulty (*m*) are given

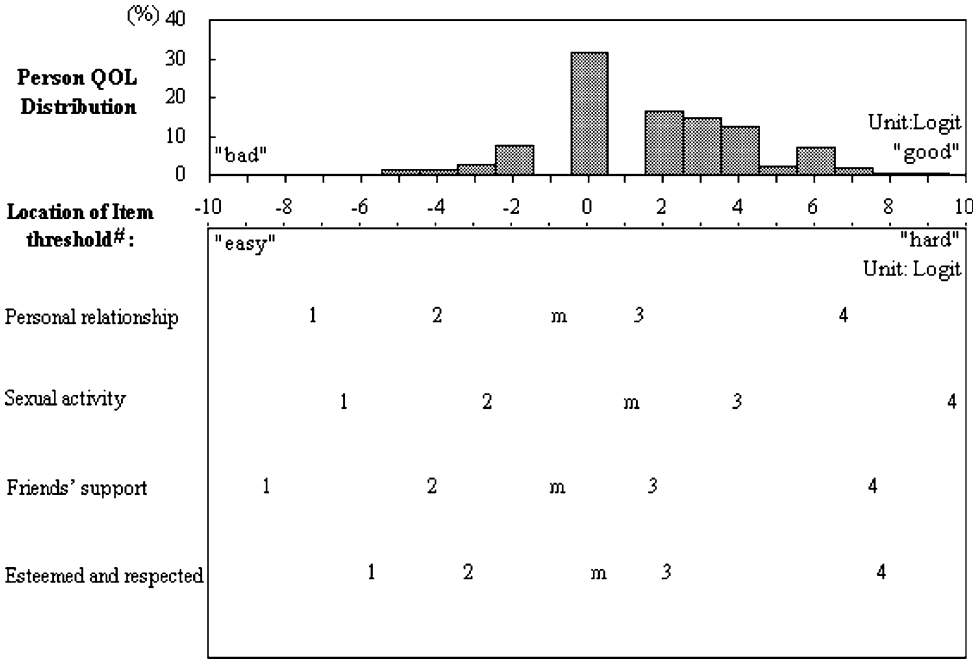

Fig. 3 Item–person map on the logit scale for the social domain of the WHOQOL-BREF. Logit values for the four thresholds (*1*, *2*, *3*, *4*, respectively) and the mean item difficulty (*m*) are given

Examination of DIF based on gender and age

Figure 5 shows that no DIF items were found across genders, with the exception of two items in the social domain. Item S21 (Satisfaction with sex life) located at the bottom right of the DIF plot indicated that it was the most difficult item for males but the least difficult one for females. Item S22 (Satisfaction with friends' support) was the least difficult item for males but the most difficult one for females. The DIF plot, shown in Fig. 6, indicated that three items (Y11, S21, and S22) displayed DIF across the two age groups. Item 'Y11 Bodily image and appearance' was more difficult for the young–old group but less difficult for the old–old group. We further examined items 'S21 Satisfaction with sex life' and 'S22 Satisfaction with friends' support' as they both exhibited DIF across genders and across age groups. In Fig. 7, there were three DIF items (i.e., 'S21 Satisfaction with sex life', 'S22 Satisfaction with friends' support', and 'S20 Personal relationships') in the young–old group but no significant DIF item in the old–old group, despite two borderline DIF items of 'S22 Satisfaction with friends' support' and 'S27 Esteem and respect'.

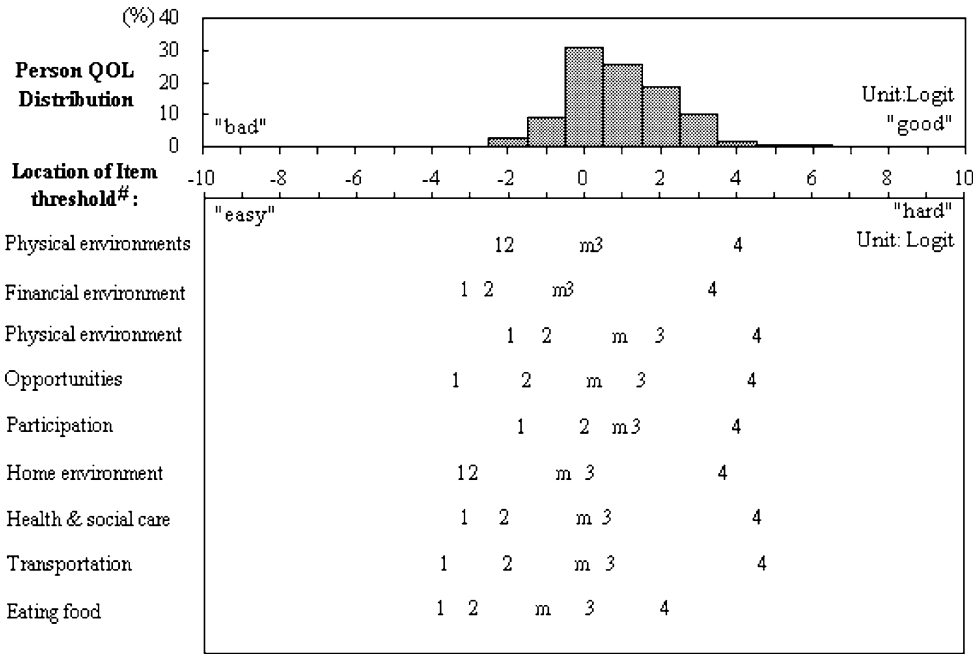

Fig. 4 Item–person map on the logit scale for the environmental domain of the WHOQOL-BREF. Logit values for the four thresholds (*1*, *2*, *3*, *4*, respectively) and the mean item difficulty (*m*) are given

Discussion

The aim of the first part of our analysis was to confirm the model fit. Each item of the WHOQOL-BREF in its respective domain had properly ordered threshold parameters and also showed a good fit to the unidimensionality specification. The person reliability indices were moderately high, with the exception of the social domain, which is similar to those found in Hwang et al. [11]. As IRT focuses more on item properties, such as difficulty, ordering, and number of items in each domain, this type of analysis provides results for a more comprehensive item evaluation.

Unlike the traditional CTT analysis, which often imposes continuous level assumptions on ordered response, the model in IRT analysis treats the response categories as ordinal levels of data [44]. The person reliability index is the proportion of observed variance that is explained by the Rasch model, while Cronbach's coefficient α is based on the covariance structure of the items within the scale. Both statistics are conceptually similar and are the percentage of observed response variance that is reproducible. Cronbach's coefficient α is constant for all scale scores, meaning that it assumes that measurement error is distributed normally and equally for all score levels; in contrast, the measure of precision for the person reliability index is estimated separately for each score level or response pattern, thereby controlling for the characteristics (e.g., difficulty) of the items in the scale. Therefore, one of the main differences between Cronbach's coefficient α and the person reliability index occurs when there are extreme scores. In such a case, Cronbach's coefficient α increases, but the person reliability decreases—and the latter is with a higher precision of measurement [16, 45, 46].

In addition, person reliability is computationally related to the person separation statistic, which is used to indicate the number of statistically different performance strata that can be identified in the sample (i.e., person separation of two indicates two strata can be identified) [16, 39]. The overall item performance also allowed us to evaluate the properties of each item in greater detail by examining their thresholds and corresponding person latent traits, which is not possible with the traditional CTT-based analysis [43, 44, 47]. For example, a person with a latent trait of three logits in the social domain is more likely to choose response category 4 (satisfied) for item 'S22 Personal relationships', 3 (moderately satisfied) for item 'S21 Sexual activity', 4 (satisfied) for item 'S22 friends' Support', and 4 (satisfied) for item 'S27 Esteemed and respected' (see Fig. 3).

The item–person map explicitly reveals the relationship between person latent trait estimates and item difficulty parameters. The positive mean values (targeting indices) of person traits for each of the four domains indicate that the average item difficulty was relatively low for the elderly, with the social domain in particular showing substantial mis-targeting, with a targeting index of 1.02. Those items tended to be scored as 'satisfied', perhaps because very elderly people may respond that they are satisfied, even though they are unsure how to answer questions about sexual life and friendship. The addition of items to the social domain that are more appropriate for the elderly would be desirable to enhance the content relevance and scale performance. It may also be worthwhile to include

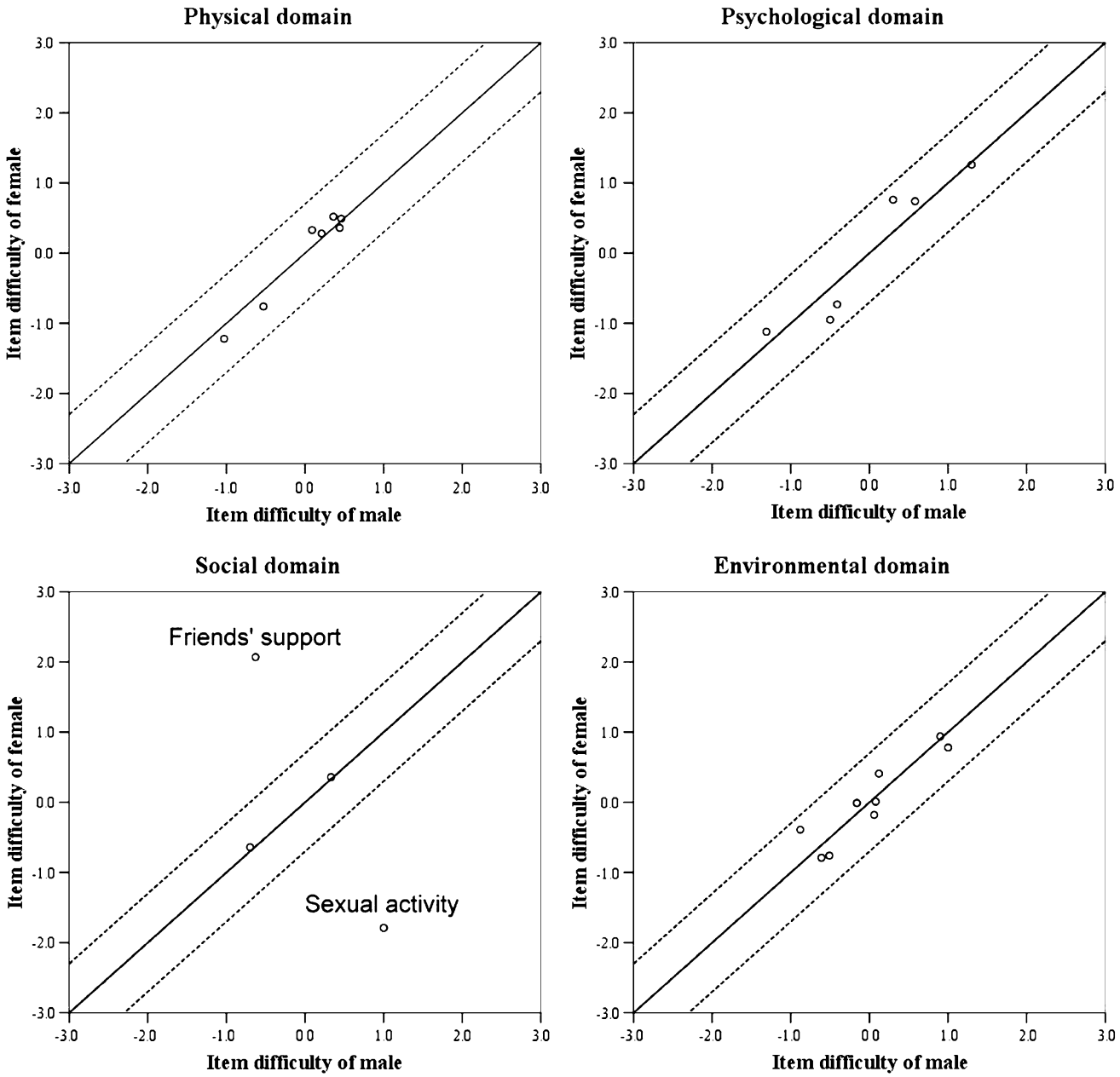

Fig. 5 Differential item functioning plot by gender and the WHOQOL-BREF domain

items in order to obtain an even distribution of thresholds on the scale of the social domain, with the aim of increasing the precision of estimates in the full range, as there were no threshold values in the intervals of −2–1 and 4–6 for the social domain [43, 47].

In the second part of our analysis, we applied DIF analysis to examine the invariance of item parameters across different subgroups and to explore influential issues within each specific construct domain. Item 'S21 Satisfaction with sexual activity' performed quite differently between males and females in young–old people, but there were only negligible differences in old–old people. This phenomenon may be attributed to the physiological differences of the aging process between males and females [48–51]. For example, females tend to have a distinct climacteric in the aging process, with apparent physiological and psychic changes as well as an end of the reproductive capacity and onset of menopause [52], while the climacteric in males is less clear, with a less-pronounced loss of libido and retained fertility [53]. However, sexual desire in both genders diminishes considerably as they age, as evidenced by the result that item 'S21 Satisfaction with sexual activity' was the least difficult item for both males and females in the old–old group.

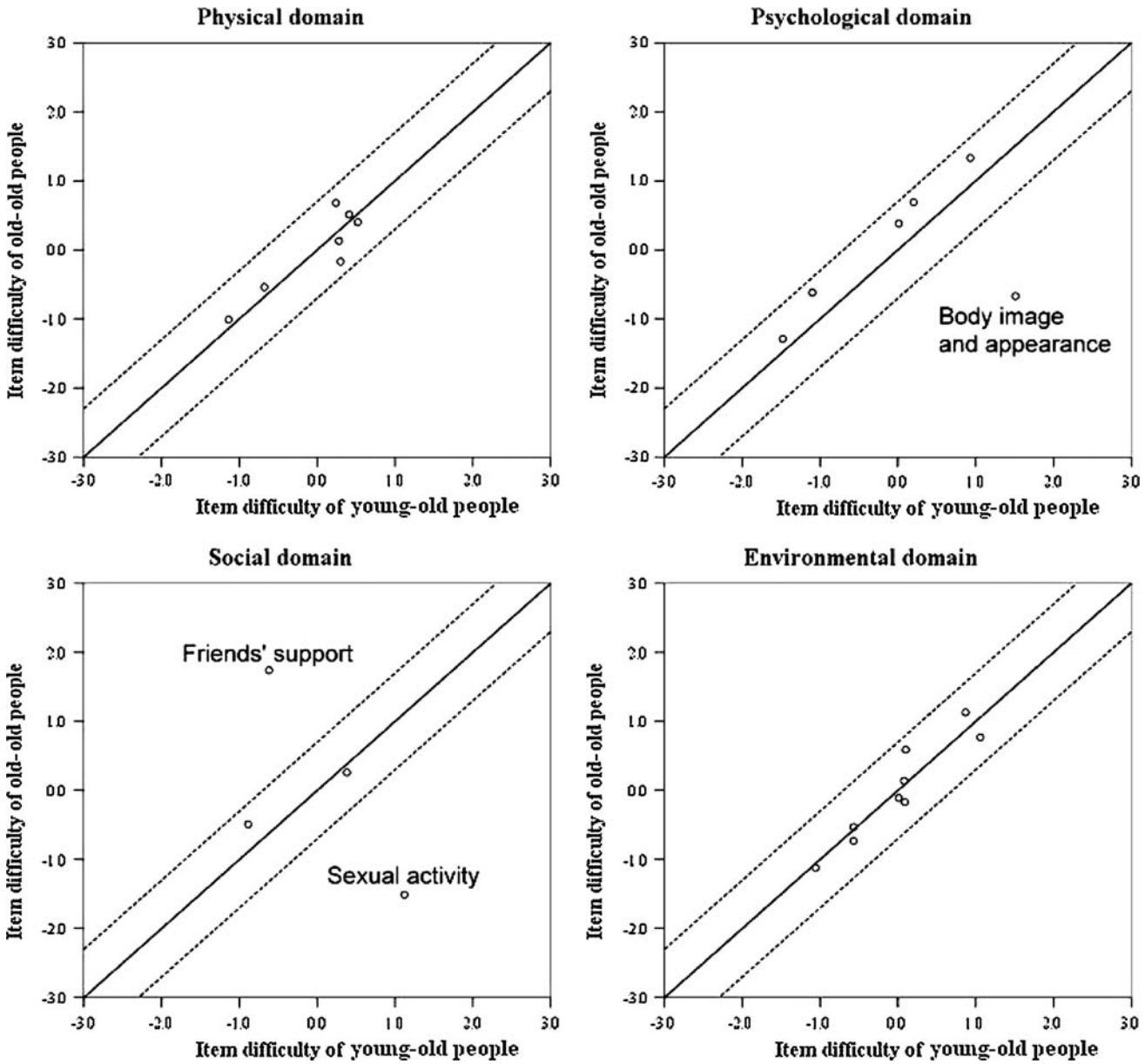

Fig. 6 Differential item functioning plot by age group and the WHOQOL-BREF domain

Another interesting item exhibiting DIF was item 'S22 Satisfaction with friends' support', which was a relatively easy item for males but the most difficult one for females, as shown in Fig. 5. Further DIF analysis across genders, stratified by different age groups as shown in Fig. 7, supported the concept that 'S22 Satisfaction with friends' support' was still a relatively difficult one for female. However, for males, 'S22 Satisfaction with friends' support' was a relatively easy item for young–old males but was the most difficult one for old–old males. The finding that females had a relatively low satisfaction with friends' support may have been due to the fact that our study subjects lived in a rural area and were more socially isolated and socially conservative. Another possible explanation is that the social circles of married women become smaller as they age and are often confined primarily to their husband and his relatives. Moreover, women often have a longer life expectancy than men and become widowed (e.g., in the study, the spouses of 16.8% of males and 39.5% of females had died). Previous studies have also found that elderly females residing in rural areas had a poor quality of social life, likely due to low levels of social contact support, high levels of isolation, being widowed, or living alone [54].

Another finding is that the item-difficulty locations of two items 'S21 Satisfaction with sex life' and 'S22 Satisfaction with friends' support' for males were calibrated in the two opposite extremes separately in the young–old and old–old groups. A noticeable movement of these two items

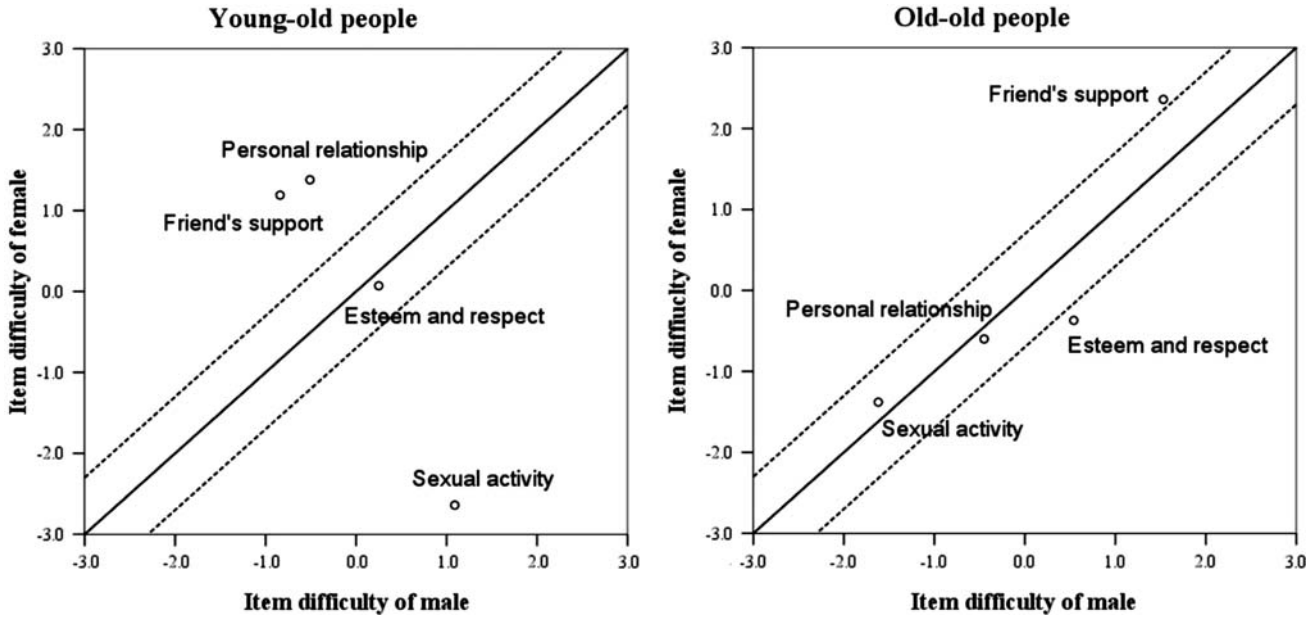

Fig. 7 Differential item functioning plot by age-stratified gender group in the social domain

in their relative locations was observed between the young–old and the old–old groups, as shown in Fig. 7. The violation of item parameter invariance for items S21 and S22 is potentially a consequence of getting older. Older people need to adjust to the aging process, such as the natural decrease in sexual activity that occurs with increased age and the greater social isolation as friends move away, become sick, or die. The relative importance of these two items in our study in the older male subjects changed, whereas the order of 'difficulty' for some other items was quite stable. For example, in both age groups, self-care tasks were easier to perform than household activities, and walking a short distance was easier to perform than stair climbing; however, items S21 and S22 did not maintain the expected order among latent traits with increasing age. The impact of DIF on older male subjects can be further investigated by examining its influence on domain scores of the WHOQOL-BREF, especially by comparing the magnitudes of scale-level differences in the social domain between the young–old and old–old groups.

There are some limitations in the study worth noting. First, the response rate of each of the 26 WHOQOL-BREF items was higher than 97.5%, with the exception of item '21 Sexual life' (83.5%). However, this limitation can be adequately mitigated by the Winstep program for the Rasch model as it can handle missing item response data. Second, the study cohort may not be representative of the general elderly population in Taiwan as compared with the general population in Taiwan, the study subjects had a lower education level and there was a higher percentage of males. The study subjects were also from a conservative rural area, and certain factors, such as the role of women in the family and job status before retirement, would be affected differently than were the subjects from urban areas. The reason why we chose this specific population group is because HRQL in rural-dwelling elderly people is less well understood and this group is in greater need of health promotion, including HRQL. Lastly, very sick older people were not available for this study; therefore, our findings may only reflect the QOL among older people who are relatively healthy.

Conclusions

Health-related quality of life has become an increasingly important measure in the clinical evaluation of the treatment and cost-effectiveness of health policies [12, 55]. As the WHOQOL-BREF expands its use as a generic HRQL instrument, it is of great importance to evaluate its psychometric properties and suitability in specific populations. Using Rasch analysis, we conducted a systematic evaluation of the WHOQOL-BREF in rural community-dwelling older people in Taiwan. While the results confirmed the suitability of the WHOQOL-BREF for this elderly population, they also identified areas for improvements, especially for items in the social domain. Differential item functioning analysis detected certain items, such as body image, sexual activity, and friends' support, as performing differently with increasing age. Researchers and clinicians need to be aware that the results for certain items, such as friends' support and sexual life, may depend on aging and

gender. If one ignores the age effect in the evaluation of a health promotion program, for example, the benefit of the program among the older elderly may be under-estimated due to their lower satisfaction with certain age- and gender-sensitive items, such as friends' support among older male individuals. Therefore, it is essential to conduct the analysis with and without these age- and gender-sensitive items to confirm consistency or with a stratified method by age and gender to avoid problems caused by DIF. We also suggest that both domain and item scores be used if there are DIF differences in order to obtain a more comprehensive understanding of HRQL in elderly populations. A particular scoring strategy is also suggested if we want to compare the HRQL by using this measure among old–old and old–young elderly Taiwanese in rural areas.

Acknowledgments This study was supported by the National Health Research Institutes (NHRI-EX98-9805PI and NHRI-EX95-9204PP), China Medical University (CMU96-225), and the National Science Council (NSC93-2320-B-039-013 and NSC95-2314-B-038-008) of Taiwan, Republic of China. We are most grateful to all of the subjects who participated in this study.

Appendix

See Table 3.

Table 3 The WHOQOL-BREF

Code	Item statement	Response choices
–	How would you rate your quality of life?	1. Very poor 2. Poor 3. Neither poor nor good 4. Good 5. Very good
–	How satisfied are you with your health?	1. Very dissatisfied 2. Dissatisfied 3. Neither satisfied nor dissatisfied 4. Satisfied 5. Very satisfied
P3	To what extent do you feel that physical pain prevents you from doing what you need to do?	1. Not at all 2. A little 3. A moderate amount 4. Very much 5. An extreme amount
P4	How much do you need any medical treatment to function in your daily life?	As above
Y5	How much do you enjoy life?	As above
Y6	To what extent do you feel your life to be meaningful?	As above
Y7	How well are you able to concentrate?	1. Not at all 2. A little 3. A moderate amount 4. Very much 5. Extremely
E8	How safe do you feel in your daily life?	As above
E9	How healthy is your physical environment?	As above
P10	Do you have enough energy for everyday life?	1. Not at all 2. A little 3. Moderately 4. Mostly 5. Completely
Y11	Are you able to accept your bodily appearance?	As above
E12	Have you enough money to meet your needs?	As above
E14	To what extent do you have the opportunity for leisure activities?	As above
P15	How well are you able to get around?	1. Very poor 2. Poor 3. Neither poor nor good 4. Good 5. Very good
P16	How satisfied are you with your sleep?	1. Very dissatisfied 2. Dissatisfied 3. Neither satisfied nor dissatisfied 4. Satisfied 5. Very satisfied
P17	How satisfied are you with your ability to perform your daily living activities?	As above
P18	How satisfied are you with your capacity for work?	As above
Y19	How satisfied are you with yourself?	As above
S20	How satisfied are you with your personal relationships?	As above
S21	How satisfied are you with your sex life?	As above
S22	How satisfied are you with the support you get from your friends?	As above
E23	How satisfied are you with the conditions of your living place?	As above
E24	How satisfied are you with your access to health services?	As above
E25	How satisfied are you with your transport?	As above
Y26	How often do you have negative feelings such as blue mood, despair, anxiety, depression?	1. Never 2. Seldem 3. Quite often 4. Very often 5. Always
S27	Do you feel respected by others?	1. Not at all 2. A little 3. A moderate amount 4. Very much 5. An extreme amount
E28	Are you usually able to get the things you like to eat?	1. Never 2. Seldem 3. Quite often 4. Very often 5. Always

References

1. Cella, D., & Chang, C. H. (2000). A discussion of item response theory and its applications in health status assessment. *Medical Care, 38*[Suppl 9], II66–72.
2. Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*[Suppl 9], II28–42.
3. Field trial WHOQOL-100. (1995). *The 100 questions with response scales*. Geneva: World Health Organization.
4. Field trial WHOQOL-100. (1995). *Scoring the WHOQOL*. Geneva: World Health Organization.
5. Ozakbas, S., Akdede, B. B., Kosehasanogullari, G., Aksan, O., & Idiman, E. (2007). Difference between generic and multiple sclerosis-specific quality of life instruments regarding the assessment of treatment efficacy. *Journal of the Neurological Sciences, 256*(1–2), 30–34.
6. Fang, C. T., Hsiung, P. C., Yu, C. F., Chen, M. Y., & Wang, J. D. (2002). Validation of the World Health Organization quality of life instrument in patients with HIV infection. *Quality of Life Research, 11*(8), 753–762.
7. Wang, W. C., Yao, G., Tsai, Y. J., Wang, J. D., & Hsieh, C. L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research, 15*(4), 607–620.
8. Naumann, V. J., & Byrne, G. J. A. (2004). WHOQOL-BREF as a measure of quality of life in older patients with depression. *International Psychogeriatrics, 16*(2), 159–173.
9. Hirayama, M. S., Gobbi, S., Gobbi, L. T., & Stella, F. (2008). Quality of life (QoL) in relation to disease severity in Brazilian Parkinson's patients as measured using the WHOQOL-BREF. *Archives of Gerontology and Geriatrics, 46*(2), 147–160.
10. Jamison, R. N., Fanciullo, G. J., Mchugo, G. J., & Baird, J. C. (2007). Validation of the short-form interactive computerized quality of life scale (ICQOL-SF). *Pain Medicine, 8*(3), 243–250.
11. Hwang, H. F., Liang, W. M., Chiu, Y. N., & Lin, M. R. (2003). Suitability of the WHOQOL-BREF for community-dwelling older people in Taiwan. *Age and Ageing, 32*(6), 593–600.
12. Pickard, A. S., Dalal, M. R., & Bushnell, D. M. (2006). A comparison of depressive symptoms in stroke and primary care: Applying Rasch models to evaluate the center for epidemiologic studies-depression scale. *Value in Health, 9*(1), 59–64.
13. Conrad, K. J., & Smith, E. V., Jr. (2004). International conference on objective measurement: Applications of Rasch analysis in health care. *Medical Care, 42*[Suppl 1], I1–6.
14. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Lawrence Erlbaum Associates.
15. Hambelton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage.
16. Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
17. Noyes, K., Dick, A. W., & Holloway, R. G. (2006). Pramipexole versus levodopa in patients with early Parkinson's disease: Effect on generic and disease-specific quality of life. *Value in Health, 9*(1), 28–38.
18. Gunther, O. H., Roick, C., Angermeyer, M. C., & Konig, H. H. (2008). The responsiveness of EQ-5D utility scores in patients with depression: A comparison with instruments measuring quality of life, psychopathology and social functioning. *Journal of Affective Disorders, 105*(1–3), 81–91.
19. Lawton, M. P., Windley, P. G., & Byerts, T. O. (1982). *Aging and the environment: Theoretical approaches*. New York: Springer.
20. Pearlman, R. A., & Uhlmann, R. F. (1988). Quality of life in chronic diseases: Perceptions of elderly patients. *Journal of Gerontology, 43*(2), M25–30.
21. Stewart, A. L., Sherbourne, C. D., & Brod, M. (1996). *Measuring health-related quality of older and demented populations*. Philadelphia: Lippincott-Raven Publishers.
22. Bindman, A. B., Keane, D., & Lurie, N. (1990). Measuring health changes among severely ill patients: The floor phenomenon. *Medical Care, 28*(12), 1142–1151.
23. Watsford, M. L., Murphy, A. J., & Pine, M. J. (2007). The effects of ageing on respiratory muscle function and performance in older adults. *Journal of Science and Medicine in Sport, 10*(1), 36–44.
24. WHOQOL-BREF. (1996). *Introduction, adminstration, scoring and generic version of the assessment–field trial version*. Geneva: World Health Organization.
25. Yao, K. P. (2002). Development and applications of the WHO-QOL-Taiwan version. *Formosan Journal of Medicine, 6*(2), 193–200.
26. Chan, S. W., Chiu, H. F., Chien, W. T., Thompson, D. R., & Lam, L. (2006). Quality of life in Chinese elderly people with depression. *International Journal of Geriatric Psychiatry, 21*(4), 312–318.
27. Yang, S. C., Kuo, P. W., Wang, J. D., Lin, M. I., & Su, S. (2006). Development and psychometric properties of the dialysis module of the WHOQOL-BREF Taiwan version. *Journal of the Formosan Medical Association, 105*(4), 299–309.
28. Yao, G., Chung, C. W., Yu, C. F., & Wang, J. D. (2002). Development and verification of validity and reliability of the WHOQOL-BREF Taiwan version. *Journal of the Formosan Medical Association, 101*(5), 342–351.
29. Van Der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.
30. Linacre, J. M. (2006). *A user's guide to WINSTEPS: Rasch-Model Computer Programs.* Chicago: www.winsteps.com.
31. Wright, B. D., & Maters, O. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.
32. Avery, L. M., Russell, D. J., Raina, P. S., Walter, S. D., & Rosenbaum, P. L. (2003). Rasch analysis of the gross motor function measure: Validating the assumptions of the Rasch model to create an interval-level measure. *Archives of Physical Medicine and Rehabilitation, 84*(5), 697–705.
33. Jöreskog, K., & Sörbom, D. (2005). *LISREL 8.72*. Chicago: Scientific Software.
34. Metz, S. M., Wyrwich, K. W., Babu, A. N., Kroenke, K., Tierney, W. M., & Wolinsky, F. D. (2006). A comparison of traditional and Rasch cut points for assessing clinically important change in health-related quality of life among patients with asthma. *Quality of Life Research, 15*(10), 1639–1649.
35. Hart, D. L., Mioduski, J. E., & Stratford, P. W. (2005). Simulated computerized adaptive tests for measuring functional status were efficient with good discriminant validity in patients with hip, knee, or foot/ankle impairments. *Journal of Clinical Epidemiology, 58*(6), 629–638.
36. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis_conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.
37. Fitzpatrick, R., Norquist, J. M., Dawson, J., & Jenkinson, C. (2003). Rasch scoring of outcomes of total hip replacement. *Journal of Clinical Epidemiology, 56*(1), 68–74.
38. Wright, B. D., & Linacre, J. M. (1996). Reasonable mean-square fit values. In J. M. Linacre (Ed.), *Rasch measurement transactions* (p. 370). Chicago: MESA.
39. Duncan, P. W., Bode, R. K., Min Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke

impact scale. *Archives of Physical Medicine and Rehabilitation, 84*(7), 950–963.
40. Gallagher, P., Horgan, O., Franchignoni, F., Giordano, A., & Maclachlan, M. (2007). Body image in people with lower-limb amputation: A Rasch analysis of the amputee body image scale. *American Journal of Physical Medicine and Rehabilitation, 86*(3), 205–215.
41. Urbach, D. R., Tomlinson, G. A., Harnish, J. L., Martino, R., & Diamant, N. E. (2005). A measure of disease-specific health-related quality of life for achalasia. *The American Journal of Gastroenterology, 100*(8), 1668–1676.
42. Wolfe, F., Michaud, K., & Pincus, T. (2004). Development and validation of the health assessment questionnaire II: A revised version of the health assessment questionnaire. *Arthritis and Rheumatism, 50*(10), 3296–3305.
43. Lai, J. S., Cella, D., Chang, C. H., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Quality of Life Research, 12*(5), 485–501.
44. Uttaro, T., & Lehman, A. (1999). Graded response modeling of the quality of life interview. *Evaluation and Program Planning, 22*(1), 41–52.
45. Vidotto, G., Bertolotti, G., Carone, M., Arpinelli, F., Bellia, V., Jones, P. W., et al. (2006). A new questionnaire specifically designed for patients affected by chronic obstructive pulmonary disease; the Italian health status questionnaire. *Respiratory Medicine, 100*(5), 862–870.
46. Vidotto, G., Carone, M., Jones, P. W., Salini, S., & Bertolotti, G. (2007). Maugeri respiratory failure questionnaire reduced form: A method for improving the questionnaire using the Rasch model. *Disability and Rehabilitation, 29*(13), 991–998.
47. Garratt, A. M. (2003). Rasch analysis of the Roland disability questionnaire. *Spine, 28*(1), 79–84.
48. Onder, G., Penninx, B. W., Guralnik, J. M., Jones, H., Fried, L. P., Pahor, M., et al. (2003). Sexual satisfaction and risk of disability in older women. *The Journal of Clinical Psychiatry, 64*(10), 1177–1182.
49. Wespes, E., & Schulman, C. C. (2002). Male andropause: Myth, reality, and treatment. *International Journal of Impotence Research, 14*[Suppl 1], S93–98.
50. Kaiser, F. E. (1992). Sexual function and the older cancer patient. *Oncology (Williston Park), 6*(2 Suppl), 112–118.
51. Yang, H., Toy, E. C., & Baker, B. (2000). Sexual dysfunction in the elderly patient. *Primary Care Update for Ob/Gyns, 7*(6), 269–274.
52. Kingsberg, S. A. (1998). Postmenopausal sexual functioning: A case study. *International Journal of Fertility and Women's Medicine, 43*(2), 122–128.
53. Peate, I. (2003). The male menopause: Possible causes, symptoms and treatment. *The British Journal of Nursing, 12*(2), 80–84.
54. Savikko, N., Routasalo, P., Tilvis, R. S., Strandberg, T. E., & Pitkala, K. H. (2005). Predictors and subjective causes of loneliness in an aged population. *Archives of Gerontology and Geriatrics, 41*(3), 223–233.
55. Tengs, T. O. (2004). Cost-effectiveness versus cost-utility analysis of interventions for cancer: Does adjusting for health-related quality of life really matter? *Value in Health, 7*(1), 70–78.