Iran J Public Health, Vol. 44, No.4, Apr 2015, pp.522-534

Original Article

Developing a Shortened Quality of Life Scale from Persian Version of the WHOQOL-100 Using the Rasch Analysis

Kamran YAZDANI [1], Saharnaz NEDJAT [1,2], Masoud KARIMLOU [3], Hojjat ZERAATI [1], Kazem MOHAMMAD [1], Akbar FOTOUHI [1]

1. *Dept. of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran*
2. *Knowledge Utilization Research Center, Tehran University of Medical Sciences, Tehran, Iran*
3. *Dept. of Statistics and Computer, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran*

***Corresponding Author:** Email: kyazdani@tums.ac.ir

(Received 20 Nov 2014; accepted 21 Jan 2015)

Abstract

Background: Few studies use modern approaches to develop a quality of life (QOL) questionnaire with acceptable construct validity, especially in Iran. Our main objective was to construct a new validated and uni-dimensional questionnaire, based on WHOQOL-100, using the Rasch analysis.

Methods: In a population-based, cross-sectional study in 2007, 500 Tehran residents aged≥18 were randomly sampled. The Persian version of WHOQOL-100 was used to measure the participants' QOL. After using targeting and person fit analysis, we performed category/threshold ordering, item fit, and differential item functioning analyses, in succession. We used outfit or infit statistics>1.5 and <0.5 for detecting underfit and overfit items/persons, respectively. We also deleted items with disordered category and/or threshold. Person Separation Index and test reliability were also calculated in the datasets.

Results: Male to female ratio was 0.98 and the mean age (SD) of participants was 35.1 (12.8) years. Initial analyses showed significant differences in quality of life between age groups (P=0.002), educational levels (P=0.001), and current health status groups (P<0.001). We eliminated 67 persons for under-fitting, 38 items for category and/or threshold disordering, 6 items for under-fitting, and 26 items for item bias. Test reliability for the final 30-item scale was 0.89.

Conclusion: We prepared a shortened version of the WHOQOL-100 that is single construct, uni-dimensional and free of item bias or any disordering, according to the Rasch model.

Keywords: Rasch analysis, Quality of life, WHOQOL-100

Introduction

In recent decades, the "Health-Related Quality of Life (HRQOL)" has been increasingly used as an important health outcome (1-3). According to the World Health Organization (WHO), the QOL is defined as a latent variable (4), so its measurement is based on each individual's performance in responding to a set of items or questions. To achieve this, two paradigms have been used: Classical Test Theory (CTT) and modern approaches (5). The classical approach is usually based on the assumption that an observed score consists of two components: a true score, and an error score (5, 6) in which the latter is not related to the former, which is one of the strengths of this model (6). However, raw scores used in this model are ordinal, and mathematical operations such as addition or subtraction cannot be used; therefore, it cannot be considered as "measurement" (7, 8). Potential

problems that can result from this have led to the use of modern approaches (5, 6), including Item Response Theory (IRT) (9) and Rasch analysis that claims to be able to achieve fundamental measurement (10, 11).

The basic Rasch model, named after its inventor the Danish mathematician, Georg Rasch, is a probabilistic model in which the natural logarithm of the odds of giving the right answer to a dichotomous item is a linear function of the difference between a person's ability (or location) and the item difficulty (or calibration):

$$\mathrm{Log}_n (P_{ni1} / P_{ni0}) = B_n - D_i$$

In this basic equation, P_{ni1} is the probability of answering "yes" or "correct" and P_{ni0} bears the opposite meaning. B_n and D_i denote the person and item measures, respectively (12). This model is believed to provide "simultaneous additive conjoint measurement" (13) that can translate raw scores to latent measures if the data fit the model (14). In this model, two parameters of ability and difficulty can be estimated independently. Then, through using some statistics, the appropriateness of the items and people's response pattern can be checked (12, 15, 16). This model can also be expanded to be used for the items with rating scales response patterns with more than two categories, e.g. the Likert scale (polytomous models: Rating Scales and Partial Credit) (15).

One of the major applications of this model is in the evaluation of the construct validity of the questionnaires. The way the model is used in these evaluations will differ with regard to the goal of the study and type of the questionnaire (17-26). Item reduction (deletion) in an available questionnaire is one of these applications. It is done in order to exclude unsuitable questions and decrease the number of questions to increase the individuals' compliance and cooperation. In addition, in this method, we can be sure that the remainder of the questions fit the model and can provide a unidimensional Rasch scale or construct (17-19) through which we can achieve an objective (fundamental) measurement (14).

Among the several questionnaires that have been designed to evaluate the QOL (27, 28), the WHOQOL-100 and the WHOQOL-BREF are two of the most famous generic scales that belong to the generic category (4, 29-33). The WHOQOL-BREF was prepared using the CTT approach and has a smaller number of questions and comparable validity with the WHOQOL-100 (33).

To evaluate the validity and psychometric properties of these WHOQOL questionnaires, Nedjat et al (34) and Karimlou et al (35) conducted studies on their Persian version, using the CTT approach, where the latter showed acceptable reliability and validity for Iranians, except for the spiritual domain.

Since no modern approach study has investigated the psychometric characteristics of the Persian version of the WHOQOL-100 questionnaire, our objective was to use Rasch analysis to select its best questions to develop a shortened unidimensional questionnaire with higher construct validity.

Methods

The methodology of this study is mainly based on the Tennant (17) recommendations and the methods used by Leplege (18) and Ecosse (19). The main objective was to delete the items that do not fulfill the Rasch measurement quality control criteria.

The data used is a part of Karimlou's study (35) database; so, the details of the sampling strategy, translation process, and data collection have been discussed elsewhere (35) and are described here in brief.

Persian version of the WHOQOL-100 questionnaire

The WHOQOL-100 questionnaire has been translated to Persian according to the WHO international guidelines (36). The steps include forward translation, review of the face and content validity, backward translation, final check, pilot study, and a final reliability check (test-retest) at a 2-week interval.

This scale comprises Physical, Psychological, Level of independence, Social, Environmental, and Spirituality domains that altogether contain 24

facets each having 4 items, plus an additional facet for the overall QOL (30). All questions address the respondent's status in the past 2 weeks, and have a 5-category response from 1 to 5 that we changed to 0 to 4.

In addition to the 100 QOL items, we also asked some socio-demographic questions including age, sex, level of education, and marital status, and also a question about the current illness status (*Are you currently ill?* Yes / No).

Participants and data gathering

In a population-based, cross-sectional study, 500 healthy residents of Tehran who were aged 18 and older were sampled, using a multi-stage sampling method, during spring 2007. The exclusion criteria were having chronic (including diabetes, cardiovascular, amputee, or autoimmune) or acute diseases, and having psychiatric or mental disorders. Initially, stratification was done according to 22 municipal districts and subsequently, proportional to each district's population, a total number of 50 blocks were selected randomly as clusters. In each systematically selected household, one eligible person was sampled randomly.

Anonymous Questionnaires were administered by well-trained, supervised interviewers using a door-to-door approach. After completion, the data were rechecked and entered into an electronic database.

Verbal consents were taken from all participants, and the data were kept confidential. The Ethics Committee of the University of Social Welfare and Rehabilitation Sciences approved the study protocol and the School of Public Health of Tehran University of Medical Sciences approved it as a PhD dissertation.

Statistical analyses

In the first round, using Rasch measurement, we achieved the primary person and item measures. By using these measures, we performed the following analyses in succession.

According to the Likert type questions and based on Linacre's recommendations about choosing one of the polytomous models to analyze (37), the Rating Scale Analysis (RSA) was selected, and only to inspect the threshold ordering, we were forced to employ the Partial Credit Model (PCM).

We also used appropriate regular descriptive and analytical statistical analyses to describe baseline data and also to compare groups. In all analyses, if applicable, we considered $P < 0.05$ as significant.

In order to analyze the data, Winsteps® 3.68.2, SPSS version 11.5 for Windows, and Microsoft Office Excel 2003 were employed.

Targeting Analysis

Considering the negative effect of great differences between the person and item measures on fit statistics (14), observations for which expected scores in responding were too high or too low were excluded and treated as missing values. By determining cuthi and cutlo values equal to 1.75 in Winsteps® software (38), observations in which the absolute difference between individual and item measures, was more than 1.75 logit, were deleted. Setting this value in the first round of the analyses caused observations with expected score values more than about 3.5 or less than about 0.5 to change to missing values.

Person Fit Analysis

In this step, outfit (unweighted) and infit (weighted) mean square indexes were used to detect and delete underfit individuals because their responses could damage the validity of all measurements. Based on Linacre's opinion (38), outfit or infit statistics more than 1.5 were considered as underfit. Overfit individuals (outfit or infit < 0.5) were not removed from the dataset.

Category and Threshold Ordering

Since ordering of response categories and also thresholds are not always identical and disturbances in both need to be considered (39), all items with category or threshold disordering were identified and eliminated. Based on the Winsteps® user's guide (38), to identify category disordering, values of "average measures" for each response category (Table 13.3 in the software outputs), and to detect threshold disordering, values of "threshold measures" (Table 3.2 in the software outputs, "structure calibration") were used.

Item Fit Analysis

According to outfit or infit mean square criteria of more than 1.5 or less than 0.5, questions of underfit and overfit were identified and eliminated, respectively. Since item reduction was the main purpose of this study, the less efficient overfit items were removed, although they do not disturb the measurement.

DIF Analysis

According to person-free item measurement in Rasch model, item calibration must be the same in different subgroups of people (lack of item bias). In this study, the existence of DIF in subgroups based on gender (male/female), age (15-24 years, 25-34 years, 35-44 years, and ≥45 years), health status in an individual's opinion (ill/healthy) and level of education (primary school, secondary school, and college) was examined. The commonly used criterion to detect DIF is the presence of a significant difference of more than 0.5 logit in item calibration (40, 41) between two groups (dichotomous variables), or between one group with all groups combined (polytomous variables) (38). According to some limitation of this criterion, we decided to use an equivalence region or interval equal to -0.5 to +0.5 logits. The 95% confidence interval of a difference was calculated using its Standard Error (SE). Provided that this confidence interval had an overlap with the equivalence region, two measures were assumed to be equivalent and if it located out of this region on each side, the item was considered to have differential functioning or bias and was deleted. Joint SE, the SE value in comparing two groups, was computed using the following formula:

$$\text{Joint SE} = \sqrt{(SE_1^2 + SE_2^2)}$$

For comparing one group with all groups together, the sub-group SE was used.

Analysis of the Remaining Items

After deleting some observations, persons and items, person and item measures were again computed totally and also in subgroups in the new instrument. In addition, the remaining items were inspected with respect to fitness to the model, and threshold and category ordering. Person separation index and test reliability were also measured.

Results

Socio-demographic characteristics, current health status in the individual's opinion, and the average QOL measure of people in the initial database and without any omission, in both logit and a 0 to 100 scale, as total and in subgroups, are shown in Table 1. Among the 500 participants of this study, male to female ratio was 0.98. Mean and SD of age were 35.1 and 12.8 years (18-69 years), respectively. There was a statistically significant difference in current disease status between the two genders (P = 0.006) and among different age groups (P < 0.001). Overall, when compared to males, females and when compared to the youth, the elderly reported being sick more often.

Mean, standard deviation, minimum and maximum values of the difficulty of questions in the primary dataset were 0.00, 0.43, -1.37, and 0.86 logit, respectively. Person separation index and test reliability were 4.54 and 0.95, respectively.

Targeting

Fig. 1(a) shows distribution of measures for individuals and items (person-item bar chart). It clearly shows that questions were slightly off-target toward easiness for study participants. However, overall, the coverage of the instrument was well and there was no redundancy or gap over the item difficulty continuum.

After targeting analysis, 1878 missing observations were identified, which accounted for about 3.8% of total observations (1 person's observations were completely eliminated). The number of missing data was 387, about 1% of total data before conducting this analysis. Fig. 1 (b) shows the person-item bar chart after targeting analysis. Because Rasch analysis is robust to missing data (42), the distribution of measures was not substantially different. Comparison of fit statistics before and after targeting analysis did not reveal any significant differences (P = 0.131 to 0.920).

Table 1: Socio-demographic characteristics, current health status, and quality of life measure of participants in initial database

Characteristics	Number (%)	Measure (logit) Mean (SD)	Measure (0 – 100) Mean (SD)	*P* value
Gender				0.76
Male	247 (49)	0.26 (0.51)	50.65 (3.71)	
Female	253 (51)	0.28 (0.53)	50.75 (3.86)	
Age groups (year)				0.002
15-24	142 (28.4)	0.40 (0.56)	51.66 (4.08)	
25-34	128 (25.6)	0.22 (0.47)	50.29 (3.42)	
35-44	103 (20.6)	0.17 (0.48)	49.93 (3.48)	
> 45	127 (25.4)	0.27 (0.53)	50.66 (3.84)	
Educational level				0.001
Primary school	67 (13.5)	0.10 (0.47)	49.47 (3.38)	
Secondary school	261 (52.4)	0.24 (0.51)	50.50 (3.73)	
College	170 (34.1)	0.38 (0.54)	51.48 (3.88)	
Missing	2 (0.0)			
Are you currently ill?				< 0.001
Yes	133 (26.6)	-0.00 (0.48)	48.73 (3.46)	
No	367 (73.4)	0.37 (0.50)	51.41 (3.64)	
Total (Min, Max)	500 (100)	0.27 (0.52) (-1.18, 2.54)	50.70 (3.78) (40.21, 67.14)	

There was a significant difference between current health status of people and the number of missed observations ($P < 0.001$). The mean number of missing observations was higher among healthy people when compared to diseased people (4.32 vs. 1.49). In this regard, no significant relationship was detected between the number of missing observations and sex ($P = 0.981$); however, this number was significantly correlated with the level of education ($P = 0.001$) and age group ($P = 0.002$). Similarly, the number of missing values was higher in groups with higher QOL.

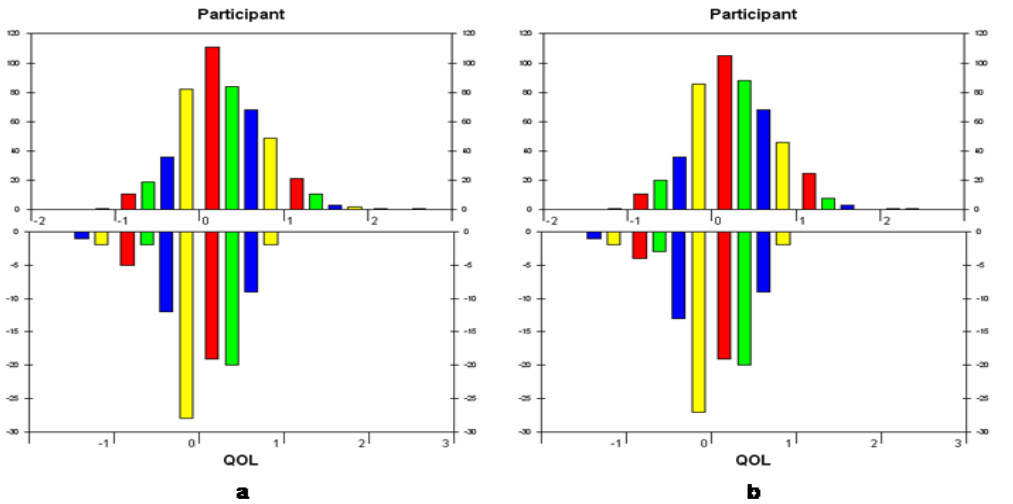

Fig. 1: (a) Item-person plot of initial data, (b) Item-person plot of data after targeting. Upper part shows distribution of individuals' measures and lower part shows questions measures. Both are in logit scale

Fig. 2 shows the relationship between the QOL measure and difficulty of questions with the number of missing data. Individuals with a better quality of life and easier items had more missing observations, which confirms being slightly off-target toward the easiness of questions in another way.

Person Fit

In person fit analysis, 67 individuals whose outfit or infit values were more than 1.5 were deleted from data. For all of these people, ZSTD (index for significance of fit statistics) value was more than 1.96, which corresponds to a p-value of less than 0.05.

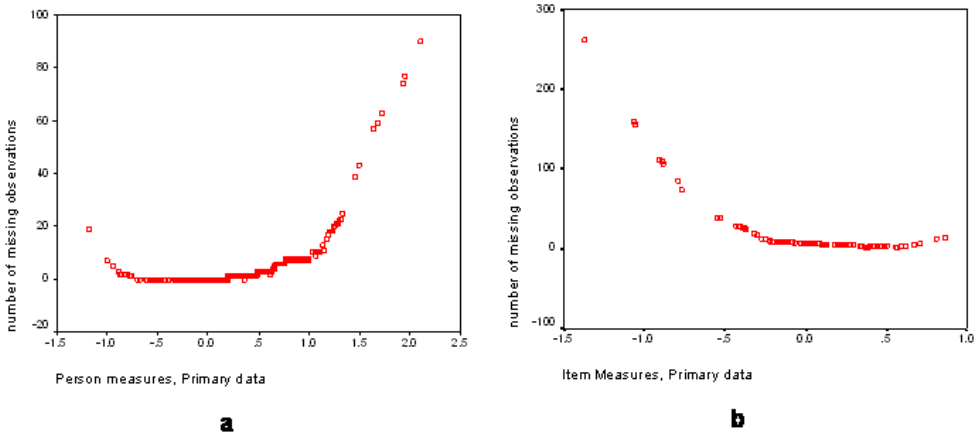

Fig. 2: Relation of person (a) and item (b) measures (before targeting) with number of missed observations (after targeting). All measures are in logit scale

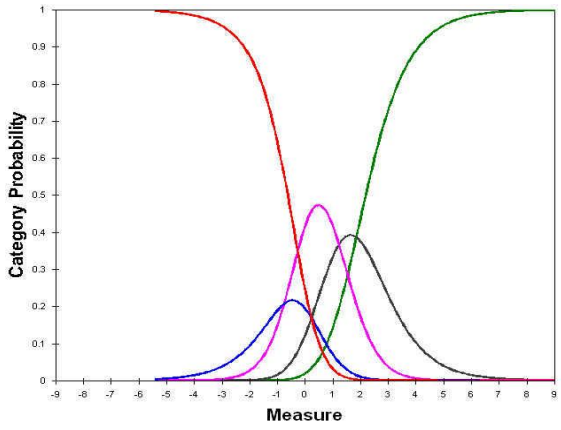

Fig. 3: Category probability curve of question No. 25 of WHOQOL-100 questionnaire; contrary to expectation, threshold No. 1 (intersection of the orange and blue lines) came after No. 2 (intersection of the blue and purple lines)

The fit index was not <0.5 for any individual. Elimination of persons was not significantly related to age, sex, QOL, and current health status variables, but significantly related to the level of education ($P = 0.003$). The proportions eliminated from primary school, secondary school, and college groups were 27%, 11%, and 13%, respectively.

Categories and Thresholds

In category and threshold analyses, 38 items, 10 due to category disordering, 24 due to threshold disordering, and 4 due to both, were deleted (Table 2). As an example, question number 25 (F15.2: How well are your sexual needs fulfilled?) is illustrated in Fig. 3. Threshold disordering is obvious in this graph.

Item Fit

In item fit analysis, 6 questions were eliminated due to being under-fit while none of the items were over-fitted (Table 2). Outfit and infit statistics in these items ranged from 1.55 to 1.85 and all were strongly significant (ZSTD ranged from 7.25 to 9.9).

DIF

In the DIF analysis, 4 items between males and females, 10 items between ill and healthy groups,

17 items between primary school subgroup and the total sample, and 6 items between age subgroups and the total sample showed bias or DIF. Some items in subgroups of two or more variables showed item bias; overall, 26 items were identified as inappropriate and were deleted in this stage (Table 2).

Table 2: All deleted items in this study considering analysis type and domain. Underlined cases are those with more than one reason for elimination

Analysis (in performed order)	D1 (F1-F3)	D2 (F4-F8)	D3 (F9-F12)	D4 (F13-F15)	D5 (F16-F23)	D6 (F24)	G	No of items deleted
				Facet.Item*				
Threshold / Category								**38**
Threshold	2.4, 3.2, 3.4	6.2, 7.2, 7.3, 7.4,	9.1, 11.1, 11.3, 11.4	13.1, 13.2, 13.4, 14.3, 14.4, 15.1,15.2, 15.3, 15.4	16.2, 17.1, 18.2, 19.2, 19.4, 23.1	-	G.1, G.2	
Category	1.2, 2.1	4.2, 5.3, 7.2,	10.1, 10.3, 11.1	15.2, 15.3	22.2, 22.3, 22.4	24.3	-	
Fit								**6**
Underfit	1.4	7.1	11.2	-	16.3, 18.4, 23.2	-	-	
Overfit	-	-	-	-	-	-	-	
DIF								**26**
Sex	1.1	6.1, 8.1	-	-	23.4	-	-	
Health status	1.1, 1.3, 2.2	6.1, 8.2	-	-	21.2	24.1, 24.2, 24.4	G.4	
Age	1.1, 2.2	8.3, 8.4	-	-		24.1, 24.4	-	
Education	1.1, 2.2	6.3,	9.3, 9.4, 10.2, 12.1, 12.2	14.1, 14.2	18.3, 19.1, 21.3, 23.3	24.2, 24.4	G.4	
No of items deleted (percent)	9 (75%)	13 (65%)	12 (75%)	11 (92%)	18 (56%)	4 (100%)	3 (75%)	**70** (70%)
No of Facets completely deleted	1	2	1	2	1	1	0	
No of Facets with more than 1 item remained	1	2	1	0	4**	0	0	

D: Domain; F: Facet; G: Overall QOL; DIF: Differential Item Functioning. / *First part is facet number and the second part is the item number in each facet; ** The facet 20 (Opportunities for acquiring new information and skills) was remained completely.

The Remaining Items

Table 3 shows the remaining items, in order of difficulty with mean and SD of 0 and 0.42 logit, respectively. The first 5 items in this list are all belong to the environmental domain. The mean difficulty was not significantly different between the remaining and the deleted items ($P = 0.922$).

Fig. 4 is the final item-person bar chart, which shows some off-targeting on both sides, more item reduction on the easier tail, a slight item redundancy, and absence of item gap.

Outfit or infit statistics were not more than 2 or less than 0.5 for any of the remaining items, and all item thresholds were ordered. Person separation index and test reliability values were 2.91 and 0.89, respectively.

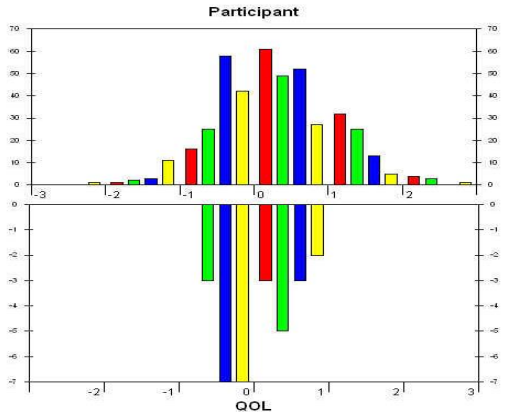

Fig. 4: Item-person plot in remained data. Upper part shows distribution of person measures and lower part shows item measures. Both are in logit scale

Table 3: Thirty remained items and their measures in logit scales

Item No	Facet.item*	Measure (logit)	Item description
50	21.1	0.93	To what extent do you have the opportunity for leisure activities?
47	18.1	0.76	Have you enough money to meet your needs?
49	20.2	0.55	To what extent do you have opportunities for acquiring the information that you feel you need?
74	20.3	0.54	How satisfied are you with your opportunities for acquiring new skills?
76	21.4	0.52	How satisfied are you with the way you spend your spare time?
75	20.4	0.43	How satisfied are you with your opportunities to learn new information?
8	4.1	0.41	How much do you enjoy life?
48	20.1	0.39	How available to you is the information that you need in your day-to-day life?
9	4.3	0.3	How positive do you feel about the future?
46	17.2	0.25	To what degree does the quality of your home meet your needs?
10	4.4	0.24	How much do you experience positive feelings in your life?
27	16.1	0.1	How safe do you feel in your daily life?
72	19.3	0.06	How satisfied are you with your access to health services?
58	5.2	-0.05	How satisfied are you with your ability to learn new information?
96	9.2	-0.12	How satisfied are you with your ability to move around?
36	22.1	-0.12	How healthy is your physical environment?
70	17.3	-0.12	How satisfied are you with the conditions of your living place?
31	17.4	-0.13	How much do you like it where you live?
20	10.4	-0.2	How much are you bothered by any limitations in performing everyday living activities?
59	5.4	-0.2	How satisfied are you with your ability to make decisions?
69	16.4	-0.29	How satisfied are you with your physical safety and security?
57	3.3	-0.33	How satisfied are you with your sleep?
83	3.1	-0.36	How well do you sleep?
84	5.1	-0.39	How would you rate your memory?
54	G.3	-0.41	In general, how satisfied are you with your life?
61	6.4	-0.42	How satisfied are you with your abilities?
91	12.4	-0.48	How satisfied are you with your capacity for work?
64	13.3	-0.54	How satisfied are you with your personal relationships?
56	2.3	-0.57	How satisfied are you with the energy that you have?
92	12.3	-0.75	How would you rate your ability to work?

*First part is facet number and the second part is the item number in each facet.

Discussion

In this study, using the Rasch analysis, 70 questions were excluded from a set of 100 questions in the Persian version of the WHOQOL-100 questionnaire due to category and/or threshold disordering, underfitting, or DIF in subgroups. The new questionnaire includes 30 items that, as the Rasch model claims, are expected to be unidimensional and measure a single construct. The Rasch model is not able to determine whether this construct is definitely QOL or not, and rather, this is usually left in the hands of the experts. Considering the extensive and worldwide research that has gone into constructing this questionnaire (3) and the great deal of expert opinions it has benefited from (29, 31), the new scale is also expected to be able to measure the QOL, like the original scale.

Targeting

In this study, first we used targeting analysis. Having on-target items not only improves the sample size efficiency, but also prevents a large decrease in the fitting of the items or persons because of a few unexpected behaviors. This effect is more pronounced with a small sample size. In our study, despite the large sample size and the relative stability of the measures, observations in which the measure difference between an individual-item

pair was greater than 1.75 logit were deleted. At the end of this analysis, the number of deleted observations was less than 2% (results are not reported), less than what is reported in studies conducted by Leplege (18) and Ecosse (19), which indicated a very favorable on-targeting.

Person Fit

In the next step, we looked for individuals who did not fit the model well, meaning that their observed responding patterns were largely different from the Guttman scale (43). Considering values from 0.5 to 1.5 as an accepted range for the fit statistics (38), we only omitted individuals with values more than 1.5 because being overfitted (fit statistic < 0.5) does not create a problem in the measurement process. We only had 67 underfitted persons, which indicate appropriate participation and response. Leplege (18) and Ecosse (19) used a more lenient criterion to detect unfit persons to prevent huge person deletion, but they had much more person elimination than we did. One reason could be employing well-trained interviewers in our setting to reduce misunderstanding and carelessness.

Categories and Thresholds

The most common problem of items in our study was disordered thresholds and/or categories. This is usually seen when there is a large number of categories, or when understanding questions and choices for responders is either difficult or different from the original concept (18). Following the Linacre's (39) recommendation, we paid attention to both, and observed that 10 items showed only category disordering. Based on our main objective, we deleted items that had category or threshold problems; however, in other cases, as has been done in some studies (44), correcting a question or its choices by combining two adjacent choices can be more appropriate.

Item Fit

Only 6 questions were deleted in item fit analysis, none of which showed over-fitting. It should be noted that because of deleting a large number of items in the previous steps, many of the unfit items had already been deleted. When compared to the study performed by Leplege (18), the total number of deleted items in the two steps of threshold and fitting analyses were almost similar: 44 questions in this study and 46 ones in the Leplege study.

DIF

We performed DIF analysis to detect item bias; for this reason, we defined an equivalence region based on which, items with bias were detected and deleted. This criterion is stricter than the usual method, but is robust to the effect of sample size on the results of such analyses. In this analysis, the items, which were evaluated, showed less bias in sex and age subgroups, but more bias in health status and especially in the level of education subgroups. The former is slightly incompatible with the specific objectivity principle but the latter can mostly be an indication of the fact that an individual's level of education has a large effect on his/her understanding of the meaning of the questions. Notably, all detected biased items in the education subgroups were due to the less educated group (primary school). This may indicate that it may be better to rephrase the questions in a way that the meaning of the questions is clear for persons with lower education as for others. This was also observed in person fit analysis in another way. In that case, deleted individuals only showed a significant difference with others in the level of education, so that a higher percentage of person deletion occurred among individuals with lower levels of education. Of course, at least to some extent, it could be due to various problems encountered during the translation process (29, 45) and may not be related to the original text of the questions.

The Remaining Items

Table 4 compares the remaining items from the WHOQOL-100 questionnaire in the current study with Leplege (18) and Ecosse (19) studies, and also the WHOQOL-BREF questionnaire. In the two mentioned studies, DIF analysis was conducted based on different cultures using databases from 4 (Argentina, France, Hong-Kong, and UK) and 6 (the 4 previous countries plus Spain and

USA) countries, respectively. The main purpose of these studies was to inspect the cultural equivalence of the desired questionnaire. In the Ecosse study, DIF analysis was not reported for sex and current health status. In the Leplege study, it was done, but item deletion was not performed based on it.

Table 4 shows that BREF, Ecosse, and Leplege had 11, 10 and 3 items in common with our items, respectively.

In the Ecosse's study, the most difficult questions (questions which showed the highest level of QOL) were about money and sex whereas in our study, they were related to environmental issues and money. On the other hand, the easiest questions in that study were about transportation and in our study about work capacity, energy and fatigue, and personal relationship facets. An important characteristic that these two studies had in common was that some facets were deleted completely whereas more than one question was spared in some of the other facets. This is somewhat expected considering the assumptions of the Rasch analysis; items within one facet are very similar and naturally behave in a similar manner. However, since all the remaining items fitted the model, based on the assumptions of the Rasch analysis (14), the remaining items were considered to be uni-dimensional, even though this assumption might not be completely correct.

In summary, considering the complexities of the Rasch model and our limited experience in performing such analyses, our study cannot be conclusive; therefore, the results should be used conservatively. On the other hand, large sample size and therefore the stability of our measurements is one of the advantages of this study.

Table 4: Comparison of remaining items in the current study with WHOQOL-BREF questionnaire and two similar studies

No.	Present Study		Leplege*	Ecosse		BREF
	Facet.item**	Measure (logit)	Facet.item**	Facet.item**	Measure (logit)	Facet.item**
1	**21.1**	0.93	**G.4**	**18.1**	0.54	G.1
2	**18.1**	0.76	18.3	**15.3**	0.52	G.4
3	20.2	0.55	22.3	15.1	0.48	1.4
4	20.3	0.54	**5.3**	22.3	0.29	11.3
5	21.4	0.52	**24.2**	**3.3**	0.24	4.1
6	20.4	0.43	20.3	**22.1**	0.24	24.2
7	**4.1**	0.41	**20.1**	14.2	0.21	5.3
8	**20.1**	0.39	23.1	20.3	0.2	16.1
9	4.3	0.3	1.2	14.1	0.18	22.1
10	17.2	0.25	**13.3**	**24.2**	0.04	2.1
11	4.4	0.24	**8.1**	5.4	0.02	7.1
12	**16.1**	0.1	**23.3**	17.2	0	18.1
13	**19.3**	0.06		12.1	-0.03	20.1
14	5.2	-0.05		20.4	-0.04	21.1
15	9.2	-0.12		**20.1**	-0.09	9.1
16	**22.1**	-0.12		**1.4**	-0.09	3.3
17	**17.3**	-0.12		9.2	-0.15	10.3
18	17.4	-0.13		8.4	-0.16	12.4
19	10.4	-0.2		**23.3**	-0.16	6.3
20	5.4	-0.2		**14.4**	-0.17	13.3
21	16.4	-0.29		**17.3**	-0.27	15.3
22	**3.3**	-0.33		13.1	-0.33	14.4
23	3.1	-0.36		**9.1**	-0.39	17.3
24	5.1	-0.39		23.4	-0.54	19.3
25	G.3	-0.41		23.2	-0.55	23.3
26	6.4	-0.42				8.1
27	**12.4**	-0.48				
28	**13.3**	-0.54				
29	2.3	-0.57				
30	12.3	-0.75				

*In Leplege study, the item measures were not reported and only ordering of them is available./ **First part is facet number and the second part is the item number in each facet./ Undelined **bold** numbers are items that are also present in the WHOQOL-BREF.

Conclusion

This paper should be considered positional and the resulted new scale should be considered as a complement to the present similar questionnaires like the WHOQOL-BREF. Considering the results of this study and the unique features of the Rasch model, it is suggested that first, this analysis be used in different populations and settings, and for different types of health related questionnaires; and second, in developing or reviewing a scale, wording of the items and also the number of choices should be noticed and revised accordingly.

Abbreviations

WHO, World Health Organization; QOL, Quality of Life; SD, Standard Deviation; HRQOL, Health-Related Quality of Life; CTT, Classical Test Theory; IRT, Item Response Theory; DIF, Differential Item Functioning; RSA, Rating Scale Analysis; PCM, Partial Credit Model; ZSTD, Standardized Z score.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This study was part of a PhD thesis supported by the School of Public Health of Tehran University of Medical Sciences (Thesis registration No. 240/5258). The data was provided by University of Social Welfare and Rehabilitation Sciences. The authors would like to thank Dr. Ali Seddighzadeh, Mrs. Homa Kashani, and Dr. Hamidreza Ghadimi for revising the manuscript. The authors declare that there is no conflict of interests.

References

1. Sanders C, Egger M, Donovan J, Tallon D, Frankel S (1998). Reporting on quality of life in randomised controlled trials: bibliographic study. *BMJ,* 317 (7167): 1191-94.
2. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R (2002). Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ,* 324 (7351): 1417-21.
3. Hubanks L, Kuyken W (WHOQOL group) (1994). *Quality of Life Assessment: An Annotated Bibliography*. WHO, Geneva, p.: 1.
4. Kuyken W, Orley J. (1994). Development of the WHOQOL: Rationale and current status. *Int J Ment Health,* 23 (3): 24-56.
5. Salkind NJ, Rasmussen K, Eds (2007). *Encyclopedia of Measurement and Statistics*. 1st ed. SAGE Publications, Thousand Oaks, pp.: 140-143.
6. Streiner DL, Norman GR, Eds (2003). *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. Oxford University Press, New York, pp.: 213-227.
7. Wright BD (1999). Common sense for measurement. *Rasch Meas Trans,* 13: 704-5.
8. Svensson E (2001). Guidelines to statistical evaluation of data from rating scales and questionnaires. *J Rehab Med,* 33 (1): 47-48.
9. Embretson SE, Reise SP (2000). *Item response theory for psychologists*. 1st ed. Lawrence Erlbaum Associates, New Jersey, pp.: 3-10.
10. Rasch G (1960 (Reprinted 1993)). *Probabilistic models for some intelligence and attainment tests*. 1st ed. MESA Press, Chicago, pp.: 1-199.
11. Andrich D (1988). *Rasch models for measurement*. 1st ed. Sage Publications, Newbury Park, California, pp.: 1-95.
12. Wright BD, Stone MH (1979). *Best Test Design*. 1st ed. MESA Press, Chicago, pp.: 1-240.
13. Luce RD, Tukey JW (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *J Math Psychol,* 1 (1): 1-27.
14. Tennant A, McKenna SP, Hagell P (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value Health,* 7 (Suppl 1): S22-26.
15. Wright BD, Masters GN (1982). *Rating Scale Analysis*. 1st ed. MESA Press, Chicago, pp.: 1-206.

16. Bond TG, Fox CM (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates, New Jersey, pp.: 29-260.
17. Tennant A, Conaghan PG (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum,* 57 (8): 1358-62.
18. Leplege A, Ecosse E, WHOQOL Rasch Project Scientific Committee (2000). Methodological issues in using the Rasch model to select cross culturally equivalent items in order to develop a Quality of Life index: the analysis of four WHOQOL-100 data sets (Argentina, France, Hong Kong, United Kingdom). *J Appl Meas,* 1 (4): 372-92.
19. Ecosse E, Leplège A, The WHOQOL Rasch Group (2005). Rasch model isolates quality of life construct in six whoqol-100 data sets (Argentina, France, Hong-Kong, Spain, USA, and UK). In: *Rasch Measurement in Health Sciences.* Bezruczko N, ed. 1st ed. JAM Press, Maple Grove, Minnesota, pp.: 411-431.
20. Smith AB, Wright P, Selby PJ, Velikova G (2007). A Rasch and factor analysis of the Functional Assessment of Cancer Therapy-General (FACT-G). *Health Qual Life Outcomes,* 5 (1): 19.
21. Rocha NS, Fleck MP (2009). Validity of the Brazilian version of WHOQOL-BREF in depressed patients using Rasch modelling. *Rev Saúde Pública,* 43 (1): 147-53.
22. Chachamovich E, Fleck MP, Trentini C, Power M (2008). Brazilian WHOQOL-OLD Module version: a Rasch analysis of a new instrument. *Rev Saúde Pública,* 42 (2): 308-16.
23. Prieto L, Alonso J, Lamarca R (2003). Classical Test Theory versus Rasch analysis for quality of life questionnaire reduction. *Health Qual Life Outcomes,* 1 (1): 27.
24. Noerholm V, Groenvold M, Watt T, Bjorner JB, Rasmussen NA, Bech P (2004). 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,* 13 (2): 531-40.
25. Kook SH, Varni JW (2008). Validation of the Korean version of the Pediatric Quality of Life Inventory 4.0 (PedsQL) generic core scales in school children and adolescents using the Rasch model. *Health Qual Life Outcomes,* 6 (1): 41.
26. Liang WM, Chang CH, Yeh YC, Shy HY, Chen HW, Lin MR (2009). Psychometric evaluation of the WHOQOL-BREF in community-dwelling older people in Taiwan using Rasch analysis. *Qual Life Res,* 18 (5): 605-18.
27. Fayers PM, Machin D (2007). *Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes*. 2nd ed. John Wiley & Sons Ltd, West Sussex, pp.: 477-508.
28. McDowell I (2006). *Measuring Health: A Guide to Rating Scales and Questionnaires*. 3rd ed. Oxford University Press, New York, pp.: 520-703.
29. Anonymous (The WHOQOL Group) (1995). The World Health Organization quality of life assessment (WHOQOL): position paper from the World Health Organization. *Soc sci med,* 41 (10): 1403-09.
30. Anonymous (The WHOQOL Group) (1998). The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. *Soc sci med,* 46 (12): 1569-85.
31. Anonymous (The WHOQOL Group) (1993). Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). *Qual Life Res,* 2 (2): 153-59.
32. Power M, Harper A, Bullinger M (1999). The World Health Organization WHOQOL-100: tests of the universality of Quality of Life in 15 different cultural groups worldwide. *Health psychol,* 18 (5): 495-505.
33. Skevington SM, Lotfy M, O'Connell KA, WHOQOL Group (2004). 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,* 13 (2): 299-310.
34. Nedjat S, Montazeri A, Holakouie K, Mohammad K, Majdzadeh R (2008). Psychometric properties of the Iranian interview-administered version of the World Health Organization's Quality of Life Questionnaire (WHOQOL-BREF): A population-based study. *BMC Health Serv Res,* 8 (1): 61.
35. Karimlou M, Zayeri F, Salehi M (2011). Psychometric properties of the Persian version

of the World Health Organization's quality of life questionnaire (WHOQOL-100). *Arch Iran Med,* 14 (4): 281-87.
36. Anonymous (1998). *WHOQOL user's manual (WHO/MNH/MHP/98.4.Rev.1).* World Health Organization, Geneva, pp.: 26-34.
37. Linacre JM (2000). Comparing "partial credit" and "rating scale" models. *Rasch Meas Trans,* 14 (3): 768.
38. Linacre JM (2009). *A User's Guide to WINSTEPS Rasch measurement computer program.* Chicago. Available from: www.winsteps.com.
39. Linacre JM (1999). Category Disordering vs. Step (Threshold) Disordering. *Rasch Meas Trans,* 13 (1): 675-8.
40. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW (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. *Qual Life Res,* 12 (5): 485-501.
41. Wang WC, Yao G, Tsai YJ, Wang JD, Hsieh CL (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Qual Life Res,* 15 (4): 607-20.
42. Linacre JM (1989). Rank ordering and Rasch measurement. *Rasch Meas Trans,* 2 (4): 41-42.
43. Guttman LA (1950). The basis for Scalogram analysis. In: *Studies in social psychology in World War II.* Stouffer SA, Guttman LA, Suchman FA, Lazarsfeld PF, Star SA, Clausen JA, Eds. Princeton University, Princeton, pp.: 60-90.
44. Pallant JF, Tennant A (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol,* 46 (1): 1-18.
45. Bowden A, Fox-Rushby JA (2003). A systematic and critical review of the process of translation and adaptation of generic health-related quality of life measures in Africa, Asia, Eastern Europe, the Middle East, South America. *Soc sci Med,* 57 (7): 1289-306.