Research Article

Psychometric Properties of the Korean Short Form-36 Health Survey Version 2 for Assessing the General Population

Seon Ha Kim, RN, PhD, [1] Min-Woo Jo, MD, PhD, [2,*] Sang-il Lee, MD, PhD [2]

[1] Department of Nursing, Eulji University, Seongnam, South Korea
[2] Department of Preventive Medicine, College of Medicine, University of Ulsan, Seoul, South Korea

SUMMARY

Purpose: To evaluate the psychometric properties of the Korean short form-36 health survey version 2 for assessing the general population and to provide normative data on the general population.
Methods: Six hundred members of the general Korean population were recruited using a multistage quota sampling method. Data quality was evaluated in terms of the completeness of the data and the response consistency index. Each psychometric property was evaluated using descriptive statistics, item internal consistency, item discriminant validity, known-group validity, internal consistency reliability, and exploratory factor analysis.
Results: The rate of missing data was low, and the rate of consistent responses was similar to conventional criteria. Item internal consistency was acceptable across all scales, whilst item discriminant validity was satisfactory for five of the eight scales. Social functioning was the least acceptable in terms of not only item discriminant validity but also item consistency reliability (Cronbach's alpha = .64). Test-retest Pearson correlation coefficients ranged from .54 to .80. In known group comparison, male sex, age <60 years, high educational status, and the absence of any comorbidities were associated with higher scores than their counterparts. Item factor analysis yields the presence of six factors, accounting for 68.8% of the variance.
Conclusion: The findings of this study generally support the use of the Korean short form-36 version 2 for evaluating the general population, although caution is recommended when interpreting the vitality, social functioning, and mental health scales. Further research is needed in Korea.

Copyright © 2013, Korean Society of Nursing Science. Published by Elsevier. All rights reserved.

Introduction

The health outcomes of a population can be measured not only in terms of classical outcomes of mortality and morbidity but also in terms of subjective outcomes such as self-reported health-related quality of life (HRQOL). Interest in HRQOL issues has increased recently, and the number of citations for "quality of life" in the medical literature has expanded exponentially (Walters, 2009). In order to measure HRQOL as an outcome measure of community-based or hospital-based nursing intervention, an HRQOL instrument was essential. Measuring it in a standardized manner was important for international comparability, while it is also adequate within individual context. There is evidence suggesting the existence of cultural differences in item interpretation

of HRQOL instruments (Thumboo et al., 2001; Tseng, Lu, & Gandek, 2003). Hence, testing for applicability and understanding of psychometric properties of the instruments should precede applying the instruments in research since these instruments developed at other countries need to be adapted to the local context.

The short form-36 health survey version 2 (SF-36 v2) is one of the most popular generic instruments for measuring HRQOL in the world and available in more than 140 translated versions. SF-36 v2 is widely used to assess population health, estimate disease burden, and evaluate intervention outcomes in clinical practice (Gandek, Sinclair, Kosinski, & Ware, 2004). There is extensive evidence of the descriptive validity of SF-36 as a measure of HRQOL for a wide range of medical conditions (Brazier et al., 1992; McHorney, Ware, Lu, & Sherbourne, 1994; McHorney, Ware, & Raczek, 1993). Han, Lee, Iwaya, Kataoka, and Kohzuki (2004) demonstrated the validity of Korean SF-36 in elderly people. However, the psychometric properties of this questionnaire have not been evaluated in the general population. Two studies have been performed in Korea to date (Han et al., 2004; Nam & Lee, 2003), however, the participants of these studies were not representative of the Korean general population.

* Correspondence to: Min-Woo Jo, Department of Preventive Medicine, College of Medicine, University of Ulsan, 86 Asanbyeongwon-gil, Songpa-gu, Seoul 138-736, Korea.

E-mail address: jominwoo@amc.seoul.kr

In addition, if there is result of the specific instrument in population, one could make use of comparison between research data in specific group and general population data.

Therefore, our aim was to evaluate the psychometric properties of the Korean version of SF-36 v2 using a general representative population and to provide SF-36 domain scores according to general characteristics in the general population.

Methods

Study design

This study was performed using individual face-to-face survey. The survey was conducted from April 15—30, 2011 by 37 trained interviewers. Respondents were asked to fill out the Korean version of SF-36 v2. Demographic factors such as age, sex, level of education, and morbidities were also collected.

Setting and samples

Of the 2,025 households that were contacted for interviews, 600 participated in successful interviews (29.6%). Total sample size was arbitrarily decided considering previous research and budget constraint. After considering the age and sex strata, a total of 100 individuals were selected for retesting from among the respondents who consented to participate in the second survey. Individuals ≥19 years of age living in South Korea were eligible to participate in this study. We randomly selected 57 of the 3,429 districts and then recruited individuals by visiting their homes after considering age and sex distributions. Starting point was randomly selected and then every fifth house was selected until the quota target in each district was achieved. The number of participants recruited from each region was proportional to the 2010 registered Korean resident population in terms of sex and age.

Ethical considerations

This study was approved by the institutional review board of Asan Medical Center (approval No.: 2011-0035), and all participants provided written informed consent.

Measurements

Our study used the official Korean SF-36 v2 provided by Quality Metrics. SF-36 v2 is a multipurpose, short-form health survey that includes 36 items and yields an 8-scale profile (physical functioning [PF], role-physical [RP], bodily pain [BP], general health [GH], vitality [VT], social functioning [SF], role-emotional [RE], and mental health [MH]) to assess functional health, well-being, two psychometrically based physical and mental health summary measures, and a preference-based health utility index (i.e., Short Form-6 Dimension[SF-6D]) (Ware, Kosinski, & Bjorner, 2007).

Data analysis

A number of criteria were evaluated, including completeness, consistency, item internal consistency, item discriminant validity, internal consistency reliability, test-retest reliability, known-group validity (including evaluations of the floor and ceiling effects for each scale) and exploratory factor analysis. These criteria are briefly presented below.

Completeness and response consistency

Completeness of the data was assessed by calculating the percentage of completed responses. Consistent responses were evaluated by examining the percentage of respondents with a Response Consistency Index (RCI) of 0, and an RCI ≥90% was considered satisfactory. RCI consists of checking the consistency between 15 pairs of items. If a pair of response is consistent, the RCI score for that pair would be 0; if a pair of responses is inconsistent, the RCI score for that pair would be 1. For example, a respondent's indication that he can "walk more than a mile" but at the same time "cannot walk 100 yards" would be an inconsistent response. The best RCI score is 0, while the worst score is 15 (Ware et al., 2007).

Item internal consistency and discriminant validity

To evaluate the scaling assumptions that underlie the scoring of SF-36, item descriptive statistics, item internal consistency, and item discriminant validity were examined. Under traditional Likert scaling criteria, the means of each item should be roughly equivalent (Gandek et al., 1998). The internal consistency of each item was evaluated by examining the percentage of items with a correlation ≥.40 on their hypothesized scale. Items with their own scale correlations were calculated after correcting for overlap. The internal consistency rate of each item was considered satisfactory if ≥90% of the hypothesized item-scale correlations were ≥.40 (Ware et al., 2007). The discriminant validity of each item was assessed using the Pearson correlation coefficient between each item and its hypothesized scale; this was also used to compare competing scales. When ≥80% of the hypothesized item-scale correlations were significantly higher than the alternative item-scale correlations, item discriminant validity was considered satisfactory (Ware et al.).

Internal consistency reliability

The percentages of respondents who achieved either the highest score (ceiling) or the lowest score (floor) were calculated because large ceiling and floor effects limit the responsiveness of SF-36 (Gandek et al., 1998). Internal consistency reliability was estimated using Cronbach's alpha. When Cronbach's alpha was ≥.70, the reliability was considered acceptable (Ware et al., 2007). To assess test-retest reliability, the Pearson correlation was used as the domain scale.

Known-group construct validity

To assess known-group construct validity, SF-36 v2 scale scores were calculated in terms of sex, age group, level of education, and health status (i.e., participants who did or did not present with any comorbidity). We assumed that the SF-36 v2 scale scores would be lower in females, older person, poorly educated persons, and those suffering from any disease based on previous publications (Brazier et al., 1992; Montazeri, Goshtasebi, Vahdaninia, & Gandek, 2005; Thumboo et al., 2001). Differences in scale scores between groups are analyzed using the student's t test or analysis of variance with Tukey's test.

Exploratory factor analysis

In order to test whether the Korean SF-36 v2 produced an original hypothesized structure, exploratory item level factor analysis was conducted using principal component analysis with varimax rotation. Factor loadings ≥.40 were considered significant (Sharma, 1996).

All statistical analyses were performed using SAS (version 9.1, SAS Institute Inc., Cary, NC, USA) and Quality Metric Health Outcomes Scoring Software 4.0 (Lincoln, RI, USA).

Results

The mean age of the participants was 44.9 years (± 15.3), and 50.5% of the participants were women. A total of 209 participants (34.8%) reported one or more morbidities. Hypertension was the

Table 1 *Demographic Characteristics of the Participants (N = 700)*

Demographic variables	First survey (n = 600)	Second survey (n = 100)	National census[a] (%)
	n (%)	n (%)	
Gender			
Male	297 (49.5)	51 (51.0)	48.9
Female	303 (50.5)	49 (49.0)	51.1
Age (yr)			
19–29	115 (19.17)	19 (19.0)	17.9
30–39	126 (21.00)	21 (21.0)	21.2
40–49	133 (22.17)	22 (22.0)	22.3
50–59	108 (18.00)	18 (18.0)	17.9
≥60	118 (19.67)	20 (20.0)	20.7
Level of education			
Elementary school	52 (8.7)	6 (6.0)	15.0
Middle school graduate	49 (8.2)	11 (11.0)	9.3
High school graduate	278 (46.3)	48 (48.0)	32.5
≥College	221 (36.8)	35 (35.0)	43.2
Morbidity			KNHANES[b]
None	391 (65.2)	69 (69.0)	56.1
One and more	209 (34.8)	31 (31.0)	43.9

Note.
[a] National Census data from 2010.
[b] Data from Korean National Health and Nutrition Examination Survey 2009. The frequencies are weighted in consideration of the population distribution.

most prevalent condition, which was reported by 92 (15.3%) participants (Table 1). In our study, the proportion of participants with a lower educational level or the presence of any comorbidity was less than that of the general population.

Data quality

Of all of the SF-36 v2 items, there were only three missing responses in item PF04 ("climbing several flights of stairs"). Therefore, the completeness of data was almost 99.99%, and there were no out-of-range values. The proportion of consistent response was 89.3%, which is close to the required criteria of 90%. The percentage of inconsistent responses increased with age and decreased with higher levels of education (Table 2). The proportion of participants who demonstrated an RCI of 1 was 6.2%, and the maximum RCI value was 6. The case where both the MH01 ("very nervous person") and MH03 items ("feel calm and peaceful") were rated as 1 (*all of the time*) and 5 (*none of the time*), respectively. At the same time, these two items were the most frequently encountered inconsistent pair of responses (23 participants; 3.8%).

Table 2 *Frequency of Short Form-36 Health Survey Version 2 Response Consistency Index according to General Characteristics (N = 600)*

Variables	Response consistency index (%)		
	0	1	≥2
Total	89.3	6.2	4.5
Gender			
Male	91.9	5.4	2.7
Female	86.8	6.9	6.3
Age group (yr)			
19–29	90.4	7.0	2.6
30–39	94.4	3.2	2.4
40–49	87.97	6.77	5.26
50–59	88.9	6.5	4.6
≥60	84.8	7.6	7.6
Level of education (yr)			
≤6	78.9	11.5	9.6
7–9	87.76	4.08	8.16
10–12	87.8	7.2	5.0
≥13	94.1	4.1	1.8

Table 3 *Correlations[a] between Short Form-36 Health Survey Version 2 Items and Scales (N = 600)*

Items (range)	M	SD	PF	RP	BP	GH	VT	SF	RE	MH
PF (1–3)										
PF01	2.46	0.67	.72	.48	.53	.56	.53	.44	.37	.37
PF02	2.66	0.59	.82	.56	.54	.54	.49	.49	.42	.41
PF03	2.67	0.57	.80	.47	.52	.48	.49	.46	.36	.39
PF04	2.62	0.60	.78	.47	.46	.49	.45	.42	.35	.38
PF05	2.82	0.48	.79	.45	.40	.37	.37	.43	.36	.33
PF06	2.72	0.55	.84	.51	.52	.49	.48	.47	.41	.42
PF07	2.73	0.58	.81	.53	.52	.47	.42	.49	.42	.38
PF08	2.85	0.43	.82	.44	.42	.34	.33	.47	.35	.33
PF09	2.88	0.39	.77	.38	.39	.32	.31	.40	.31	.29
PF10	2.91	0.37	.63	.34	.29	.25	.21	.30	.29	.22
RP (1–5)										
RP01	4.55	0.93	.47	.82	.48	.39	.33	.53	.70	.35
RP02	4.52	0.96	.50	.82	.51	.40	.34	.55	.72	.37
RP03	4.61	0.82	.57	.86	.58	.49	.46	.64	.70	.46
RP04	4.59	0.85	.56	.87	.54	.46	.42	.60	.72	.41
BP (1–6)										
BP01	5.31	1.11	.55	.54	.87	.54	.50	.57	.44	.49
BP02	5.12	1.20	.55	.57	.87	.51	.50	.61	.49	.50
GH (1–5)										
GH01	3.62	0.92	.52	.43	.52	.66	.60	.41	.31	.46
GH02	4.08	0.99	.39	.33	.38	.59	.50	.38	.26	.47
GH03	3.74	1.09	.36	.37	.37	.60	.46	.38	.29	.42
GH04	3.35	1.13	.37	.31	.34	.54	.51	.33	.23	.39
GH05	3.49	1.08	.46	.39	.48	.73	.59	.42	.30	.52
VT (1–5)										
VT01	3.48	1.24	.48	.39	.44	.62[b]	.61	.35	.30	.52
VT02	3.56	1.15	.42	.27	.40	.58	.61	.38	.24	.62[b]
VT03	4.26	0.91	.33	.32	.35	.41	.46	.47[b]	.34	.58[c]
VT04	3.51	0.97	.30	.28	.35	.43	.47	.34	.31	.45
SF (1–5)										
SF01	4.64	0.80	.43	.50[b]	.56[c]	.34	.35	.47	.48[b]	.40
SF02	4.49	0.87	.49[b]	.58[c]	.49[b]	.51[b]	.52[c]	.47	.55[b]	.56[c]
RE (1–5)										
RE01	4.60	0.83	.42	.73	.41	.32	.34	.52	.82	.40
RE02	4.56	0.88	.41	.71	.45	.33	.36	.58	.82	.41
RE03	4.60	0.80	.41	.71	.47	.34	.39	.56	.83	.48
MH (1–5)										
MH01	4.44	0.86	.38	.41	.40	.43	.50	.48	.45	.62
MH02	4.42	0.85	.35	.39	.35	.38	.46	.53	.48	.57
MH03	3.58	1.11	.27	.23	.36	.41	.53[c]	.30	.21	.48
MH04	4.28	0.88	.33	.33	.36	.41	.52	.43	.36	.60
MH05	3.67	1.03	.30	.28	.42	.52	.61[c]	.37	.27	.56

Note. PF = Physical Functioning; RP = Role Physical; BP = Bodily Pain; GH = General Health; VT = Vitality; SF = Social Functioning; RE = Role Emotional; MH = Mental Health.
[a] Item-scale correlations are corrected for overlap (relevant items were removed from its own scale when assessing correlation).
[b] Item-scale correlation is insignificantly higher on the competing scale than the hypothesized scale.
[c] Item-scale correlation is significantly higher on the competing scale than the hypothesized scale.

Testing the scaling assumptions

The descriptive statistics and the correlation coefficients of the SF-36 v2 items and their scales are shown in Table 3. Item means were roughly similar to their own scales. On the PF scale, the most difficult item (PF01: vigorous activities) demonstrated the lowest mean of 2.46, and the easiest item (PF10: bathing and dressing) demonstrated the highest mean of 2.91. Regarding item internal consistency, all items were correlated with their hypothesized scale at values ≥.40. Each item and its hypothesized scale demonstrated a correlation between .46 and .87. In terms of item discriminant validity, most items were more highly correlated with their hypothesized scales than competing scales, but there were a few exceptions. First, vitality items 1, 2, and 3 ("full of life", "a lot of energy", and "worn out", respectively) were more highly correlated with the GH, SF, or MH scales than with the VT scale. Moreover, SF items 1 and 2 ("extent [to which] health problems interfered" and

Table 4 *Short Form-36 Health Survey Version 2 Descriptive Statistics and Psychometric Results (N = 600)*

Scales	M	SD	Floor (%)	Ceiling (%)	Cronbach's alpha	Test-retest Pearson (r)
PF	86.52	21.68	0.7	48.8	.94	.80
RP	89.15	20.26	0.5	65.5	.93	.62
BP	84.23	22.34	0.5	55.0	.89	.61
GH	66.35	20.02	0.3	3.5	.82	.80
VT	67.52	20.13	0.3	5.8	.74	.69
SF	89.00	17.91	0.3	61.2	.64	.62
RE	89.65	19.26	0.3	69.0	.91	.54
MH	76.95	17.42	0.2	7.3	.78	.66

Note. PF = Physical Functioning; RP = Role Physical; BP = Bodily Pain; GH = General Health; VT = Vitality; SF = Social Functioning; RE = Role Emotional; MH = Mental Health.

"frequency [with which] health problems interfered", respectively) were more highly correlated with most competing scales than the SF scale. In addition, the correlation coefficients for mental health items 3 and 5 ("calm and peaceful", and "been a happy person", respectively) were higher (.53 and .61, respectively) on the VT scale than their hypothesized MH scale. The percentage of items that were more significantly correlated with their hypothesized scale than competing scales was 93.5%.

Properties of each scale

Table 4 presents the descriptive statistics, percentages of floor and ceiling effects, Cronbach's alpha , and test-retest Pearson correlation coefficients for the eight scales that were included on the Korean SF-36 v2. The floor effect was <1% for each scale, whereas the ceiling effect was considerably higher on the PF, RP, BP, SF and RE scales. Internal consistency reliability coefficients were ≥.70 (the minimum standard) on most scales, except for the SF scale which demonstrated a coefficient of .64. The test-retest coefficient ranged from .54 on the RE scale to .80 on the PF and GH scales.

Known-group construct validity

The scale scores of the Korean SF-36 v2, according to sex, age group, and health status, are shown in Table 5. As expected, the

scale scores of female respondents tend to be lower than those of male respondents, and the difference in terms of sex was statistically significant on the PF, BP, GH, VT, and MH scales. The oldest age group (≥60 years) demonstrated a significantly lower value than the other age groups on all scales when applying post hoc Tukey's test. More highly educated people tended to report higher values than less educated people on all scales. Those without comorbidity demonstrated significantly higher scores than those with comorbidities on all eight scales (Table 5).

Exploratory factor analysis

Item factor analysis yields the presence of six factors, accounting for 68.8% of the variance. The results are presented in Table 6. PF, BP, and GH items loaded onto factor 1, 6, and 3, respectively while RP and RE items loaded onto the same factor, that is, factor 2. In contrast, VT, MH, and SF items separately loaded onto more than one factor.

Discussion

Our study investigated the data quality and psychometric properties of the Korean version of SF-36 v2 using a nationally representative sample and provides normative data on the general population. The rate of missing data was very low in our study, and the rate of consistent responses from our representative population almost reached the standard norm of 90%. The consistent response rate in our study was somewhat lower than the consistency reported in a Danish study (94.3%; Bjorner, Damsgaard, Watt, & Groenvold, 1998). More than half of the inconsistent responses were due to the misinterpretation of negative (e.g., Did you feel worn out?) and positive questions (e.g., Did you have a lot of energy?). Although SF-36 v2 was self-administered, we suggest that interviewers confirm the answers with opposite directional questions, particularly when interviewing older or poorly educated participants.

The psychometric properties of the Korean SF-36 v2 generally satisfied the conventional psychometric criteria. All of the correlations between items and their hypothesized scales are >.40. However, the findings regarding item discriminant validity suggest that further research on the Korean SF-36 v2 is required. Some of

Table 5 *Short Form-36 Health Survey Version 2 Scores according to General Characteristics and Health Status (N = 600)*

Variables	PF		RP		BP		GH		VT		SF		RE		MH	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Sex																
Male	90.15*	19.00	90.72	18.42	87.08*	20.71	68.93*	19.73	71.76*	17.98	90.24	16.23	90.85	17.85	78.94*	16.20
Female	82.97	23.50	87.60	21.83	81.42	23.53	63.83	20.00	63.37	21.27	87.79	19.37	88.48	20.50	75.00	18.36
Age group (yr)[†]																
19–29	96.39	9.77	94.84	13.04	89.85	16.68	77.03	16.12	76.03	15.31	93.04	13.47	92.54	16.16	82.26	13.17
30–39	91.79	16.09	91.42	15.45	89.13	17.67	71.31	16.32	71.68	17.74	91.87	12.48	91.27	17.70	79.13	15.76
40–49	91.28	17.74	92.72	16.12	88.96	18.98	67.56	17.24	69.64	18.45	91.07	15.28	92.61	16.37	79.17	14.93
50–59	86.43	17.99	89.93	19.44	84.56	21.97	64.00	20.01	67.25	18.70	88.08	18.83	89.89	19.77	75.83	17.70
≥60	66.02	28.45	76.43	28.73	67.85	27.38	51.46	21.04	52.65	22.11	80.51	24.65	81.57	23.72	67.97	21.46
Level of education (yr)[†]																
≤6	58.08	28.77	69.95	31.12	60.29	30.09	45.44	21.65	43.75	21.29	73.80	29.00	77.88	24.42	65.29	23.23
7–9	77.35	26.22	84.06	26.05	74.73	27.86	55.08	23.11	58.67	21.94	83.16	24.81	83.50	24.91	69.39	18.59
10–12	89.32	18.64	91.19	18.08	87.13	19.33	67.55	17.90	70.66	17.55	91.05	14.69	92.15	17.31	79.10	15.84
≥13	91.73	15.90	92.22	15.02	88.31	18.07	72.26	17.13	71.13	18.15	91.29	14.12	90.65	17.57	78.67	16.00
Health status																
Healthy	92.34*	15.22	93.08*	15.93	89.51*	17.10	71.75*	16.71	71.99*	17.88	92.52*	13.22	92.75*	16.46	79.96*	15.01
Unhealthy [a]	75.64	27.13	81.79	24.95	74.33	27.17	56.26	21.74	59.15	21.44	82.42	23.00	83.85	22.53	71.32	20.06

Note. PF = Physical Functioning; RP = Role Physical; BP = Bodily Pain; GH = General Health; VT = Vitality; SF = Social Functioning; RE = Role Emotional; MH = Mental Health.
* $p < .05$ according to the student's *t* test.
[†] $p < .05$ according to analysis of variance in all scale scores.
[a] Respondent has been diagnosed with ≥1 disease.

Table 6 *Factor Loadings of Short Form-36 Health Survey Version 2 Items after Varimax Rotation*

Items	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
PF						
PF01	.51		.57			
PF02	.64					
PF03	.67					
PF04	.51					
PF05	.81					
PF06	.74					
PF07	.74					
PF08	.88					
PF09	.87					
PF10	.78					
RP						
RP01		.81				
RP02		.80				
RP03		.75				
RP04		.79				
BP						
BP01						.71
BP02						.74
GH						
GH01			.66			
GH02			.57			
GH03			.50			
GH04			.68			
GH05			.63			
VT						
VT01			.56		.51	
VT02			.41		.72	
VT03				.73		
VT04				.57		
SF						
SF01						.65
SF02				.46		
RE						
RE01		.82				
RE02		.80				
RE03		.79				
MH						
MH01				.68		
MH02				.77		
MH03					.79	
MH04				.75		
MH05					.77	

Note. PF = Physical Functioning; RP = Role Physical; BP = Bodily Pain; GH = General Health; VT = Vitality; SF = Social Functioning; RE = Role Emotional; MH = Mental Health.

the items on the VT, MH and SF scales are more highly correlated with other competing scales than their own hypothesized scales. Social functioning, in particular, is the least acceptable in terms of item discriminant validity. When reanalyzing only consistent data sets ($n = 536$), the items that fail the item discriminant validity test include VT01, SF 01, MH03, and MH05, but the correlations between their own scales and competitive scales are not statistically significant. The findings of this study demonstrate similar patterns as the evaluation of the Chinese version of SF-36, evaluated in Chinese Americans (Ren, Amick, Zhou, & Gandek, 1998), and a previous evaluation of the Korean SF-36 v2 evaluated in employees of governmental research institutes (Nam & Lee, 2003). Tseng et al. (2003) also reported that almost all of the items on the VT and SF subscales overlap with the items on the MH subscale. In contrast, in a Danish study, only two items (PF01 and MH03) did not significantly support item discriminant validity, but in both of these cases the correlations between each item and its own scale were still higher than their correlations with other scales (Bjorner et al., 1998). Item discriminant validity was satisfied on all scales in a Thai study (Lim, Seubsman, & Sleigh, 2008); however, the Thai study was limited to a student population. As described in other

studies (Ren et al., 1998; Tseng et al.), the perception of social functioning in Asian cultures may be different from that of Western cultures.

The floor effects of <1% on both the RP and RE scales are unique findings compared with the results of studies performed in other countries. Ten Western countries and Japan demonstrated floor effects >5% on both scales (Gandek et al., 1998); a Thai study reported floor effects of 4.5% and 7.4% on the RP and RE scales, respectively (Lim et al., 2008). The higher ceiling effects that were calculated on the RP and RE scales in our study are similar to the patterns observed in other countries (Gandek et al., 1998; Laguardia et al., 2011; Lim et al.).

Internal consistency reliability was >.70 for all scales, except for the SF scale. However, when we evaluated internal consistency reliability using a consistent data set, Cronbach's alpha was satisfactory at .73 for the SF scale. The reliability of this scale (.64) in our study is relatively lower than that reported for other Western countries (Gandek et al., 1998). This finding, however, is similar to the findings reported for other Asian countries. For example, Cronbach's alpha coefficients for the SF scale were .68 in Japan (Gandek et al., 1998), .57 in Taiwan (Tseng et al., 2003), .55 in Thailand (Lim et al., 2008), and .51 in China (Rui et al., 2011). Social functioning items should be cautiously interpreted due to cultural differences regarding the concept of social functioning. Test-retest reliability in our study was similar to that reported in the Brazier et al. (1992) study (range: .60–.81) and slightly lower than that reported in the Han et al. (2004) study (range: .71–.90). Differences in the SF-36 v2 scale scores in terms of sex, age, educational level, and health status showed evidence of known-group construct validity.

Factor analysis of individual items produced partly matching of items to their hypothesized scales, however, the loading of the items in the VT, MH and SF scales differed from the original one. This was consistent with results of item discriminant validity in our study. This pattern was considerably similar to other studies in Asian countries (Hu, Gruber, & Hsueh, 2010; Thumboo et al., 2001; Wang, Lopez, Ying, & Thompson, 2006). The unique feature of our study was that RE and RP items were loaded onto the same factor and these two domains may be combined as "role limitation" similar to SF-6D. The different factor loading seems to reveal a different view in the Korean population.

There are some limitations to our study. We had recruited respondents nationwide. However, the response rate was around 30% and we were unable to identify the characteristics of the individuals who refused to participate. Therefore, the generalizability of the sample could be limited. The age and sex distributions of our sample were similar to those reported in the 2010 national census, however the proportion of individuals with a poor educational status in this study was lower than that reported in the 2010 census. This educational distribution may impact HRQOL by producing high scale scores and a low floor effect. In contrast, the proportion of participants who had any disease in our sample was higher than the KNHANES data; thus, the result of KNHANES participants is possibly lower than the actual norm. Our study did not explore concurrent validity, therefore, further research on the concurrent validity of SF-36 v2 was required. In addition, a more specific norm (e.g., the disease group) could not be produced due to the limited sample size.

Quality of life is an essential component in nursing field. A variety of HRQOL outcome measures have been used in nursing research. Before applying HRQOL instruments, evidence on psychometric properties of each instrument should be considered. In general, our study provides support for the applicability and validity of SF-36 v2 in public health and nursing research. However, the concept of social functioning may be more appropriate in

Western countries and less so in Korea. Avoiding social activities because of health problems seems to be less acceptable for Korean people than it is for Western people. Further research regarding the conceptualization of social functioning and the responsiveness of the Korean SF-36 v2 is required.

Conclusion

The Korean SF-36 v2 generally seems to be a practical, valid, and reliable instrument for assessing the general population. The finding that the social functioning items were significantly associated with the other scales should be cautiously interpreted in light of the cultural framework. Further research on the vitality, social functioning, and mental health scales in Korea is recommended.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgments

References

Bjorner, J. B., Damsgaard, M. T., Watt, T., & Groenvold, M. (1998). Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. *Journal of Clinical Epidemiology, 51*(11), 1001–1011. http://dx.doi.org/10.1016/S0895-4356(98)00092-4

Brazier, J. E., Harper, R., Jones, N. M., O'Cathain, A., Thomas, K. J., Usherwood, T., et al. (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *British Medical Journal, 305*(6846), 160–164.

Gandek, B., Sinclair, S. J., Kosinski, M., & Ware, J. E., Jr. (2004). Psychometric evaluation of the SF-36 health survey in medicare managed care. *Health Care Financing Review, 25*(4), 5–25.

Gandek, B., Ware, J. E., Jr., Aaronson, N. K., Alonso, J., Apolone, G., Bjorner, J., et al. (1998). Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: results from the international quality of life assessment project. *Journal of Clinical Epidemiology, 51*(11), 1149–1158. http://dx.doi.org/10.1016/S0895-4356(98)00106-1

Han, C. W., Lee, E. J., Iwaya, T., Kataoka, H., & Kohzuki, M. (2004). Development of the Korean version of short-form 36-item health survey: Health related quality of life of healthy elderly people and elderly patients in Korea. *Tohoku Journal of Experimental Medicine, 203*(3), 189–194. http://dx.doi.org/10.1620/tjem.203.189

Hu, J., Gruber, K. J., & Hsueh, K. H. (2010). Psychometric properties of the Chinese version of the SF-36 in older adults with diabetes in Beijing, China. *Diabetes Research & Clinical Practice, 88*(3), 273–281. http://dx.doi.org/10.1016/j.diabres.2010.03.005

Laguardia, J., Campos, M. R., Travassos, C. M., Najar, A. L., Anjos, L. A., & Vasconcellos, M. M. (2011). Psychometric evaluation of the SF-36 (v2) questionnaire in a probability sample of Brazilian households: results of the survey Pesquisa Dimensoes Sociais das Desigualdades, Brazil, 2008. *Health and Quality of Life Outcomes, 9*(61). http://dx.doi.org/10.1186/1477-7525-9-61

Lim, L. L., Seubsman, S. A., & Sleigh, A. (2008). Thai SF-36 health survey: tests of data quality, scaling assumptions, reliability and validity in healthy men and women. *Health and Quality of Life Outcomes, 6*(52). http://dx.doi.org/10.1186/1477-7525-6-52

McHorney, C. A., Ware, J. E., Jr., Lu, J. F., & Sherbourne, C. D. (1994). The medical outcomes study 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care, 32*(1), 40–66.

McHorney, C. A., Ware, J. E., Jr., & Raczek, A. E. (1993). The medical outcomes study 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care, 31*(3), 247–263.

Montazeri, A., Goshtasebi, A., Vahdaninia, M., & Gandek, B. (2005). The short form health survey (SF-36): translation and validation study of the Iranian version. *Quality of Life Research, 14*(3), 875–882. http://dx.doi.org/10.1007/s11136-004-1014-5

Nam, B. H., & Lee, S. W. (2003). Testing the validity of the Korean SF-36 health survey. *Journal of the Korean Society of Health statistics, 28*(2), 3–24.

Ren, X. S., Amick, B., 3rd, Zhou, L., & Gandek, B. (1998). Translation and psychometric evaluation of a Chinese version of the SF-36 health survey in the United States. *Journal of Clinical Epidemiology, 51*(11), 1129–1138. http://dx.doi.org/10.1016/S0895-4356(98)00104-8

Rui, W., Cheng, W., Ma, X. Q., Zhao, Y. F., Yan, X. Y., & Jia, H. (2011). Health-related quality of life in Chinese people: a population-based survey of five cities in China. *Scandinavian Journal of Public Health, 39*(4), 410–418. http://dx.doi.org/10.1177/1403494810395817

Sharma, S. (1996). *Applied multivariate techniques.* New York: John Wiley & Sons.

Thumboo, J., Fong, K. Y., Machin, D., Chan, S. P., Leon, K. H., Feng, P. H., et al. (2001). A community-based study of scaling assumptions and construct validity of the English (UK) and Chinese (HK) SF-36 in Singapore. *Quality of Life Research, 10*(2), 175–188. http://dx.doi.org/ 10.1023/A:1016701514299

Tseng, H. M., Lu, J. F., & Gandek, B. (2003). Cultural issues in using the SF-36 health survey in Asia: results from Taiwan. *Health and Quality of Life Outcomes, 1*(72). http://dx.doi.org/10.1186/1477-7525-1-72

Walters, S. J. (Ed.). (2009). *Quality of life outcomes in clinical trials and health-care evaluation: a practical guide to analysis and interpretation).* West Sussex, UK: John Wiley & Sons.

Wang, W., Lopez, V., Ying, C. S., & Thompson, D. R. (2006). The psychometric properties of the Chinese version of the SF-36 health survey in patients with myocardial infarction in mainland China. *Quality of Life Research, 15*(9), 1525–1531. http://dx.doi.org/10.1007/s11136-006-0012-1

Ware, J. E., Kosinski, M., & Bjorner, J. B. (2007). *User's manual for the SF-36v2 health survey* (2nd ed.). Lincoln: RI. Quality Metric.