Original Article

The Iranian SF-12 Health Survey Version 2 (SF-12v2): Factorial and Convergent Validity, Internal Consistency and Test-retest in a Healthy Sample

Camelia Rohani, PhD. Student
Karolinska Institutet, NVS Development, Nursing Devision, Stockholm, Sweden

Heidari Ali Abedi, PhD., Associate Professor
Islamic Azad University, Khorasgan Branch, Nursing & Medwifery Faculty, Isfahan, Iran

Ann Langius – Eklöf, PhD., Professor
Örebro University, School of Health and Medical Sciences, Örebro, Sweden

Objectives: Within the context of Health-Related Quality of Life studies, it was necessary to translate and evaluate the psychometric properties of the SF-12 Health Survey version 2. The aim of this study was to investigate the factorial structure, convergent validity and reliability of this instrument in a healthy Iranian sample, following translation and establishment of content and face validity.

Method: In this cross-sectional study the translated instrument was administered together with the Sense of Coherence Scale and Health Index in a convenience sample of healthy people (n = 289) aged 17 to 76 years old between June and September 2006. A test-retest was conducted one month later.

Results: Content Validity Index for Scale (85.6%) and face validity of the instrument were acceptable. The results of Exploratory Factor Analysis and Structural Equation Modeling by four models verified the existing two-factor structure, a physical and a mental component summary. All models exceeded the goodness of fit indices and showed a resemblance with the original instrument, except for models 3 and 4 (allowing cross-loadings). Although in these two models the General Health item and scale was loaded to the mental component rather than the physical component, construct validity of the instrument was confirmed. Also, physical and mental component summaries were significantly ($p < 0.001$) correlated to the Sense of Coherence Scale ($r = 0.27$, $r = 0.68$) and Health Index ($r = 0.49$, $r = 0.67$). Cronbach's alpha values and the intra-class correlation coefficients were ≥ 0.70 and ≥ 0.60, respectively.

Conclusion: The Iranian SF-12 Health Survey version 2 was a psychometrically sound instrument, implying that it is suitable for use with large-scale surveys in Iranian population, both in clinical and rehabilitation settings or at a public level.

Keywords: SF-12 Health Survey version 2, health-related quality of life, instrument translation, psychometric tests, exploratory factor analysis, structural equation modeling, Sense of Coherence Scale, Health Index

Introduction

The concept of Health-Related Quality of Life (HRQoL) is regarded as a sensitive outcome variable in health outcome measurement studies (1, 2). HRQoL is a multidimensional concept that refers to function and well-being on various dimensions of health, including physical, emotional, social and spiritual aspects of life (3, 4). To cover the different aspects of HRQoL in research studies, the use of multiple instruments is required (2, 5). Applying a generic instrument in combination with a disease-specific questionnaire can be useful (6). The Short Form 12-item Health Survey (SF-12) is a brief, generic, well-tested instrument used worldwide. It is derived from the SF-36 that is also widely accepted and which was developed to assess subjective physical and mental health status in large surveys, as well as longitudinal studies in both general and clinical populations (7, 8). Ware and co-workers (9) released a new revised version with several improvements, the Short Form 12-item Health Survey version 2 (SF-12v2). The SF-12v2 differs

* All correspondences to: Camelia Rohani, E-mail: camelia.rohani@ki.se

from the original version of the SF-12 (SF-12 version 1/ SF-12v1) with respect to instructions, layout of questions and answers, response alternatives, and a larger font size throughout the instrument. Also, the SF-12v2 is more international due to improvements in item wording following the process of translation and adaptation of the SF-36 and the SF-12 in other countries. Recently, a study on factor structure and internal consistency of the Iranian version of the SF-12v1 when used in a general population showed promising results (10). To our knowledge, the present study is the first study on the Iranian version of the SF-12v2. The purpose of this study was to investigate factorial structure, convergent validity and reliability (in the form of internal consistency, stability and robustness) of the SF-12v2 in a healthy Iranian sample, following translation and establishment of content and face validity.

Methods

The study is cross-sectional with a baseline (T1) and a one month test-retest (T2).

Participants and data collection

Sample for face validity

Twenty voluntary participants including 10 healthy persons drawn from one urban health center and 10 breast cancer patients from one educational hospital were recruited in the study.

Main sample

A convenience sample of 289 healthy respondents from nine urban health centers (n=210), one university (n=30), and a private company (n=49) in Tehran participated in this study between June and September 2006. They were recruited with a letter of invitation that was posted on a wall directed to specific personnel at each location, with the inclusion criteria listed in the letter (see below). At the health centers, the letter was addressed to the health care providers, clerical personnel and community volunteers who were trained to deliver health care to their neighbors. At the university, clerical employees, and at the private company, clerical and technical employees were targeted. Subjects met general inclusion criteria if they were free from all chronic conditions, were of Iranian nationality and at least 18 years old and were able to read and write the Persian language. The sampling was performed in various settings to allow easy access to healthy people with different levels of education and socio-economic classes with mixed gender. From a total of 289 healthy participants, 252 persons (87.2%) answered the instruments at baseline (T1) and 203 persons (70.2%) at the one-month follow-up (T2). Therefore, the final sample consisted of the 203 healthy persons who participated at both T1 and T2. In addition, voluntariness and confidentiality were emphasized and interested participants provided their written informed consent before participating.

Ethical permission to conduct the study was obtained from the Ministry of Health and Medical Education of Iran (P/391-31, July 2005), and the Isfahan University of Medical Sciences.

Translation procedure:

The translation process included two forward translations from English to Persian and two blind back-translations were conducted based on the standard guidelines (11, 12). All versions of the translated SF-12v2 were reconciled by the authors.

Instruments

Except for the SF-12v2, two additional instruments, the Sense of Coherence (SOC) Scale and the Health Index (HI), were used for evaluation of convergent validity. All questionnaires were self-administered.

Short Form 12-Item Health Survey version 2 (SF-12v2)

The SF-12v2 is a multi-purpose Short Form (SF) generic measure of health status that uses a Likert scale format (9). In the present study, the standard four-week recall period version was used. Validity and reliability of the SF-12v2 has been demonstrated in several studies (9, 13-15). The SF-12v2 is comprised of a 12-item subset of the SF-36 version 2 (SF-36v2) categorized in eight domains: Bodily Pain (BP), General Health (GH), Vitality (VT), and Social Functioning (SF) with one item each. In addition, Physical Functioning (PF), Mental Health (MH), Role Physical (RP), and Role Emotional (RE) domains are represented with two items each (7, 9). Based on a theoretical test of the original model, the PF, RP, BP, and GH scales yield a Physical Component Summary (PCS) measure, and the MH, RE, VT, and SF scales reveal a Mental Component Summary (MCS) measure. These scales need to show maximum loading on the respective component (7, 9, 16). Theoretically, cross-loadings are not supposed to occur, but they have been found with regard to the GH, VT, and SF scales in some studies (17-19). All twelve items are used to

calculate both PCS and MCS measures scores by applying scoring algorithms with weighted item responses. Calculation of scores for the eight scales is performed using the transformed scores (range: 0-100) and summary measures are standardized to produce mean of 50 with a standard deviation of 10 for the United States (US) population (norm-based scoring); the higher the score, the better the perceived health (7, 9). When using the US standard scores, results can easily be compared across various countries and settings. Also, comparisons within one country can be carried out by parallel analyses or country specific scoring. An advantage of the SF-12v2 is the availability of more up-to-date norms from the general US population compared to the version 1 (9). A License for using the SF-12v2 was acquired from QualityMetric Incorporated (# 25762, May 2006).

Sense of Coherence Scale (SOC-13)

The Sense of Coherence scale is an orientation to life instrument which measures the sense of coherence concept. This concept is defined as an individual's global view of life based on how comprehensible, manageable and meaningful life appears to him/her. The scale has a semantic-differential format ranging from one to seven points with two anchoring responses. The scoring range is 13-91; the higher the score, the stronger the sense of coherence (20). The SOC-13 has been translated and tested in an Iranian sample (21).

Health Index (HI)

The HI has been developed and tested in Sweden and measures general well-being with nine items: energy, temper, fatigue, loneliness, sleep, dizziness, bowel function, pain and mobility (22, 23). Each item has a verbal category scale format ranging from 1 to 4. The items are summarized to a total index with a possible total score ranging from 9 to 36. The higher the score, the better the perceived general health (22). The HI has been translated and tested in an Iranian sample (21).

Data analysis

All statistical analyses were conducted using the Statistical Package for Social Science (SPSS) version 16 and Lisrel 8.80 (24). In Lisrel analyses, the Maximum Likelihood Estimation (MLE) was applied as the most commonly used estimation method (25). The twelve items, eight scales and two summary measures of the SF-12v2 were evaluated by the P-P plots and the normality assumption of the variables was not violated. Floor and ceiling effects were evident if more than 15% of the respondents rated the lowest or highest possible score, respectively (26).

Content and face validity

Content Validity Index for Scale (S-CVI) was determined by an expert panel consisting of teachers (n = 10) from Isfahan University of Medical Sciences Nursing and Midwifery Faculty. They evaluated relevance, clarity and simplicity of the individual items using CVI assessment form on a four point scale. A scale-level CVI score of 0.80 or higher indicates a good content validity (27). The final version of the SF-12v2 was pre-tested for face validity with 20 voluntary subjects.

Construct validity

Construct validity of the instrument was assessed through factorial validity and convergent validity, according to the literature (28).

Factorial validity

Factorial validity of the SF-12v2 was estimated by both Exploratory Factor Analysis (EFA) and Structural Equation Modeling (SEM) at T1. Thus, EFA and SEM were conducted on the twelve items and eight scales of the SF-12v2 to test each factor structure, respectively. Based on the original SF-12 conceptual model (7) and studies on items/scales cross-loadings (19), we expected that a model with the following characteristics would fit with the SF-12v2 data: (1) the model would include a two-factor structure, consisting of the PCS and MCS measures (2) loadings of the items and scales on the factors would be similar to the original version (PCS: GH, PF, RP, BP and MCS: RE, MH, VT, SF), and (3) cross-loadings of the GH, VT, and SF items and scales might be shown. Cross-loadings were considered substantial if they were greater than 0.40 (19). EFA was run using Principal Component Analysis (PCA) with varimax and oblique rotations. For more specific testing of the configuration of the factor structure, a Confirmatory Factor Analysis (CFA) by SEM was performed with four models. According to the theoretical model of the SF-12 (7), model 1 included two latent inter-correlated factors (PCS and MCS), and each factor correlated to the six specific items, respectively. Model 2 consisted of two latent inter-correlated factors and each factor correlated with the four specific scales, respectively (PCS: GH, PF, RP, BP and MCS: RE, MH, VT, SF). For assessment of cross-loadings of the items and

scales, models 3 and 4 were created in the same way as models 1 and 2 with cross-loadings. Sometimes, produced models by CFA make some items/scales to have loading of zero on the opposite component, which prevents inconsistency (29). Because of running the Lisrel program and fixing the models with corresponding factor loadings, two paths were excluded in models 3 and 4, respectively (in model 3, two paths from the PCS measure to the MH1 and from the MCS measure to the PF1; and in model 4 two paths from the PCS measure to the MH scale and from the MCS measure to the PF scale).

Some specific indices and cutoff points were selected for CFA by SEM analyses. The indices were Chi-square to the degrees of freedom ratio (criteria: ratio < 6), Comparative Fit Index (CFI) (criteria: > 0.90), Standardized Root Mean Square Residual (SRMR) (criteria: < 0.08), Non-Normed Fit Index (NNFI) or Tucker-Lewis Index (criteria: > 0.90), and Incremental Fit Index (IFI) or BL89 (criteria: > 0.90). Also, improvements in the models fit were evaluated by a decrease in Akaike Information Criterion (AIC) and Expected Cross-Validation Index (ECVI) (30-33).

Convergent validity

In support of convergent validity, correlations between PCS and MCS measures scores with SOC and HI scores were estimated by Pearson product moment correlation coefficient. Based on the results of earlier studies (34-37), slight to moderate positive associations between the concepts of sense of coherence and well-being with the PCS and MCS were hypothesized. A correlation below 0.20 was considered low, between 0.20-0.35 slight, 0.36-0.65 moderate, 0.66-0.85 high, 0.86 and above was considered very high (38).

Reliability

For reliability, internal consistency and stability were assessed. Internal consistency was measured by Cronbach's alpha coefficient. An expected Cronbach's alpha coefficient equal to or greater than 0.70 would be considered satisfactory (39). Stability was assessed by intra-class correlation coefficient (ICC). Most QoL instruments fail to attain a demanding level for ICC, so some authors suggest that values around 0.60 and above are reasonable (5). Furthermore, statistical changes in the means of two summary scores and eight scales of the SF-12v2 were estimated by a paired t-test between T1 and T2.

Results

Descriptive data

Socio-demographic characteristics of the sample are summarized in Table 1. The age of the study sample ranged from 17 to 76 years old (mean = 37.90, SD = 11.7). Floor and ceiling effects, transformed and standardized norm-based means values are shown in Tables 2 and 3, respectively.

Table 1. Socio-demographic characteristics of the healthy Iranian sample (n=252).

Variables	n (%)
Gender	
Female	182 (72)
Male	70 (28)
Marital status	
Single	53 (21)
Married	186 (74)
Divorced	7 (3)
Widowed	6 (2)
Job status	
Housewife	95 (38)
Employed	131 (52)
Student	19 (7)
Retired	7 (3)
Educational level	
Primary school	20 (8)
Secondary school	29 (12)
High school	23 (9)
Diploma	71 (28)
University	109 (43)

Content and face validity

The S-CVI score was 85.6%. An evaluation of the linguistic appropriateness of the SF-12v2 items by 20 voluntary participants rendered small changes in the wording of some items for more clarification.

Construct validity

Factorial structure by Exploratory Factor Analysis

The EFA results with PCA by both varimax and oblique rotations including twelve items and eight scales explained a two-factor conceptual structure, namely the physical and

Table 2. The SF-12v2 items, floor and ceiling effects in the healthy Iranian sample at baseline (n=252) and one month later (n=203).

SF-12v2 items	Response frequencies at baseline (%)					Response frequencies at one month later (%)				
	Floor (%)				Ceiling (%)	Floor (%)				Ceiling (%)
	1	2	3	4	5	1	2	3	4	5
GH[1]	1.2	24.2	45.6	20.6	8.3	1.5	24.1	48.8	17.7	7.9
PF1[2]	3.2	31.3	65.5	NA*	NA*	4.9	28.6	66.5	NA*	NA*
PF2	4.0	32.1	63.9	NA*	NA*	4.4	35.5	60.1	NA*	NA*
RP1[3]	4.0	9.5	30.2	31.0	25.4	2.0	8.4	39.9	29.1	20.7
RP2	3.2	8.3	25.0	29.8	33.7	1.5	6.4	31.0	36.0	25.1
BP[4]	1.6	5.2	23.0	32.1	38.1	2.0	5.4	20.2	27.6	44.8
RE1[5]	3.2	14.7	34.5	28.6	19.0	1.0	10.8	34.5	24.1	29.6
RE2	1.2	12.3	34.1	31.7	20.6	2.5	9.9	31.0	35.0	21.7
MH1[6]	4.4	11.1	34.9	37.7	11.9	3.0	10.8	31.0	40.4	14.8
MH2	3.2	15.9	33.3	27.8	19.8	3.4	13.3	26.6	32.5	24.1
VT[7]	2.4	17.1	34.9	36.5	9.1	2.0	16.7	37.4	34.0	9.9
SF[8]	4.0	15.1	27.8	22.2	31.0	3.0	15.3	27.1	21.2	33.5

[1] General Health, [2] Physical Functioning, [3] Role Physical, [4] Bodily Pain, [5] Role Emotional,
[6] Mental Health, [7] Vitality, [8] Social Functioning.
* NA: Not applicable due to limitation of these items to 3 alternatives.

mental component summaries (Table 4). The PCS and MCS measures together with twelve items and eight scales explained 59.3% and 64.0% of the total variance, respectively. The results of EFA with varimax rotation showed that the PF, RP, and BP items and scales were more highly loaded on the physical component and the RE, MH, VT, and SF items and scales were more highly loaded on the mental component. Further, the GH item and scale was loaded higher on the mental component and a cross-loading was observed in the RP items and scale. The oblique rotation showed that the GH item and scale was more loaded on the mental than the physical component as well, and no cross-loading was observed.

Table 3. Mean (SD), Internal consistency and stability of the SF-12v2 at baseline (Time 1) and one month later (Time 2) (n = 203).

SF-12v2 Scales/ Summary Measures	Mean (SD) [1] (Time 1)	Mean (SD) [1] (Time 2)	p Value[2]	Cronbach's Alpha (Time 1)	Cronbach's Alpha (Time 2)	ICC[3] (95% CI) [4]
GH[5]	59.3 (24.5)	58.2 (24.2)	0.424	NA*	NA*	0.80 (0.74-0.85)
PF[6]	81.0 (25.6)	79.3 (26.8)	0.394	0.83	0.83	0.61 (0.50-0.67)
RP[7]	69.3 (25.3)	66.9 (22.7)	0.176	0.87	0.88	0.62 (0.50-0.71)
BP[8]	75.4 (24.7)	75.4 (24.7)	0.377	NA	NA	0.64 (0.53-0.73)
RE[9]	63.4 (24.0)	66.7 (23.5)	**0.046**	0.88	0.81	0.65 (0.54-0.73)
MH[10]	60.5 (21.8)	64.2 (22.8)	**0.007**	0.61	0.71	0.76 (0.68-0.82)
VT[11]	58.1 (24.0)	58.2 (23.4)	0.943	NA*	NA*	0.64 (0.53-0.73)
SF[12]	65.9 (29.6)	66.7 (29.4)	0.681	NA*	NA*	0.66 (0.55-0.74)
PCS[13]	49.4 (8.1)	48.2 (8.2)	**0.048**	0.82	0.84	0.64 (0.53-0.73)
MCS[14]	42.4 (10.5)	44.2 (10.8)	**0.005**	0.89	0.89	0.77 (0.70-0.83)

[1] Transformed mean was estimated for the eight SF-12v2 scales and normed-based mean was used for the SF-12v2 summary measures (PCS & MCS).
[2] Student's paired t-test, [3] Intra-class Correlation Coefficient, [4] CI: Confidence Interval.
[5] General Health, [6] Physical Functioning, [7] Role Physical, [8] Bodily Pain, [9] Role Emotional,
[10] Mental Health, [11] Vitality, [12] Social functioning, [13] Physical Component Summary, [14] Mental Component Summary.
* NA: Not applicable due to limitation of them to one item.

Factorial structure by SEM analysis

SEM analyses with four models of the Iranian version of the SF-12v2 were conducted to confirm the EFA results. Table 5 summarizes the results of goodness of fit indices for them. SEM results showed that all models exceeded the sensitivity criteria of the goodness of fit. In addition, the factor loadings in these models showed that all variables were more highly loaded on the respective components, excluding model 3 (two-factor with twelve items and cross-loading between them) and model 4 (two-factor with eight scales and cross-loading between them) which demonstrated that the GH item and scale was more loaded on the mental component. Furthermore, the results of the items and scales loadings illustrated that all the regression coefficients in the four models had moderate to strong associations with the respective component, except for the GH item and scale in models 3 and 4, which was weakly Loaded on the physical component and moderately on the mental component. The variance explained by the four models of the Iranian version of the SF-12v2 was 0.85, 0.94, 0.87 and 0.98, respectively.

Also, correlation between the PCS and MCS measures was positive and significant in the four models with regression coefficients of 0.71, 0.79, 0.34, and 0.43, correspondingly. Models 2 and 4 showed the best fit according to AIC and ECVI results.

Table 4. The results of Exploratory Factor Analysis of the SF-12v2 with 12 items and 8 scales using two rotation methods in the healthy Iranian sample (n =252).

	Varimax		Oblique			Varimax		Oblique	
SF-12v2 items	**Factor I (PCS)**	**Factor II (MCS)**	**Factor I (PCS)**	**Factor II (MCS)**	**SF-12v2 scales**	**Factor I (PCS)**	**Factor II (MCS)**	**Factor I (PCS)**	**Factor II (MCS)**
GH[1]	0.29	**0.59**	0.13	**0.58**	GH[1]	0.31	**0.62**	0.14	**0.61**
PF1[2]	**0.83**	0.08	**0.92**	-0.19	PF[2]	**0.84**	0.10	**0.94**	-0.18
PF2	**0.79**	0.06	**0.88**	-0.19	RP[3]	**0.73**	0.43	**0.69**	0.24
RP1[3]	**0.70**	0.43	**0.66**	0.25	BP[4]	**0.76**	0.29	**0.77**	0.07
RP2	**0.70**	0.41	**0.66**	0.23	RE[5]	0.30	**0.72**	0.09	**0.72**
BP[4]	**0.65**	0.35	**0.62**	0.18	MH[6]	0.12	**0.85**	-0.16	**0.94**
RE1[5]	0.29	**0.73**	0.09	**0.73**	VT[7]	0.17	**0.76**	-0.06	**0.81**
RE2	0.22	**0.74**	0.01	**0.78**	SF[8]	0.30	**0.70**	0.11	**0.70**
MH1[6]	0.03	**0.74**	-0.21	**0.83**					
MH2	0.13	**0.79**	-0.11	**0.85**					
VT[7]	0.23	**0.68**	0.04	**0.70**					
SF[8]	0.31	**0.67**	0.12	**0.67**					

Notes: Strong association ($r \geq 0.66$), Moderate association ($0.36 \leq r \leq 0.65$), and Weak association ($r \leq 0.35$).
[1] General Health, [2] Physical Functioning, [3] Role Physical, [4] Bodily Pain, [5] Role Emotional,
[6] Mental Health, [7] Vitality, [8] Social Functioning.

Convergent validity

The PCS and MCS measures scores were significantly ($p < 0.001$) correlated to SOC scores ($r = 0.27$, $r = 0.68$) and HI scores ($r = 0.49$, $r = 0.67$), respectively.

Reliability

Cronbach's alpha coefficients, test-retest (ICC) and paired t-test results of the SF-12v2 are presented in Table 3.

Discussion

Based on the S-CVI score and the judgment of the 20 participants involved in the linguistic reformation of the translated SF-12v2, content and face validity of the Iranian version of the SF-12v2 was supported. There was no floor effect at baseline or one month later. However, a rather high ceiling effect for some items appeared and was most probably related to the healthy sample characteristics, which did not pose a threat to the instrument validity. Ware *et al.* (9) found that in spite of the changes in the SF-12v1, a

Table 5. Goodness of fit indices in four models of the Iranian SF-12v2 by SEM analyses in the healthy Iranian sample (n = 252).

Model	Facture Structure	χ^2/df ; ratio	CFI[1]	SRMR[2]	NNFI[3]	IFI[4]	AIC[5]	ECVI[6]
1	2 latent variables with 12 items	256.79/53; 4.84	0.93	0.078	0.91	0.93	315.79	1.26
2	2 latent variables with 8 scales	64.98/19; 3.42	0.97	0.055	0.95	0.97	98.98	0.39
3	2 latent variables with 12 items & cross-loadings	222.14/43; 5.17	0.94	0.053	0.91	0.94	292.14	1.16
4	2 latent variables with 8 scales & cross-loadings	25.74/13; 1.98	0.99	0.028	0.98	0.99	71.74	0.29

[1] CFI: Comparative Fit Index, [2] SRMR: Standardized Root Mean Square Residual, [3] NNFI: Non-Normed Fit Index, [4] IFI: Incremental Fit Index, [5] AIC: Akaike Information Criterion, [6] ECVI: Expected Cross-Validation Index.

small ceiling effect still remained in the SF-12v2 within the general population. Nevertheless, the SF-12v2 in the present study captured the full range of response alternatives in all items. It supports sensitivity and responsiveness of the instrument (5).

The results of EFA and SEM analyses confirmed the existing two-factor structure of the physical component and mental component summary measures respectively, consistent with US and European studies (7, 40-42). Hence, construct validity is supported in the Iranian version of the SF-12v2. By EFA analysis, these two factors, which underlie twelve items or eight scales, commonly accounted for 59.3% and 64.0% of the variance respectively, similar to the results of the Iranian SF-36 (43) and SF-12v1 (10). No cross-loadings were observed in the GH, VT, and SF items and scales, neither with EFA nor with SEM. The original SF-12 conceptual model (7) suggests no cross-loadings between the items and scales, even though cross-loadings in the GH, VT, and SF items and scales have been reported in some studies (18, 19). The RP items and scales were cross-loaded to both the physical and mental components by the varimax rotation, but disappeared in the oblique rotation. It must be noted that oblique rotation is suggested to be the most optimal way to perform factor analysis. Oblique rotation permits correlation between factors and thereby provides more useful information than varimax rotation (44). Furthermore, with oblique rotation, all items and scales were loaded on the respective component as expected, except for the GH item and scale that was loaded to the mental component higher than the physical component. This was also confirmed by the SEM analyses in models 3 and 4 (allowing cross-loadings). Even though the original underlying conceptual model hypothesizes that the GH item and scale should be loaded to the physical component, studies in the US and some European countries indicate that the GH item and scale may be cross-loaded to both components (17-19). Some studies in Asian countries (45, 46) and the Iranian SF-36 (43), also report that the GH item and scale was loaded to the mental component than to the physical component. It is rather apparent that the GH item and scale has mixed factor content (17, 19), which might be reflection of physical and mental aspects. Furthermore, the results of the original study and some other studies showed that the GH item and scale is not the best predictor of the physical component (7, 14). Hann & Reeves (47) found that removing the three scales of the GH, VT, and SF from an oblique model of the SF-12 does not pose a threat to the predictive power or reliability of the components. It can be discussed whether the pattern in respondents' ratings of General Health reflects a cultural bias, as the same phenomenon appeared in several other Asian studies (45, 46).

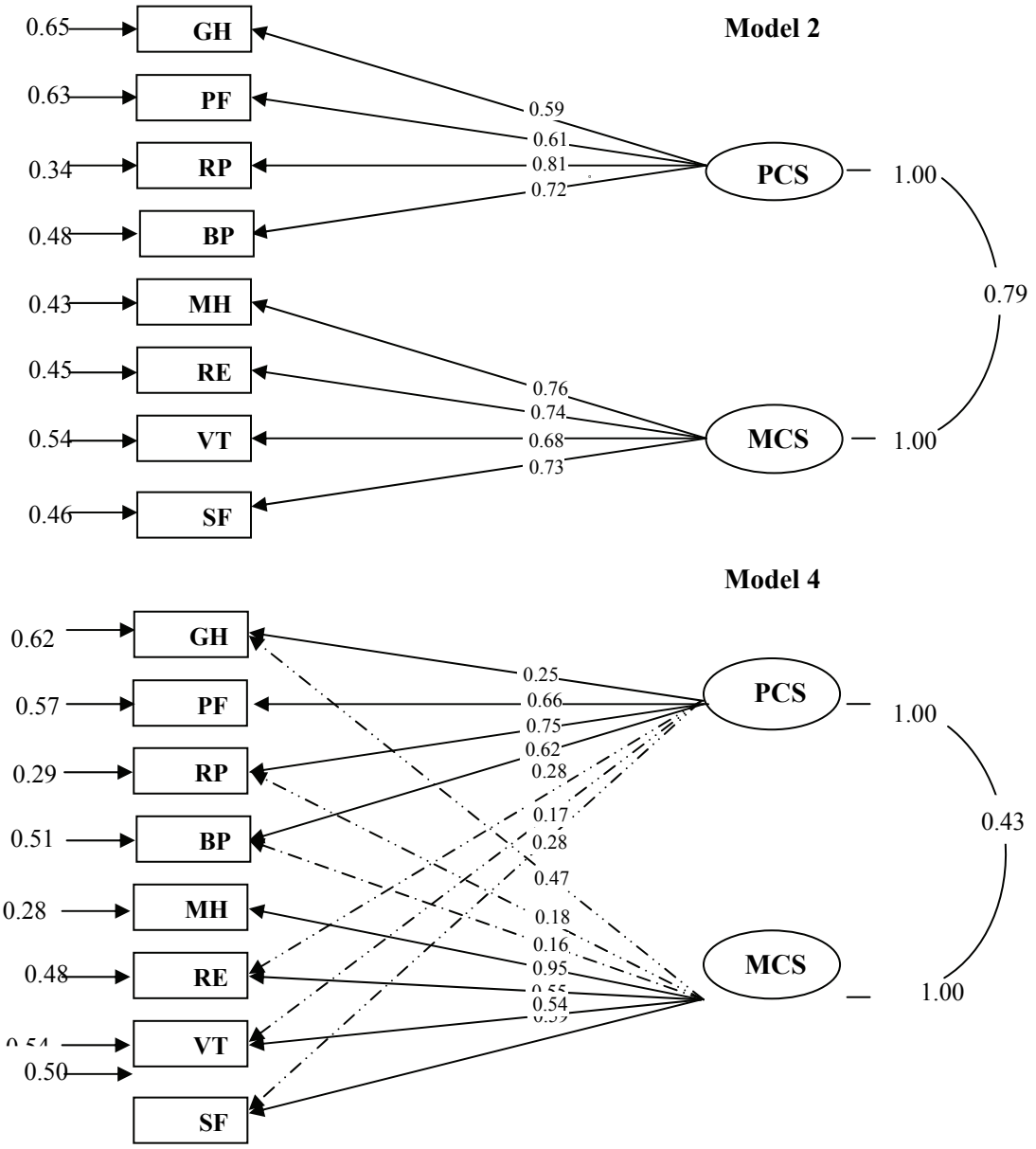

Figure 1. Structure of the Iranian versions of the SF-12v2 with two-factor and eight scales with and without cross-loadings (models 2 and 4) based on SEM analysis.
Rectangles show observed variables and ellipses present latent variables. Numbers in the middle of the one-way arrows reflect the factor loadings and errors in measured variables are located in the left of boxes.

However, this was not eported in the recently published study on the SF-12v1 among an Iranian population (10). The SEM analysis confirmed the construct validity of the Iranian SF-12v2 based on the original SF-12 conceptual model by models 1 (two factors and twelve items) and 2 (two factors and eight scales). There was even an observed improvement in most of the scales loading of the SF-12v2 for model 2, when compared with the Iranian model of the SF-12v1 (two-factor structure and eight scales) (10). This can be related to the improvements made in version 2, particularly the rephrasing of some response alternatives in the scales (the RP, RE, MH and VT scales) and the enhancement made in the psychometric study results (9). For models 3 and 4, the model fit improved when cross-loadings were added. Both models 3 and 4 showed reasonable evidence to support construct validity of the Iranian version of the SF-12v2, as well. However, when the models were compared with each other, according to the study's criteria, all models were acceptable, though models 2 and 4, with a two-factor structure

and eight scales together with and without cross-loadings were best. Designation of different models can be used in future scoring of the Iranian SF-12v2 and comparing the results with the US standard scoring or different algorithms in various Iranian populations.

There was a bivariate significant correlation between the PCS and MCS measures in all models as also shown in previous studies (10, 29, 41, 48, 49). It is important to note that physical and mental health could not be considered independently, though they do not measure the same concept. This might indicate that people do not make a clear distinction between their physical and mental health (48). This is in agreement with Ware *et al.*'s discussion (9), in which they argue for a connection between rated physical and mental health as those with the better physical health probably are in general happier, socially active or energetic.

As expected, based on previous studies (34-37), convergent validity of the Iranian version of the SF-12v2 was supported by the correlations between PCS and MCS measures scores with SOC and HI scores. Thus, the better physical and mental health was rated, the better general well-being and a higher sense of coherence were reported. But, the stronger correlations were related to the mental health.

Internal consistency reliability of the eight scales and the two summary measures of the SF-12v2 was satisfactory. The ICC of the PCS and MCS measures met the study's criteria (5), but when the scores were evaluated by paired t-test, PCS and MCS measures scores showed a significant change from baseline to one month follow-up. The physical health decreased, while the mental health increased. At the same time, the participants' ratings of the eight scales showed no statistical changes, except for the RE and MH scales. On the one hand, it is interesting to note that these changes are considered to be statistical changes and not clinical changes; on the other hand, this contrast can be attributed to negative scoring coefficients used in computing the summary scores that could have produced some inconsistencies between scale and summary scores (50, 51). Therefore, following the recommendations of Ware and Kosinski (52), it is suggested that the summary scores should be reported together with the eight scales scores. However, as a state of "health" can change, it might also indicate sensitivity to change. As no such criterion was used, it is not possible to speculate further. Nonetheless, it is suggested that the ICC is superior analysis to a t-test when estimating robustness because; ICC takes into account the variations within and between individuals. It is also sensitive to both random variation and systematic deviation (5, 53). It is, therefore, suggested that the SF-12v2 in the present study is reliable.

This study has strengths and limitations. The use of both EFA and SEM, designation of the SF-12v2 in four models and comparisons of the hypothesized conceptual models with each other, is important. Also, in this study the SF-12v2 was applied as a self-rated instrument in contrast to a recent publication on the Iranian SF-12v1, where all participants were interviewed. There are several reasons for why self-administration is preferable to interviews, including a greater willingness of the participants to disclose sensitive information, as well as a limited bias towards positive responses to the items and respondents' tendencies to present themselves in the best possible condition (54). One methodological consideration regarding to sampling should be considered. Even though the sample size was reasonable for EFA and SEM analyses, both based on the total numbers of the subjects and in terms of the subject to the item ratio (10 subjects for each item) (25), it should be emphasized that the sample is not a representative sample of the Iranian population. The sample included a greater proportion of women and highly educated people than the general population. Therefore, the results should be considered with some caution and it is suggested that the SF-12v2 further tested with other populations and in different settings.

Conclusion

Construct validity of the Iranian SF-12v2 was confirmed by all factorial models in this study and convergent validity. Whilst paired t-test results were shown to be fair in this study, the results of the internal consistency and test-retest by Cronbach's alpha, and ICC were satisfactory. The findings support the content, face and construct validity of the Iranian version of the SF-12v2, as well as reliability for this sample. It is a psychometrically sound instrument, implying that the Iranian version of the SF-12v2 is suitable for use with large-scale surveys and for cross-cultural HRQoL comparisons. However, one must be aware that the single loading of General Health (the GH item and scale) to the mental component is in contrast to the results of original SF-12 and some other studies found single loading to the physical component or cross-loading to both components.

Acknowledgement
We are grateful to Dr. Ahmad Saidee from Shahid Beheshti University for his statistical advice. We also thank all participants for their cooperation in this study. The project was partially supported by Isfahan University of Medical Sciences Faculty of Nursing and Midwifery.

References:

1. Anderson KL, Burckhardt CS. Conceptualization and measurement of quality of life as an outcome variable for health care intervention and research. J Adv Nurs. 1999; 29(2): 298-306.
2. O'Connor R. Measuring Quality of Life in Health. Edinburgh: Churchill Livingstone; 2004.
3. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of Life. A conceptual model of patient outcomes. JAMA. 1995; 273(1): 59-65.
4. King CR, Hinds PS. Quality of Life from Nursing and Patient Perspectives. Boston: Jones and Bartlett publishers; 2003.
5. Fayers PM, Machine D. Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes. Chichester, UK: John Wiley & Sons Ltd; 2007.
6. Fayers P, Hays R. Assessing Quality of Life in Clinical Trials: Methods and Practice. New York: Oxford University Press; 2005.
7. Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 996; 34(3): 220-233.
8. Medical Outcomes Trust. SF-12® Health Survey. [internet]. USA: Medical Outcomes Trust; 2006 Jan [cited 2010 Mar 30]; [about 1 screen]. Available from: http://www.outcomes-trust.org/instruments.htm#SF-12
9. Ware JE Jr, Kosinski M, Turner-Bowker DM, Gandek B. How to Score Version 2 of The SF-12® Health Survey (With a Supplement Documenting Version 1). Lincoln, RI: QualityMetric Incorporated; 2002.
10. Montazeri A, Vahdaninia M, Mousavi SJ, Omidvari S. The Iranian version of 12-item Short Form Health Survey (SF-12): factore structure, internal consistency and construct validity. BMC Public Health, 2009; 9: 341.
11. Jones PS, Lee JW, Phillips LR, Zhang XE, Jaceldo KB. An adaptation of Brislin's translation model for cross-cultural research. Nurs Res. 2001; 50(5): 300-304.
12. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health. 2005; 8(2): 94-104.
13. Llewellyn CD, McGurk M, Weinman J. Head and neck cancer: to what extent can psychological factors explain differences between health-related quality of life and individual quality of life? Br J Oral Maxillofac Surg, 2006; 44(5): 351-357.
14. Lee CE, Browell LM, Jones DL. Measuring health in patients with cervical and lumbosacral spinal disorders: is the 12-item short-form health survey a valid alternative for the 36-item short-form health survey? Arch Phys Med Rehabil. 2008; 89(5): 829-833.
15. Cheak-Zamora NC, Wyrwich KW, McBride TD. Reliability and validity of the SF-12v2 in the medical expenditure panel survey. Qual Life Res. 2009; 18(6): 727-735.
16. QualityMetric. SF-12v2 Health Survey. [internet]. USA: QualityMetric Incorporated. 2005 [cited 2005 Jan 17]; [about 1 screen]. Available from: http://www.qualitymetric.com/products/surveys/pdf/SF-12v2-Acute-Sample.pdf
17. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993; 31(3): 247-263.
18. Ware JE Jr, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, et al. The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol. 1998; 51(11): 1159-1165.
19. Ware JE Jr. SF-36 Health Survey update. Spine. 2000; 25(24): 3130-3139.
20. Antonovsky A. Unraveling the Mystery of Health: how people manage stress and stay well. San Francisco: Jossey-Bass Publishers; 1987.
21. Rohani C, Khanjari S, Abedi H-A, Oskouie F, Langius-Eklöf A. Health Index, Sense of Coherence Scale, Brief Religious Coping Scale, and Spiritual Perspective Scale: Psychometric Properties. J Adv Nurs. 2010; 66(12): 2796-2806.
22. Forsberg C, Björvell H. Swedish population norms for the GHRI, HI and STAI- state. Qual Life Res. 1993; 2: 349-356.
23. Nordström G, Nyman CR, Theorell T. Psychosocial adjustment and general state of health in patients with ileal conduit urinary diversion, Scand J Urol Nephrol. 1992; 26(2): 139-147.
24. Jöreskog K, Sörbom D. LISREL 8: User's reference guide. Chicago: Scientific Software International; 1996.
25. Munro BH. Statistical Methods for Health Care Research. Philadelphia: Lippincott Williams & Wilkins; 2005.
26. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995; 4(4): 293-307.
27. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res Nurs Health. 2006; 29(5): 489-497.
28. McDowell I. Measuring Health: A guide to Rating Scales and Questionnaires. New York: Oxford University Press Inc; 2006.
29. Fleishman JA, Selim AJ, Kazis LE. Deriving SF-12v2 physical and mental health summary scores: a comparison of different scoring algorithms. Qual Life Res. 2010; 19(2): 231-241.
30. Hurley AE, Scandura TA, Schriesheim CA, Brannick MT, Seers A, Vandenberg RJ, Williams LJ. Exploratory and confirmatory factor analysis: guidelines, issues and alternatives. J Organ Behav. 1997; 18(6): 667-683.
31. Hu L, Bentler PM. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol Methods. 1998; 3(4): 424-453.
32. Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Model. 2007; 14(3): 464-504.
33. Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. Sociol Methods Res. 2008, 36(4): 462-494.

34. Söderman AC, Bagger-Sjöbäck D, Bergenius J, Langius A. Factors influencing quality of life in patients with Meniere's disease identified by a multidimensional approach. Otolo Neurotol. 2002; 23(6): 941-948.
35. Wettergren L, Björkholm M, Axdorph U, Langius-Eklöf A. Determinants of health-related quality of life in long term survivors of Hodgkin's Lymphoma.Qual Life Res. 2004; 13(8): 1369-1379.
36. Von Lengerke T, Janssen C, John J. Sense of coherence, health locus of control, and quality of life in obese adults: physical limitations and psychological normalcies. Int J Public Health. 2007; 52(1): 16-26.
37. Lee A, Browne MO, Villanueva E. Consequences of using SF-12 and RAND-12 when examining levels of well-being and psychological distress. Aust N Z J Psychiatry. 2008; 42(4): 315-323.
38. Cohen L, Morrison K, Manion L. Research Methods in Education. London: Falmer Press; 2000.
39. Nunnally JC, Bernstien IH. Psychometric Theory. New York: McGraw Hill; 1994.
40. Gandek B, Ware JE, Aaronson NK, Apolone G, Bjorner JB, Brazier JE, et al. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol. 1998; 51(11): 1171-1178.
41. Maurischat C, Ehlebracht-König I, Kühn A, Bullinger M. Factorial validity and norm data comparison of the Short Form 12 in patients with inflammatory-rheumatic disease. Rheumatol Int. 2006; 26(7): 614-621.
42. Anagnostopoulos F, Niakas D, Tountas Y. Comparison between exploratory factor-analytic and SEM-based approaches to constructing SF-36 summary scores. Qual Life Res. 2009; 18(1): 53-63.
43. Montazeri A, Goshtasebi A, Vahdaninia M, Gandek B. The Short Form Health Survey (SF-36): translation and validation study of the Iranian version. Qual Life Res. 2005; 14(3): 875-882.
44. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods. 1999; 4(3): 272-299.
45. Fukuhara S, Ware JE Jr, Kosinski M, Wada S, Gandek B. Psychometric and clinical tests of validity of the Japanese SF-36 Health Survey. J Clin Epidemiol. 1998; 51(11): 1045-1053.
46. Tseng HM, Lu JF, Gandek B. Cultural issues in using the SF-36 Health Survey in Asia: results from Taiwan. Health Qual Life Outcomes. 2003; 1:72.
47. Hann M, Reeves D. The SF-36 scales are not accurately summarized by independent physical and mental component scores. Qual Life Res. 2008; 17(3): 413-423.
48. Maurischat C, Herschbach P, Peters A, Bullinger M. Factorial validity of the Short form 12 (SF-12) in patients with diabetes mellitus. Psychol Sci Q. 2008; 50(1): 7-20.
49. Mols F, Pelle AJ, Kupper N. Normative data of the SF-12 health survey with validation using postmyocardial infarction patients in the Dutch population. Qual Life Res. 2009; 18(4): 403-414.
50. Simon GE, Revicki DA, Grothaus L, Vonkorff M. SF-36 summary scores: are physical and mental health truly distinct? Med Care. 1998; 36(4): 567-572.
51. Wilson D, Parsons J, Tucker G. The SF-36 summary scales: problems and solutions. Soz Praventivmed. 2000; 45(6): 239-246.
52. Ware JE, Kosinski M. Interpreting SF-36 summary health measures: a response. Qual Life Res. 2001; 10(5): 405-413.
53. Pellis L, Franssen-Van Hal NL, Burema J, Keijer J. The intraclass correlation coefficient applied for evaluation of data correction, labeling methods, and rectal biopsy sampling in DNA microarray experiments. Physiol Genomics. 2003; 16(1): 99-106.
54. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J of Public Health. 2005; 27 (3): 281-291.