Quality of Life Research (2006) 15: 607–620
DOI 10.1007/s11136-005-4365-7

Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis

Wen-Chung Wang[1], Grace Yao[2], Yih-Jian Tsai[3], Jung-Der Wang[4] & Ching-Lin Hsieh[5]
[1]*Department of Psychology, National Chung Cheng University, Chia-Yi, 621, Taiwan (E-mail: psywcw@ccu.edu.tw);* [2]*Department of Psychology, National Taiwan University, Taiwan;* [3]*Population and Health Research Center, Bureau of Health Promotion, Department of Health;* [4]*Institute of Occupational Medicine and Industrial Hygiene, College of Public Health, National Taiwan University, Taiwan;* [5]*School of Occupational Therapy, College of Medicine, National Taiwan University, Taiwan*

Accepted in revised form 20 October 2005

Abstract

Objective: This study examined the construct validity, and improved the test reliability and the estimation accuracy for the correlation between domains of the WHOQOL-BREF using multidimensional Rasch analysis. *Method*: A total of 13,083 adults were administered the 28-item WHOQOL-BREF Taiwan version, which consists of 4 subscales (domains). The multidimensional form of the partial credit model was used to examine the fit of the 4 subscales. For comparison, each subscale individually was also fitted to the unidimensional partial credit model. Standard item fit statistics and analysis of differential item functioning (DIF) were used to check model-data fit. *Results*: After excluding 2 overall items and deleting 7 DIF items, the remaining items of each subscale in the WHOQOL-BREF constituted a single construct. The test reliabilities and correlations between domains obtained from the multidimensional approach, (0.82–0.86) and (0.79–0.89), respectively, were much higher than those obtained from the unidimensional approach, (0.67–0.75) and (0.53–0.65), respectively. *Conclusion*: The 19-item WHOQOL-BREF measures more succinct latent traits than the original design. The multidimensional approach yields not only more accurate estimates for the correlation between domains but also substantially higher reliabilities, than the standard unidimensional approach.

Key words: Bandwidth-fidelity dilemma, Multidimensional item response model, Quality of life, Rasch measurement, Test reliability

Introduction

The World Health Organization (WHO) has developed an instrument for assessing quality of life (QOL) [1], which is called the WHOQOL-100. It consists of 100 items representing 24 aspects in six domains. The WHOQOL-BREF, the abbreviated version of the WHOQOL-100, contains one item from each of the 24 aspects of the WHOQOL-100, plus two benchmark items from the general aspect on overall QOL and general health. The WHOQOL-100 and WHOQOL-BREF have been developed for use in Taiwan according to the WHO international guidelines [2]. The 28-item WHOQOL-BREF Taiwan version consists of 2 overall items measuring general QOL and health and 24 items that are universally adopted for the WHOQOL-BREF in four domains (subscales): physical health, psychological health, social relationships, and environment, plus 2 additional items that are more specific to the culture of people in Taiwan: being respected/accepted among people (social relationships domain), and eating what one loves to eat (environmental domain). Respondents are asked to evaluate the selected attributes of QOL over the previous two weeks on 5-point rating scales. High scores indicate good QOL.

Emphasis on the measurement properties of QOL questionnaires has increased in recent years. The WHOQOL-BREF has been validated using classical test theory [3–6]. However, the measurement properties of the WHOQOL-BREF have rarely been explored using modern test theory (e.g., Rasch analysis). In a recent study of the WHOQOL-BREF [7], the item responses to the 5-point rating scales were dichotomiedso that the dichotomous Rasch model could be fitted. Dichotomizing responses will in general cause a loss of information and lead to less accurate measures. In fact, the partial credit model and the rating scale model [8], preserving the ordered nature of polytomous responses, are more appropriate for polytomous items such as rating scales, Likert-type items, or essay questions.

Traditional analysis inappropriately treats raw scores or their linear transformations and item responses to rating scales as interval data. Rasch analysis is a statistical technique that can be applied to dichotomous or polytomous items to transform ordinal scores into interval measures [9, 10]. If data do not fit the Rasch model's expectation, unidimensionality is not preserved, so the presumed latent trait is not quantified successfully. Standard Rasch analysis is based on unidimensional Rasch models where a single latent trait is assumed to determine individuals' performances on the test. If a test consists of several unidimensional tests (e.g., the four subscales of the WHOQOL-BREF), it could be calibrated using standard Rasch analysis procedures. The test could be either analyzed as a whole, or the unidimensional Rasch model could be applied to each subscale separately, one test at a time. The first approach ignores the claims for the subscale structure of the test, the second approach, shown in Figure 1a ignores the potential intercorrelations between related, but not identical latent traits. This approach is likely to yield unnecessarily imprecise measurements, especially when the tests are short [11].

To take the correlations between latent traits into account, one needs a multidimensional model that simultaneously calibrates all the tests and thus utilizes the correlations to increase measurement precision. In reality, there are always non-zero correlations between latent traits, meaning that at least in theory the multidimensional approach (Figure 1b) is more appropriate than the unidimensional one (Figure 1a). In addition, the greater the correlations, the greater the measurement precision using the multidimensional approach [11]. In other words, even short tests can yield precise measurements if the multidimensional approach is used, given that the latent traits are not uncorrelated. If the tests are long enough, the measures of each individual latent trait obtained from the unidimensional approach will be accurate enough, and thus the multidimensional approach will yield little improvement in measurement precision. On the other hand, if the tests are too short for the unidimensional approach to yield precise measures, the use of the multidimensional approach will squeeze as much information as possible from the whole data for multiple dimensions to provide measures that are more precise.

In many cases, the correlations between latent traits are also of great interest (e.g., the correlations among the four subscales of the WHOQOL-BREF). Using the unidimensional approach, one

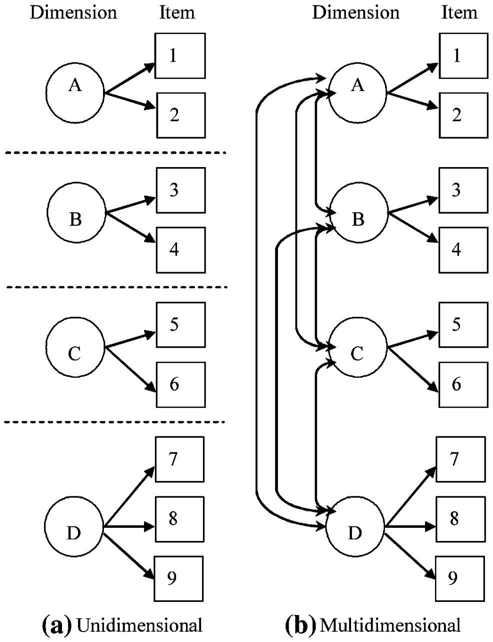

Figure 1. Graphical representations of the unidimensional and multidimensional approaches. *Note*: The arrow between dimension and item indicates that the dimension (e.g., A) is assessed by the item (e.g., Item 1). The arrow between dimensions in (b) indicates that the correlations between the two dimensions can be directly estimated, whereas the dashed line in (a) indicates that the correlation can only be indirectly estimated.

can compute the Pearson correlation on the person measures for two latent traits. Because the person measures contain certain amounts of measurement error, the computed correlation underestimates the true correlation between latent traits. Direct estimation of the correlation between latent traits is possible only for the multidimensional approach, rather than for the unidimensional one. Thus the multidimensional approach is useful in estimating correlations between latent traits.

Multidimensional Rasch analysis can be useful for the WHOQOL-BREF, given that the lengths of the four subscales are short and that the correlations among them are generally high [3, 5]. In this study, we aimed to validate the dimensionality of the WHOQOL-BREF, directly estimate the correlations between domains, and increase the test reliabilities of the four domains, using multidimensional Rasch analysis.

Methods

Subjects

The National Health Interview Survey (NHIS) in Taiwan used a multi-stage stratified systematic sampling scheme in 2001. The 359 townships/ districts of Taiwan were divided into 7 strata according to geographic location and an urbanization index. Townships or districts in each stratum were selected with selection probability proportional to their household population sizes (PPS) registered on 16th January, 2001. In each selected township/district, lins (the smallest administrative unit in Taiwan) were selected with PPS. Four households were selected randomly from each selected lin. Every member of the selected household was interviewed [12].

Instrument

The first two items of the 28-item WHOQOL-BREF [2, 13] assess general QOL and health. The remaining 26 items consist of 24 items universally adopted for the WHOQOL-BREF in four subscales: physical health (7 items), psychological health (6 items), social relationships (3 items), and environment (8 items), plus two additional items that are more specific to the culture of people in Taiwan: “Do you feel respected by others?” (social subscale), and “Are you usually able to get the things you like to eat?” (environmental subscale) [2]. The two general items were removed from the analyses in this study, because they do not belong to any of the four subscales.

Procedures

The respondents, aged 20–65, were asked to evaluate the selected attributes of QOL in the previous two weeks. To ensure consistency in the interviews, all interviewers received pre-job training. Basically, the interviewees filled in the WHOQOL-BREF themselves, according to the original design. A total of 13 senior Bureau of Health Promotion staff closely supervised the interview process, reviewed all completed questionnaires, and randomly verified interviewees’ responses in a telephone follow-up. They made a cross-item comparison as well between the scale and other corresponding items in the NHIS. The WHOQOL-BREF was not returned to an interviewer for a revisit even if there were any missing items, errors, or contradictions. No proxies were allowed, even if an interviewee was frail, mentally ill, or unable to communicate.

Data analysis

The multidimensional Rasch model used in this study is called the multidimensional random coefficients multinomial logit model (MRCMLM) [14]. Let person n's levels on the L latent traits (in this study $L = 4$) be denoted as $\boldsymbol{\theta}_n^{\mathrm{T}} = (\boldsymbol{\theta}_{n1}, \ldots, \boldsymbol{\theta}_{nL})$, which is considered to represent a random sample from a population with a multivariate density function $g(\boldsymbol{\theta}_n; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ indicates a vector of parameters that characterize the distribution. In this study, g is assumed to be normal so that $\boldsymbol{\alpha} \equiv (\boldsymbol{\mu}, \Sigma)$. The probability of a response in category j of item i for person n is

$$p_{nij} = \frac{\exp\left(\mathbf{b}_{ij}^{\mathrm{T}}\boldsymbol{\theta}_n + \mathbf{a}_{ij}^{\mathrm{T}}\boldsymbol{\xi}\right)}{\sum_{u=1}^{K_i} \exp\left(\mathbf{b}_{iu}^{\mathrm{T}}\boldsymbol{\theta}_n + \mathbf{a}_{iu}^{\mathrm{T}}\boldsymbol{\xi}\right)}, \quad (1)$$

where K_i is the number of categories in item i (in this study, $K_i = 5$ for every item); $\boldsymbol{\xi}$ is a vector of location parameters that describe the items; $\mathbf{b}_{ij}$ is a score vector given to category j of item i across the

L latent traits, which can be collected across items into a scoring matrix $\mathbf{B}$ for the whole test; and $\mathbf{a}_{ij}$ is a design vector given to category j of item i that describes the linear relationship among the elements of ξ, which can be collected across items into a design matrix $\mathbf{A}$ for the whole test. Equation 1 can be expressed as

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \left(\mathbf{b}_{ij}^{\mathrm{T}} - \mathbf{b}_{i(j-1)}^{\mathrm{T}}\right)\boldsymbol{\theta}_n + \left(\mathbf{a}_{ij}^{\mathrm{T}} - \mathbf{a}_{i(j-1)}^{\mathrm{T}}\right)\xi \equiv \mathbf{b}_{ij}^{*\mathrm{T}}\boldsymbol{\theta}_n + \mathbf{a}_{ij}^{*\mathrm{T}}\xi, \tag{2}$$

which is more consistent with the standard expression of the family of Rasch models.

Using $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ (or equivalently $\mathbf{a}_{ij}^*$ and $\mathbf{b}_{ij}^*$) to define the relationship between items and persons allows a general model to be written that includes most of the existing unidimensional Rasch models, such as the simple logistic model [9], the linear logistic test model [15], the rating scale model [8], the partial credit model [16], the partial order model [17], the facet model [18], and the linear partial credit model [19]. More importantly, the definitions allow the specification of a range of multidimensional models by imposing linear constraints on the item parameters, such as multidimensional forms of the simple logistic model, the rating scale model, the partial credit model (to be used in this study), and the linear partial credit model. For details on how to manipulate $\mathbf{a}_{ij}$ and $\mathbf{b}_{ij}$ to form the above models and other customized models, the reader is referred to Adams et al. [14] and Adams and Wilson [20].

In this study, each of the four subscales of the WHOQOL-BREF is treated as unidimensional so that, as a whole, the WHOQOL-BREF is four-dimensional. Because the items in the WHOQOL-BREF are not judged on the same kind of rating scales, the multidimensional form of the partial credit model, rather than the rating scale model, is fitted to the data. Under the (unidimensional) partial credit model, the log-odds of being in category j over category $j - 1$ in item i for a person n with latent trait θ_n are

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - (\delta_i + \tau_{ij}), \tag{3}$$

where δ_i is called the (overall) difficulty of item i, and τ_{ij} is called the categorical boundary (or step) parameter of category j relative to category $j - 1$ for item i.

For the rating scale model, Equation 3 reduces to:

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - (\delta_i + \tau_j), \tag{4}$$

where all items are constrained to share the same set of the categorical boundary parameters τ_j. For the simple logistic Rasch model, Equation 3 reduces to:

$$\log\left(\frac{p_{nij}}{p_{ni0}}\right) = \theta_n - \delta_i, \tag{5}$$

where each item has one difficulty parameter δ_i. Comparing Equations 3, 4 and 5 with Equation 2, one recognizes that the partial credit model, the rating scale model, and the simple logistic Rasch model are all special cases of the MRCMLM.

The MRCMLM, being a member of the exponential family of distributions, can be viewed as a generalized linear mixed model [21–25]. Several computer programs can be used to calibrate parameters in the MRCMLM, including ConQuest [26], SAS NLMIXED [27, 28], STATA gllamm [29, 30], MIXOR [31] and MIXNO [32]. In the authors' experiences with the multidimensional approach, ConQuest takes only a few minutes to converge, whereas the other programs may take several hours to converge (or sometimes even fail to converge). Hence, ConQuest was used for all analyses in this study. The appendix shows in detail how the parameters in the MRCMLM are estimated using ConQuest. Because all the model parameters in the MRCMLM $\xi, \boldsymbol{\mu}$ and Σ) are simultaneously estimated, measurement errors in the latent traits $\boldsymbol{\theta}$ are directly taken into account.

To check if items fit the model's expectation, two kinds of analyses were performed. One was item fit statistical analysis, and the other was analysis of differential item functioning (DIF) [10, 33–35]. Regarding item fit statistics, the outfit mean square error (MNSQ) in which residuals are directly summated, is sensitive to unexpected behavior by persons on items far from the person's proficiency level; the infit MNSQ in which residuals are weighted before summation, is sensitive to unexpected behavior affecting responses to items near the person's proficiency measure. When the data fit the model's expectation, the infit and outfit MNSQ statistics have an expected value of unity. The magnitudes of the MNSQ statistics show the amount of distortion of the measurement system. Values less than unity (also called over-fit) indicate that observations are too predictable (redundancy). Values greater than unity (called under-fit) indicate unpredictability (unmodeled noise). Statistically, the MNSQ statistics are chi-square statistics divided by their degrees of freedom. For rating scales, a range of (0.6, 1.4) is recommended as the critical range for the MNSQ statistics [36]. Items with infit or outfit MNSQ statistics beyond this range are usually regarded as misfitting. A more stringent range of (0.7, 1.3) was used in this study.

It is a great challenge to develop a test that is suitable for many groups. Self-report inventories, in particular, face this difficulty, because different groups may have different linguistic interpretations of test items and category labels. DIF analysis is a means of verifying construct equivalence over groups [34]. If construct equivalence does not hold over groups, meaning that different groups have different perspectives on the items, the derived measures are not directly comparable over groups. If, for example, different genders or age groups have different perspectives on the four subscales of the WHOQOL-BREF, normative data of the WHOQOL-BREF for the general population is not possible. To obtain comparable measures over groups, all the items have to be DIF-free or at least DIF-trivial.

Statistically, an item is considered to exhibit DIF if the response probabilities for that item cannot be fully explained by the latent trait and a set of difficulty parameters for that item. DIF analysis identifies items that appear to be too difficult or too easy, after having controlled for differences in the latent trait levels of the reference and focal groups. There were 3 main demographic characteristics of our participants, including gender (2 groups), education (classified as 3 groups: elementary, secondary, and higher education), and age (classified as 5 groups: 20–29, 30–39, 40–49, 50–59, and over 60). Individually, we compared differences in the overall item difficulties between men and women, between five age groups, and between three education levels. Once a difference was found between men and women, between any two of the five age groups, or between any two of the three education levels, the item was considered as exhibiting DIF.

To resolve the scale indeterminacy problem in DIF analysis, the mean item parameters were set to be equal (zero) over groups so that the differences in the parameter estimates between groups can be directly compared. By setting the mean item parameters identical across groups, the "impact" of the difference in latent trait levels on DIF analysis is eliminated [10, 33]. This procedure has been implemented on several popular computer programs, such as BILOG-MG [37], WINSTEPS [38], and ConQuest. There are other procedures to solve scale indeterminacy [39], for example, if a set of items are believed to have no DIF, they can served as the anchors so that the other items can be detected for the evidence of DIF. Unfortunately, this procedure was not applicable in this study, because no prior knowledge was available to claim which items were indeed DIF-free.

Because the sample sizes of the groups were very large (several thousands), a trivial DIF could be identified as statistically significant. In this study, a difference larger than or equal to 0.5 logits in the estimates between any groups was treated as a sign of substantial DIF. A difference of 0.5 logits is equal to an odds ratio of 1.65 ($=2.718^{0.5}$). Once a DIF item was identified, it was removed from further analysis. The multidimensional form of the partial credit model was again fitted to the new data set. The analyses stopped when all the infit and outfit MNSQ statistics were located within the (0.7, 1.3) critical range and no DIF items were identified.

Results

Participants

There were 27,160 eligible participants living in 7357 households who were found as a result of sampling. They were representative of the national population in age, gender, and urbanization [40]. From late August 2001 to January 2002, the 2001 NHIS data were collected from 25,464 persons who were aged from 11 to 98 and lived in 6721 households, with response rates of 93.8% by person and 91.4% by household, respectively [41]. Among the 15,425 participants aged 20–65, 13,083 (85%) participants completed the WHOQOL-BREF and their data were analyzed in this study. These subjects had a higher proportion among those aged 50–59 and a lower proportion among those aged 20–39, as compared with the registered Taiwan population in the 2000 census [12].

Model-data fit

DIF analysis was conducted to assess the model-data fit. Table 1 lists the maximum differences in the estimates of item difficulties across groups. We took a difference lager than or equal to 0.5 logits as a sign of substantial DIF. None of the six items of the psychological subscale exhibited substantial DIF. For the physical subscale, only item 15 (How well are you able to get around?) exhibited substantial DIF between young and old adults (older than 50). For the social subscale, item 21 (How satisfied are you with your sex life?) showed substantial DIF between young (below 40) and old adults. Finally, for the environmental subscale, items 9 (How healthy is your physical environment?), 12 (Have you enough money to meet your needs?), 13 (How available to you is the information that you need in your day-to-day life?), and 14 (To what extent do you have the opportunity for leisure activities?) had substantial DIF, mainly between low and high educational groups, and

Table 1. Maximum differences in the estimates for item difficulties (in absolute value) over gender, age, and education level, and infit and outfit MNSQ statistics

Subscale/item	Gender	Age	Education	Outfit	Infit
Psychological					
5 Positive feelings	0.20	0.40	0.29	1.10	1.10
6 Spirituality/religion/personal beliefs	0.09	0.08	0.02	0.92	0.92
7 Thinking, learning, memory, and concentration	0.19	0.25	0.07	0.94	0.94
11 Bodily image and appearance	0.02	0.12	0.18	1.00	1.01
19 Self-esteem	0.02	0.09	0.09	0.86	0.87
26 Negative feelings	0.14	0.31	0.20	1.17	1.15
Physical					
3 Pain and discomfort	0.08	0.11	0.29	1.27	1.19
4 Dependence on medical substances and medical aids	0.17	0.38	0.10	1.30	1.18
10 Energy and fatigue	0.20	0.05	0.18	0.92	0.92
15 Mobility	0.08	0.53*	0.22		
16 Sleep and rest	0.14	0.32	0.36	1.02	1.02
17 Activities of daily living	0.03	0.24	0.24	0.81	0.83
18 Work capacity	0.02	0.35	0.35	0.95	0.96
Social					
20 Personal relationships	0.07	0.21	0.13	0.95	0.96
21 Sexual activity	0.03	0.51*	0.24		
22 Practical social support	0.13	0.41	0.03	0.95	0.97
27 Being respected/accepted	0.03	0.49	0.15	1.08	1.09
Environmental					
8 Freedom, physical safety and security	0.23	0.17	0.46	1.12	1.11
9 Physical environment: pollution/noise/traffic/climate)	0.04	0.16	0.83*		
12 Financial resources	0.09	0.09	0.75*		
13 Opportunities for acquiring new information and skills	0.05	0.19	0.89*		
14 Participation in and opportunities for recreation/leisure activities	0.06	0.35	0.52*		
23 Home environment	0.08	0.22	0.45	1.00	1.01
24 Health and social care: accessibility and quality	0.04	0.06	0.44	0.98	0.98
25 Transport	0.01	0.19	0.32	0.94	0.95
28 Eating/food	0.23	0.50*	0.40		

Note: *Substantial DIF (a difference in item difficulties larger than or equal to 0.5 logits between groups); Gender: 1 = Men, 2 = Women; Age: 1 = 20-29, 2 = 30-39, 3 = 40-49, 4 = 50-59, 5 = Over 60; Education: 1 = Elementary, 2 = Secondary, 3 = Higher education.

item 28 (Are you usually able to get the things you like to eat?) had substantial DIF between young and old adults. After deleting these 7 DIF items from the respective subscales and reanalyzing the data, we found that none of the remaining items exhibited substantial DIF. The right-hand side of Table 1 shows the infit and outfit MNSQ statistics for the remaining 19 items across the four subscales, ranging from 0.81 to 1.30. As they were allocated within the (0.7, 1.3) critical range and no substantial DIF was found, we concluded that these 19 items fit the model's expectation fairly well. That is, the original 28-item WHOQOL-BREF Taiwan version (excluding the 2 general items), in its complete form, comprised more than four dimensions. When these 7 DIF items were deleted, the remained 19 items assessed four latent traits that corresponded to the original test construction.

The measures and standard errors for the item difficulties and the category boundary parameters for the four subscales are listed in Table 2. The ranges of the category boundary parameters for the four subscales were sufficiently large, as compared to the distributions of the person measures. Furthermore, the ordered natures for the category boundary parameters were held, indicating that the 5-point rating scales were appropriate. However, the category boundary parameters were not very similar within or between subscales, even after the standard errors were taken into consideration. For example, the four category boundary parameters of item 3 (To what extent do you feel that (physical) pain prevents you from doing what you need to do?) were −1.93, −0.25, −0.16, and 2.35, respectively, whereas those for item 10 (Do you have enough energy for everyday life?) were −2.72, −1.71, 0.65 and 3.79, respectively. That is, there was an interaction between category labels and items: The participants treated the labels of the five response categories differently for different items.

Correlations between the four subscales

Table 3 shows the direct estimates of the variance-covariance and correlation matrices for the four

Table 2. Measures and standard errors (in parentheses) for the item difficulties and category boundary parameters

Subscale/item	Difficulty	Step 1	Step 2	Step 3	Step 4
Psychological					
5 Positive feelings	1.32 (0.02)	−3.00 (0.04)	−1.27 (0.03)	1.00 (0.03)	3.26 (0.04)
6 Spirituality/religion/personal beliefs	−0.21 (0.02)	−2.47 (0.05)	−0.68 (0.04)	0.24 (0.03)	2.90 (0.05)
7 Thinking, learning, memory, and concentration	0.03 (0.02)	−2.80 (0.06)	−1.27 (0.04)	0.41 (0.03)	3.66 (0.06)
11 Bodily image and appearance	−0.57 (0.02)	−2.16 (0.08)	−1.51 (0.05)	0.73 (0.03)	2.94 (0.08)
19 Self-esteem	−0.54 (0.03)	−2.71 (0.10)	−1.50 (0.05)	0.32 (0.04)	3.89 (0.10)
26 Negative feelings	−0.03 (0.02)	−2.25 (0.06)	−1.53 (0.04)	0.42 (0.03)	3.36 (0.06)
Physical					
3 Pain and discomfort	−0.45 (0.02)	−1.93 (0.09)	−0.25 (0.05)	−0.16 (0.04)	2.35 (0.09)
4 Dependence on medical substances and medical aids	−0.77 (0.02)	−1.33 (0.09)	−0.11 (0.06)	0.21 (0.04)	1.23 (0.09)
10 Energy and fatigue	0.61 (0.02)	−2.72 (0.07)	−1.71 (0.04)	0.65 (0.03)	3.79 (0.07)
16 Sleep and rest	0.61 (0.02)	−2.65 (0.06)	−1.15 (0.04)	0.19 (0.03)	3.61 (0.06)
17 Activities of daily living	−0.10 (0.04)	−2.92 (0.16)	−1.73 (0.07)	0.46 (0.06)	4.19 (0.15)
18 Work capacity	0.09 (0.02)	−2.67 (0.10)	−1.47 (0.05)	0.25 (0.04)	3.90 (0.11)
Social					
20 Personal relationships	−0.22 (0.03)	−3.59 (0.11)	−1.98 (0.05)	0.71 (0.04)	4.86 (0.14)
22 Practical social support	−0.39 (0.03)	−3.27 (0.13)	−2.53 (0.07)	0.53 (0.5)	5.28 (0.21)
27 Being respected/ accepted	0.61 (0.03)	−3.54 (0.07)	−2.08 (0.04)	0.83 (0.03)	4.79 (0.09)
Environmental					
8 Freedom, physical safety and security	0.43 (0.02)	−2.58 (0.05)	−1.63 (0.04)	0.18 (0.03)	4.03 (0.07)
9 Physical environment: (pollution/noise/traffic/climate)	−0.27 (0.02)	−2.81 (0.08)	−1.67 (0.04)	0.47 (0.03)	4.01 (0.09)
24 Health and social care: accessibility and quality	0.08 (0.02)	−2.86 (0.07)	−2.09 (0.04)	0.56 (0.03)	4.39 (0.09)
25 Transport	−0.25 (0.02)	−2.89 (0.09)	−1.94 (0.05)	0.37 (0.03)	4.46 (0.11)

Table 3. Variances, covariances, and correlations for the four subscales (N = 13,083)

Subscale	Psychological	Physical	Social	Environmental
Psychological	1.22	1.05	1.48	1.10
Physical	*0.88*	1.16	1.30	1.04
Social	*0.89*	*0.80*	2.26	1.54
Environmental	*0.81*	*0.79*	*0.84*	1.49

Note: Values in the lower triangles (italics) and higher triangles are the correlations and covariances, respectively.

subscales. The correlations ranged from 0.79 to 0.89, suggesting that they were highly correlated. The high correlations were consistent with what has been found by Skevington et al. [5], in which the factor loadings of the four subscales to the common factor of QOL were between 0.83 and 0.95. The high correlations reflected convergent validity of the WHOQOL-BREF, as the four subscales were designed to assess a broad concept of QOL. For comparison, the unidimensional partial credit model was also fitted to each of the four subscales separately, one subscale at a time. The correlations on the resulting person measures ranged from 0.53 to 0.65, indicating that the four subscales were only moderately correlated. The correlation estimates were actually attenuated by measurement error in the person measures [42]. Classical analyses usually ignore measurement error, leading to the correlations being underestimated. This might explain why the four subscales are often found to be only moderately correlated in the literature (e.g., between 0.61 and 0.76 in Hsiung et al. [3]). Using the multidimensional approach, we took measurement error into account so that the derived correlation estimates would reflect more appropriately the strength of association among the four subscales [43].

Test reliability

The test reliabilities for the four subscales obtained from the multidimensional analysis ranged from 0.82 to 0.86, as shown in Table 4. Even though the four subscales consisted of only three to six items, their reliabilities appeared to be very satisfactory. These high reliabilities were achieved because the high correlations between the four subscales were considered *via* the multidimensional analysis. When the correlations were ignored and the unidimensional partial credit model was fitted to each of the four subscales separately, the resulting test reliabilities ranged from 0.67 to 0.75. Apparently, the multidimensional approach yielded substantially higher reliabilities than the unidimensional one. According to the Spearman-Brown Prophecy formula [44], the test would have to be increased 103% in length (i.e., from 6 items to 12.2 items) in order to achieve a reliability of 0.83, up from 0.75 for the psychological subscale; 114% in length in order to achieve a reliability of 0.84, up from 0.71 for the physical subscale; and so on for the other subscales, when the unidimensional analysis is conducted. An increment of 103% or 114% in test length depicts the relative efficiency of the multidimensional analysis over the unidimensional analysis.

The composite unidimensional approach

One might question whether these four subscales should be treated as unidimensional, because they were so highly correlated. To reply to this question statistically, we also conducted a unidimensional analysis in which all the 26 items were treated as

Table 4. Test reliabilities obtained from the multidimensional and unidimensional approaches and increment in test length

Subscale	Test length	Reliability (Multidimensional)	Reliability (Unidimensional)	Increment in test length
Psychological	6	0.86	0.75	103%
Physical	6	0.84	0.71	114%
Social	3	0.83	0.67	138%
Environmental	4	0.82	0.68	120%

measuring a single latent trait, even though the multidimensional approach is more consistent with the construction and usual practice of the WHOQOL-BREF than the "composite" unidimensional approach. It turned out that a total of 8 items (including those aforementioned 7 DIF items and item 4 "How much do you need any medical treatment to function in your daily life?") had substantial DIF. When they were removed, the remaining 18 items fitted the model's expectation fairly well. The test reliability was 0.87, which was only slightly higher than those obtained from the multidimensional analysis for the four individual subscales, which fell in the range between 0.82 and 0.86.

The norms

Table 5 lists the means and standard deviations of the person measures for the 10 groups as well as the whole group, both on the Rasch scale and the transformed 0–100 scale. As the population was assumed to follow the normal distribution, one can easily obtain percentiles using the mean and standard deviation of the normal distribution. Generally, there was little gender difference on the four subscales. Considerable decline on the psychological and physical subscales was found for older groups, whereas little age difference was found for the social and environmental subscales. Finally, persons with higher education levels gave higher ratings on all four subscales.

Discussion

We followed standard perspectives in treating the WHOQOL-BREF as four dimensional and analyzed all of the four dimensional data jointly, using the multidimensional approach. After excluding 2 overall items and deleting 7 items

Table 5. Means and standard deviations on the Rasch and 0–100 transformed measures for the four subscales across groups

Group		Rasch scale				0–100 scale			
		Psychological	Physical	Social	Environmental	Psychological	Physical	Social	Environmental
Gender									
Men (n = 6628)	M	0.66	1.19	1.05	0.66	59.99	50.75	56.42	42.63
	SD	1.11	1.07	1.54	1.26	13.69	15.63	14.66	15.41
Women (n = 6455)	M	0.51	1.01	1.09	0.62	58.14	48.12	56.80	42.14
	SD	1.07	1.02	1.46	1.19	13.20	14.90	13.90	14.55
Age									
20–29 (n = 3591)	M	0.67	1.21	1.25	0.72	60.11	51.04	58.32	43.36
	SD	1.06	1.02	1.50	1.27	13.07	14.90	14.28	15.53
30–39 (n = 3684)	M	0.66	1.19	1.10	0.60	59.99	50.75	56.89	41.89
	SD	1.07	1.05	1.52	1.24	13.20	15.33	14.47	15.17
40–49 (n = 3418)	M	0.56	1.10	1.01	0.66	58.75	49.43	56.04	42.63
	SD	1.12	1.08	1.49	1.19	13.81	15.77	14.18	14.55
50–59 (n = 1837)	M	0.42	0.91	0.87	0.51	57.03	46.66	54.70	40.79
	SD	1.10	1.05	1.52	1.18	13.57	15.33	14.47	14.43
60 + (n = 553)	M	0.42	0.66	0.99	0.69	57.03	43.01	55.85	42.99
	SD	1.15	1.07	1.58	1.28	14.18	15.63	15.04	15.66
Education status									
Elementary (n = 2410)	M	0.21	0.68	0.70	0.46	54.44	43.30	53.09	40.18
	SD	1.06	1.04	1.46	1.19	13.07	15.19	13.90	14.55
Secondary (n = 6881)	M	0.57	1.13	1.02	0.64	58.88	49.87	56.13	42.38
	SD	1.08	1.06	1.48	1.24	13.32	15.48	14.09	15.17
College (n = 3791)	M	0.83	1.33	1.40	0.78	62.08	52.79	59.75	44.09
	SD	1.05	1.00	1.49	1.23	12.95	14.60	14.18	15.04
Total (N = 13,083)	M	0.59	1.10	1.06	0.64	59.12	49.43	56.51	42.38
	SD	1.11	1.07	1.50	1.22	13.69	15.63	14.28	14.92

from the 28-item WHOQOL-BREF Taiwan version, we conclude that each of the four subscales assesses a single latent trait. The four latent traits are very highly correlated. Using the unidimensional approach, one may mistakenly conclude that they are only moderately correlated, as measurement error is ignored. When the high correlations between the four latent traits are taken into account *via* the multidimensional approach, the test reliabilities for the four subscales are increased to a range of 0.82 to 0.86, which might be accurate for diagnosing individuals. In contrast, when the high correlations are ignored and the unidimensional analysis is conducted, the resulting test reliabilities are between 0.67 and 0.75, which are too low for use in diagnosing individuals. Clearly, the multidimensional approach not only yields more appropriate estimates for the association of four latent traits, but also improves the usefulness of the 19-item WHOQOL-BREF substantially.

Through the infit and outfit fit statistics and DIF analysis, we remove 7 misfitting items from the WHOQOL-BREF, not because these 7 misfitting items do not constitute significant aspects of QOL, but because they do not assess the same latent traits as the remaining 19 items for the general population in Taiwan. From DIF analysis, we gain a better understanding about why these 7 items do not fit the model's expectation. This understanding may be of value in item revision or scale development. For example, item 21 (How satisfied are you with your sex life?) in the social subscale has substantial DIF between young and old groups. Removing item 21 from the social subscale does not mean that "sex life" is a trivial aspect of QOL. Rather, "sex life" does not work harmoniously with the other three items (How satisfied are you with your personal relationships? How satisfied are you with the support you get from your friends? Do you feel respected by others?), especially between young and old adults. If satisfaction with one's sex life indeed contributes to QOL, then either the linguistic statement of item 21 should be revised to better reflect the "social" aspect of QOL and work harmoniously with the other three items, or a stand-alone "sex life" subscale should be developed. Future studies may be conducted to revise the linguistic statements of these 7 DIF items, or to develop stand-alone scales for these important but left-behind aspects of QOL.

The means and standard deviations for the 10 groups as well as the whole sample are provided. Under the normality assumption, one can use the mean and standard deviation to obtain percentiles. For the general population in Taiwan, little gender difference is found on the four subscales of QOL. The older the participants, the less satisfaction they report in the psychological subscale. In addition, persons older than 50 reported considerably lower satisfaction in the physical subscale. Little age difference exists in either the social or environmental subscale. Finally, persons with higher education levels give higher ratings across the four subscales. The norms can be used as reference for the interpretation of data from general samples.

From the point of view of test developers, incorporating many diverse aspects in a test increases construct bandwidth. For example, each of the 9 items in the environmental subscale was designed to tap an important environmental factor. Deleting any of these items may threaten the construct validity that the test developers intended. However, whether all the items of a test tap a single latent trait should be examined empirically. Through Rasch analysis, we diagnosed unexpected disturbances in the data. Only when the data fit the model's expectation can the underlying latent trait be quantified. In accordance with this logic, misfitting items or persons should be identified, revised, or removed. After the removal of 5 misfitting items from the 9-item environmental subscale, the remaining 4-item environmental subscale reflects a succinct construct about "residential environment", which is of course narrower than the bandwidth of the original 9-item test. However, according to the Rasch analysis, the 9-item test does not constitute a single latent trait, so the derived measures are meaningless.

We followed the standard procedures in treating the WHOQOL-BREF as four dimensional. The correlations of the four latent traits were directly estimated *via* the multidimensional approach and found to be very high. A high correlation between latent traits does not necessarily mean they are the same latent trait; for example, height and weight are highly correlated, but they are two different concepts. For the reader's convenience, we also adopted the composite unidimensional approach to treat the WHOQOL-BREF as measuring a single latent trait. After removing 8 DIF items, one may conclude that the remaining 18 items constitute a single latent trait, whose meaning is not yet clear and is not consistent with the construction and usual practice of the WHOQOL-BREF. Application and implication of the 18-item version need further investigation. Besides, with the use of the composite unidimensional approach, the information about the four individual subscales is invisible because only a composite score can be obtained. On the contrary, with the use of the multidimensional approach, the profile information (scatter pattern) about the individual latent traits is preserved. Furthermore, the reliability of the 18-item test is only slightly higher than those obtained from the multidimensional analysis for the 3- to 6-item tests. Apparently, the multidimensional approach is preferable because it corresponds to the structure with which the WHOQOL-BREF was developed and the standard practice in how the WHOQOL-BREF is analyzed, yielding high reliability for the four subscales even when they consist of only 3 to 6 items.

The multidimensional Rasch analysis in this study is confirmatory rather than exploratory in that the structure of dimensionality is specified to meet the construction and usual practice of the WHOQOL-BREF. When data conform to the model's expectation, the pre-specified dimensionality is not rejected. Confirmatory approaches allow researchers to formulate a number of theories about dimensionality and item weights, based on the researcher's theoretical knowledge and previous research findings. Exploratory approaches enable researchers to get an impression of the number of latent traits and of which items are indicative of which latent traits. Although "within-item" multidimensional analysis (in which some items assess more than one latent trait simultaneously) is possible under the MRCMLM framework [14, 45], it is not appropriate in this study due to the lack of theoretical within-item dimensionality for the WHOQOL-BREF.

In developing tests, there is an inevitable conflict between the goal of attaining precision (reliability) in measurement and the goal of attaining breadth, which is referred to as the bandwidth-fidelity dilemma [46, 47]. Bandwidth refers to the amount of information that is contained in a message, while fidelity refers to the accuracy with which the information is conveyed. The greater the amount of information to be conveyed (bandwidth), the less accurately it can be conveyed (fidelity). For any given test, a choice must be made between measuring a very specific attribute with a high degree of accuracy, and measuring a broader range of attributes with less accuracy. In practice, the time and resources available for testing are typically limited. Consequently, testers sometimes have to sacrifice accuracy (i.e., one long test for a single latent trait) and develop several short tests to cover as many important attributes as the testing time allows. In this study, it is shown that the multidimensional approach can increase test reliability for short tests to a more satisfactory level. Therefore, the bandwidth-fidelity dilemma is at least partially resolved *via* the multidimensional approach.

In unidimensional Rasch models, raw scores are sufficient statistics for Rasch person measures, meaning that persons with identical raw scores will always have the same person measures (assuming that they respond to the same set of items). When the multidimensional approach is adopted, raw scores of any subscale alone are no longer sufficient statistics for person measures on that subscale. That is, two persons with identical raw scores on a subscale may have different person measures on that subscale. In fact, person measures of a subscale depend on not only raw scores of that subscale, but also raw scores of the other subscales. Therefore, vectors of raw scores (including all raw scores over subscales) are sufficient statistics for person measures over subscales. Only when two persons have exactly the same raw score patterns over subscales will they have identical person measures, assuming that they respond to the same set of items. Consequently, the look-up table for transforming raw scores to person measures will be very long. A computer program was written to yield the transformation and is available on request.

When multiple latent traits are measured, score comparisons across latent traits are sometimes found. We warn the readers that score comparisons should be interpreted very cautiously, if they have to be done at all. Take the scores in Table 5, for example. Inspecting the mean Rasch scores for the four subscales, one may draw the conclusion that people in Taiwan score the highest (i.e., have the highest degree of satisfaction) on the physical subscale and the lowest on the psychological subscale. This comparison is actually made on the basis of the characteristics of those items that constitute the subscales. It is certainly possible to create items for the physical subscale that are extremely difficult to endorse (e.g., 100-m dash in 10 s) so that the mean score of the physical subscale is the lowest among the four subscales. The four subscales are four different latent traits. Hence, the four scales are actually qualitatively different. In this regard, we do not recommend cross-subscale comparison. If cross-subscale comparisons have to be made, we advise that the practitioner take into account the characteristics of those items that constitute the subscales.

Cross-subscale comparisons are not only item-dependent but also scale-dependent. For example, the mean scores of the psychological subscale are the lowest on the Rasch scale ($M = 0.59$), but the highest on the 0–100 scale ($M = 59.12$). Apparently, different scales lead to different rankings. If one of these two scales has to be chosen for cross-subscale comparison, we recommend the Rasch scale, as the mean Rasch scores across subscales can be interpreted on the basis of equal mean item measures (i.e., the mean item measures of every subscale are set at zero for identification of the parameters). In contrast, the 0–100 scale can be seriously affected by the range of score distributions within subscales. The 0–100 scale is obtained from the Rasch scale *via* the following linear transformation:

$$Y = \frac{100(X - \text{Min})}{\text{Max} - \text{Min}}, \tag{6}$$

where Y is the score on the 0–100 scale; X is the score on the Rasch scale; Max and Min are the maximum and minimum on the Rasch scale, respectively. Clearly, a large range (Max−Min) on the Rasch scale, which may due to a single person who scores extremely low, can cause a significant decline for the mean score on the 0–100 scale. This is the reason why the physical subscale has the highest mean on the Rasch scale ($M = 1.10$), but only the third highest on the 0–100 scale ($M = 59.43$).

In summary, the 19-item WHOQOL-BREF measures more succinct latent traits than the original design. The multidimensional approach not only improves the estimation accuracy for the correlation between subscales but also yields substantially higher reliabilities than the standard unidimensional approach. The multidimensional approach is very general and can be easily applied to any test that contains subtests, any scale that contains subscales, any test battery that contains multiple tests, or multiple tests that do not belong to the same test battery.

Appendix. Parameter Estimation Procedures for the MRCMLM

Marginal maximum likelihood estimation with Bock and Aitkin's [48] formulation of the EM (Expectation-Maximization) algorithm [49] is implemented in ConQuest to estimate all the parameters in the MRCMLM ($\boldsymbol{\xi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) simultaneously, so that measurement errors in $\boldsymbol{\theta}$ are directly taken into account. Based on the assumption of conditional independence among items and persons, the probability of a response vector $\mathbf{x}$ conditioned on the random quantities $\boldsymbol{\theta}$ is

$$p(\mathbf{X} = \mathbf{x}; \xi | \boldsymbol{\theta}) = \frac{\exp[\mathbf{x}'(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\xi)]}{\Psi(\boldsymbol{\theta}, \xi)}, \tag{A.1}$$

with

$$\Psi(\boldsymbol{\theta}, \xi) = \sum_{z \in \Omega} \exp[\mathbf{z}'(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\xi)], \tag{A.2}$$

where is the set of all possible response vectors. The marginal density of the response $\mathbf{x}$ is

$$p(\mathbf{X} = \mathbf{x}) = \int_{\boldsymbol{\theta}} \frac{\exp[\mathbf{x}'(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\xi)]}{\Psi(\boldsymbol{\theta}, \xi)} \mathrm{d}G(\boldsymbol{\theta}, \boldsymbol{\alpha}), \tag{A.3}$$

where G is the cumulative distribution of g. The likelihood for a set of N response vectors is

$$\Lambda(\xi, \boldsymbol{\alpha} | \mathbf{X}) = \prod_{n=1}^{N} \int_{\boldsymbol{\theta}} \frac{\exp[\mathbf{x}_n'(\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\xi)]}{(\boldsymbol{\theta}, \xi)} \mathrm{d}G(\boldsymbol{\theta}, \boldsymbol{\alpha}). \tag{A.4}$$

The likelihood equations for the item parameters are

$$\frac{\partial \log \Lambda(\boldsymbol{\xi}, \boldsymbol{\alpha}|\mathbf{X})}{\partial \boldsymbol{\xi}} = \sum_{n=1}^{N} \int_{\boldsymbol{\theta}} \frac{\partial \log p(\mathbf{x}'_n; \boldsymbol{\xi}, \boldsymbol{\alpha})}{\partial \boldsymbol{\xi}} \mathrm{d}H(\boldsymbol{\theta}; \boldsymbol{\xi}, \boldsymbol{\alpha}|\mathbf{x}_n) = \mathbf{0}, \tag{A.5}$$

where $H(\boldsymbol{\theta}; \boldsymbol{\xi}, \boldsymbol{\alpha}|\mathbf{x}_n)$ is the cumulative posterior marginal distribution of $\boldsymbol{\theta}$ given $\mathbf{x}_n$, with a density function

$$h(\boldsymbol{\theta}; \boldsymbol{\xi}, \boldsymbol{\alpha}|\mathbf{x}_n) = \frac{p(\mathbf{x}_n; \boldsymbol{\xi}|\boldsymbol{\theta}) g(\boldsymbol{\theta}; \boldsymbol{\alpha})}{p(\mathbf{x}_n; \boldsymbol{\xi})}. \tag{A.6}$$

Assuming the distribution of the latent traits is multivariate normal so that $\boldsymbol{\alpha} \equiv (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the likelihood equations for the mean and variance-covariance matrix are

$$\frac{\partial \log \Lambda(\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})}{\partial \boldsymbol{\mu}} = \sum_{n=1}^{N} \int_{\boldsymbol{\theta}} \frac{\partial \log g(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} \mathrm{d}H(\boldsymbol{\theta}; \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_n) = \mathbf{0}, \tag{A.7}$$

and

$$\frac{\partial \log \Lambda(\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X})}{\partial \boldsymbol{\Sigma}} = \sum_{n=1}^{N} \int_{\boldsymbol{\theta}} \frac{\partial \log g(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} \mathrm{d}H(\boldsymbol{\theta}; \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{x}_n) = \mathbf{0}. \tag{A.8}$$

Note that only the form of the multivariate normal distribution is assumed and the corresponding mean vector and variance-covariance matrix are empirically estimated, which is referred to as the empirical Bayes method [50]. After the model parameters are calibrated, point estimates for individual persons can be obtained from either the mean vector of the marginal posterior distribution, Equation A.6, called the expected a posteriori estimates [51], or the maximum point of conditional likelihood, called the maximum likelihood estimates.

Acknowledgements

This study was partly supported by research grants from the National Health Research Institute (HNRI-EX92-9204PP) and the National Science Council (NSF93-2413-H-194-004). The authors would be grateful to National Health Research Institute and Bureau of Health promotion, Department of Health, Taiwan, ROC for conducting the National Health Interview Survey and granting the data.

References

1. The WHOQOL group. The World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. Soc Sci Med 1998; 46: 1569–1585.
2. The WHOQOL-Taiwan Group. User Manual of the WHOQOL-BREF Taiwan Version., 1st edn., Taipei, Taiwan: Institute of Occupational Medicine and Industrial Hygiene, National Taiwan University College of Public Health Press, 2000.
3. Hsiung PC, Fang CT, Chang YY, Chen MY, Wang JD. Comparison of WHOQOL-BREF and SF-36 in patients with HIV infection. Qual Life Res 2005; 14: 141–150.
4. Hwang HF, Liang WM, Chiu YN, Lin MR. Suitability of the WHOQOL-BREF for community-dwelling older people in Taiwan. Age Ageing 2003; 32: 593–600.
5. Skevington SM, Lotfy M, O'Connell KA. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. Qual Life Res 2004; 13: 299–310.
6. Noerholm V, Bech P. The WHO Quality of Life (WHOQOL) Questionnaire: Danish validation study. Nord J Psychiatry 2001; 55: 229–235.
7. Noerholm V, Groenvold M, Watt T, Bjorner JB, Rasmussen NA, Bech P. 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: 531–540.
8. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978; 43: 561–573.
9. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Institute of Educational Research, 1960.
10. Wright BD, Stone MH. Best Test Design. Chicago: Measurement, Evaluation, Statistics, and Assessment Press, 1979.
11. Wang WC, Chen PH, Cheng YY. Improving measurement precision of test batteries using multidimensional item response models. Psychol Methods 2004; 9: 116–136.
12. Shih YT, Hung YT, Chang HY, et al. The design, contents, operation and characteristics of the respondents of the 2001

National Health Interview Survey in Taiwan. Taiwan J Public Health 2004; 22: 419–430.

13. Yao G, Chung CW, Yu CF, Wang JD. Development and verification of validity and reliability of the WHOQOL-BREF Taiwan version. J Formos Med Assoc 2002; 101: 342–351.
14. Adams RJ, Wilson M, Wang WC. The multidimensional random coefficients multinomial logit model. Appl Psychol Meas 1997; 21: 1–23.
15. Fischer GH. The linear logistic test model as instrument in educational research. Acta Psychol 1973; 37: 359–374.
16. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–174.
17. Wilson MR. The partial order model: an extension of the partial credit model. Appl Psychol Meas 1992; 16: 309–325.
18. Linacre JM. Many-facet Rasch Measurement. Chicago: Measurement, Evaluation, Statistics, and Assessment Press, 1989.
19. Fischer GH, Pononcy I. An extension of the partial credit model with an application to the measurement of change. Psychometrika 1994; 59: 177–192.
20. Adams RJ, Wilson MR. Formulating the Rasch model as a mixed coefficients multinomial logit. In: Englhard G, Wilson M (eds.), Objective Measurement: Theory into Practice 3. Norwood, NJ: Ablex, 1996.
21. De Boeck P, Wilson MR (eds), Explanatory item response models: a generalized linear and nonlinear approach. New York: Springer-Verlag, 2004.
22. McCullagh P, Nelder JA. Generalized Linear Models., 2nd edn., London: Chapman & Hall, 1989.
23. McCulloch CE, Searle SR. Generalized, Linear, and Mixed Models. New York: Wiley, 2001.
24. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A 1972; 135: 370–384.
25. Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P. A nonlinear mixed model framework for item response theory. Psychol Methods 2003; 8: 185–205.
26. Wu ML, Adams RJ, Wilson MR. ConQuest. Camberwell, Victoria, Australia: Australian Council for Educational Research, 1998.
27. SAS Institute. The NLMIXED procedure. Cary, NC: Author, 1999.
28. Wolfinger RD, SAS Institute Inc. Fitting nonlinear mixed models with the new NLMIXED procedure: Proceedings of the 99 Joint Statistical Meetings, 1999.
29. StataCrop. Stata Statistical Software: Release 8.0. College Station. TX: StataCorp LP, 2003.
30. Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton, FL: Chapman & Hall/ CRC Press, 2004.
31. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. Comput Methods Programs Biomed 1996; 49: 157–176.
32. Hedeker D. MIXNO. Chicago: University of Illinois Press, 1999.
33. Lord FM. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum, 1980.
34. Holland PW, Wainer H. Differential Item Functioning. Hillsdale, NJ: Erlbaum, 1993.
35. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 2000.
36. Wright BD, Linacre JM, Gustafson J-E, Martin-Lof P. Reasonable mean-square fit values. Rasch Measurement Transactions. Rasch Meas Trans 1994; 8: 370.
37. Zimowski MF, Muraki E, Mislevy RJ, Bock RD. BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items. Chicago, IL: Scientific Software international, 1996.
38. Linacre J. A User's Guide to WINSTEPS MINISTEP Rasch-model Computer Programs. Chicago: John M. Linacre, 2003.
39. Wang WC, Yeh YL. Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. Appl Psychol Meas 2003; 27: 479–498.
40. Hung YT. Sampling design of the National Health Interview Survey: NHIS Brief Communication No. 2; 2002.
41. Lin SH. Field collection and completeness of data in the National Health Interview Survey: NHIS Brief Communication No. 4; 2002.
42. Spearman C. "General intelligence" objectively determined and measured. Am J Psychol 1994; 15: 201–293.
43. Wang WC. Direct estimation of correlation as a measure of association strength using multidimensional item response models. Educ Psychol Meas 2004; 64: 937–955.
44. Lord FM, Novick M. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.
45. Hoijtink H, Rooks G, Wilmink FW. Confirmatory factor analysis of items with a dichotomous response format using the multidimensional Rasch model. Psycho Methods 1999; 4: 300–314.
46. Cronbach LJ, Gleser G. Psychological Tests and Personnel Decision., 2nd edn., Urbana, IL: University of Illinois Press, 1965.
47. Shannon C, Weaver W. The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1949.
48. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 1981; 46: 443–459.
49. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 1977; 39: 1–38.
50. Lee PM. Bayesian Statistics: An Introduction. New York: Oxford University Press, 1989.
51. Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas 1982; 6: 431–444.

Address for correspondence: Wen-Chung Wang, Department of Psychology, National Chung Cheng University, Chia-Yi, 621, Taiwan
E-mail: psywcw@ccu.edu.tw