THEORY AND METHODS

Chinese SF-36 Health Survey: translation, cultural adaptation, validation, and normalisation

L Li, H M Wang,Y Shen

J Epidemiol Community Health 2003;**57**:259–263

See end of article for authors' affiliations

Correspondence to:
Professor L Li, Zhejiang University School of Medicine, 353 Yan'an Road, Hangzhou, 310006, Zhejiang Province, China; lilu@cmm.zju.edu.cn

Accepted for publication 11 September 2002

Study objective: To develop a self administered Chinese (mainland) version of the Short-Form Health Survey (SF-36) for use in health related quality of life measurements in China.
Design: A three stage protocol was followed including translation, tests of scaling construction and scoring assumptions, validation, and normalisation.
Setting: 1000 households in 18 communities of Hangzhou.
Participants: 1688 respondents recruited by multi-stage mixed sampling.
Main results: The assumption of equal intervals was violated for the vitality and mental health scales. The recoded item values were used to calculate scale scores. The clustering and ordering of item means was the same as that of the source and other two Chinese versions. The items in each scale had similar standard deviations except those in the physical functioning, boduily pain, social functioning scales. The item hypothesised scale correlations were identical for all except the social functioning and vitality scales. Convergent validity and discriminant validity were satisfactory for all except the social functioning scale. Cronbach's α coefficients ranged from 0.72 to 0.88 except 0.39 for the social functioning scale and 0.66 for the vitality scale. Two weeks test-retest reliability coefficients ranged from 0.66 to 0.94. Factor analysis identified two principal components explaining 56.3% of the total variance. The Chinese SF-36 could distinguish known groups.
Conclusions: This study suggested that the Chinese (mainland) version of the SF-36 functioned in the general population of Hangzhou, China quite similarly to the original American population tested. Caution is recommended in the interpretation of the social functioning and vitality scales pending further studies.

An epidemiological transition from predominantly communicable diseases to chronic diseases has taken place since the middle of the past century.[1] In mainland China, long term diseases became the main death causes of urban residents in the 1950s, and those of rural residents in the 1960s.[2] The improved longevity suggests that health status can no longer be well assessed by population mortality statistics; there is a consensus to view health in terms of people's subjective assessment of wellbeing and ability to perform social roles.[3–6] The centrality of people's point of view in monitoring health related quality of life has led to the proliferation of instruments and a rapid development of theoretical literature.[7 8]

The 36-item Short Form Health Survey is a brief self administered questionnaire that generates scores across eight dimensions of health: physical functioning (PF), role limitations due to physical problems (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role limitations due to emotional problems (RE), mental health (MH), and one single item scale on health transition. It has proved useful in monitoring population health, estimating the burden of different diseases, monitoring outcomes in clinical practice, and evaluating treatment effects.[9] In 1991, the SF-36 was selected as the instrument in the International Quality of Life Assessment (IQOLA) Project.[9–16] At the time of this writing, the SF-36 has been translated and tested in more than 40 countries and normed in 12 countries. Several Chinese versions (American Chinese, Hong Kong) have been developed and tested,[17–19] but its acceptability or validity on Chinese in mainland China is not known.

In this article, we report the development of a Chinese (mainland) SF-36 Health Survey and report the results of psychometric testing among the general population in Hangzhou, the capital of Zhejiang Province, southeast of mainland China. We expect the study will stimulate further researches to establish the reliability, validity, and application of the SF-36 among various regions of China so that it can eventually be applicable to all Chinese.

METHODS

Translation of SF-36

The study developed a three stage process to produce a cross culturally comparable translation of the SF-36 with the standard protocol as a reference.[20] Firstly, two postgraduates of social medicine translated the original SF-36 into written Chinese independently. Translators had experience in questionnaire translation but were not familiar with the SF-36. The initial versions were administered to a convenience sample of 21 university students. The translators met in person with the principal investigator to agree on a common primary translation. Secondly, two English teachers rated the translation quality. The principal investigator discussed with the translators and eight professionals on questionnaire survey and developed a revised version. Finally, the revised version was pilot tested in a convenience sample of 28 subjects. Some minor changes were made to develop a final version.

Study setting

A multi-stage mixed sampling was conducted to select a representative sample of the general population. During the first stage, six "Jiedao" (a sub-district neighbourhood administration) were selected from Xiacheng district (central area) and Gongshu district (sub-central area) of Hangzhou, three for each. During the second stage, three communities were selected from each "Jiedao". Equal distance sampling was used. During the third stage, every household in a community had the same probability to be sampled that was equal to the

Table 1 Summary results of tests of item convergent and discriminant validity (n=1316)

		Range of correlations		Internal consistency tests§		Discriminant validity tests¶	
Scale	k*	Item-internal consistency†	Item-discriminant validity‡	#Success/ Total	Success Rate (%)	#Success/ Total	Success Rate (%)
PF	10	0.42–0.72	0.01–0.41	10/10	100.0	79/80	98.8
RP	4	0.70–0.78	0.09–0.45	4/4	100.0	32/32	100.0
BP	2	0.72	0.24–0.43	2/2	100.0	16/16	100.0
GH	5	0.43–0.57	0.14–0.49	5/5	100.0	39/40	97.5
VT	4	0.39–0.49	0.11–0.49	3/4	75.0	28/32	87.5
SF	2	0.28	0.09–0.37	0/2	0.0	2/16	12.5
RE	3	0.72–0.78	0.04–0.48	3/3	100.0	24/24	100.0
MH	5	0.43–0.59	0.04–0.46	5/5	100.0	39/40	97.5

*Number of items and number of convergent validity tests per scale. †Correlations between items and hypothesised scale corrected for overlap. ‡Correlations between items and other scales. §Number of correlations between items and hypothesised scale corrected for overlap ⩾0.40/total number of convergent validity tests. ¶Number of correlations significantly higher/total number of discriminant validity tests.

Table 2 Comparison of Cronbach's α coefficients in studies using different Chinese SF-36 versions

	Cronbach's α		
Scale	Hangzhou (n=1316)	Hong Kong (n=236)*	American Chinese (n=156)†
PF	0.87	0.78	0.92
RP	0.88	0.83	0.82
BP	0.80	0.87	0.78
GH	0.72	0.71	0.82
VT	0.66	0.74	0.73
SF	0.39	0.65	0.54
RE	0.87	0.77	0.88
MH	0.75	0.77	0.74

*Lam *et al.*[18] †Ren *et al.*[17]

fixed sample size n (1000 households) divided by the total households in the two districts represented as N. Family members in a sampled household, aged 18 and older, with the ability to read were eligible subjects. They were asked to complete a survey by self administration. The Myer's index was used to detect preference for all terminal digits from 0 to 9. The theoretical range of Myer's index is from 0 to 90. An index of 0 represents no heaping and an index of 90 represents a heaping of all reported ages at a single digit.[21] The differences were analysed between respondents and non-respondents by the monovariate method and the logistic regression model. Fifty seven subjects were randomly sampled for test-retest study after two weeks.

Scoring of scales

When one half or fewer of the items in a scale were missing, the mean of the non-missing items was used to represent the scale. A scale score was declared missing when more than one half of the items were missing.[9 12] Means and standard deviations of all scale scores were calculated.

Psychometric tests

The SF-36 scale scores were constructed using the method of summated ratings based on five assumptions[12 22 23]: (1) Categorical item responses should be on an interval scale. When the assumption is violated, the responses should be recoded to suit actual differences. This assumption could be checked only for scales that had more than two items with multiple choices: GH, PF, VT, MH. We computed, for each response of an item, the average value of the remaining items in the same scale. Then, we assigned empirical scores to each response level in the following fashion: the lowest response level was given the score of 1, the highest response level the score of K (for total K response levels), and the intermediate response levels were assigned scores that reflected intervals.[14] (2) Items of a given scale should have approximately equal variances and means. (3) Item-scale correlations should be roughly equal for all items in a given scale. (4) Convergent validity: the correlation of each item with its hypothesised scale, corrected for overlap should be 0.40 or above. (5) Discriminant validity: the correlation of each item with its hypothesised scale should be significantly higher than correlations of the same item with competing scales (*t* test for correlation coefficients[24]).

Table 3 Factor loadings expected in the SF-36 measurement model and the actual loadings obtained (n=1688)

	Hypothesised association†		Factorial analysis: Rotated principal components			Relative validity‡	
			Correlations with:				
Scale	Physical	Mental	Physical	Mental	Variance explained	Physical	Mental
PF	+	–	0.59	0.25	0.42	0.49	0.09
RP	+	–	0.84	0.07	0.70	1.00	0.01
BP	+	–	0.48	0.45	0.43	0.33	0.29
GH	*	*	0.35	0.68	0.59	0.17	0.67
VT	*	*	0.16	0.83	0.72	0.04	1.00
SF	*	+	0.52	0.42	0.45	0.38	0.26
RE	–	+	0.74	0.11	0.56	0.78	0.02
MH	–	+	0.06	0.79	0.63	0.00	0.90

†+ strong association ($r \geqslant 0.70$); *moderate association ($0.30 < r < 0.70$); – weak association ($r \leqslant 0.30$). ‡The ratio of explained variance of a given scale in principal component to that of the best scale.

Table 4 Comparison of SF-36 scale scores for the general population of Hangzhou by age

Scale	18–44	45–64	⩾65	*F* value	p Value	MES*
PF	86.0 (18.0)	82.0 (17.4)	68.5 (24.5)	78.424	0.000	0.88
RP	85.3 (29.0)	80.4 (34.3)	68.3 (42.8)	23.984	0.000	0.52
BP	85.0 (17.8)	78.4 (21.8)	75.3 (23.6)	30.898	0.000	0.47
GH	60.0 (19.8)	54.0 (19.4)	50.3 (20.9)	28.720	0.000	0.48
VT	53.3 (20.3)	51.2 (21.1)	48.4 (22.1)	5.510	0.004	0.28
SF	84.2 (16.9)	82.8 (17.6)	79.3 (20.9)	7.139	0.001	0.28
RE	85.3 (30.5)	85.1 (32.2)	79.5 (38.8)	3.069	0.047	0.18
MH	57.9 (21.4)	61.3 (23.5)	62.4 (25.1)	5.694	0.003	0.20

*Maximal effect size (MES) = Δ/SD; where Δ came from the difference between the maximal scale score and the minimal scale score, and SD came from the general population.

Table 5 Comparison of SF-36 scale scores for the general population of Hangzhou by sex

Scale	Male	Female	*t* value	p Value	ES*
PF	84.4 (18.6)	79.9 (20.7)	4.638	0.000	0.23
RP	82.4 (32.6)	79.9 (34.5)	1.495	0.135	0.08
BP	83.0 (19.0)	79.9 (21.8)	3.047	0.002	0.15
GH	58.0 (19.9)	55.2 (20.4)	2.737	0.006	0.14
VT	53.8 (20.9)	50.1 (20.7)	3.532	0.000	0.21
SF	83.1 (17.5)	82.9 (18.1)	0.222	0.824	0.01
RE	84.3 (32.3)	84.5 (32.5)	–0.149	0.881	0.01
MH	60.3 (23.0)	59.1 (22.4)	1.046	0.296	0.05

*Effect size (ES) = Δ/SD; where SD came from the general population.

Reliability was estimated using the test-retest method and the internal consistency method (Cronbach's α). A minimum Cronbach's α coefficient of 0.7 is considered satisfactory for group level comparisons.[9] Validity was assessed using convergent and discriminant validity checks, factor analysis, and construct validity. Factor analysis was expected to yield two principal components named as physical health and mental health. In test of construct validity, or known groups validity, scale scores were compared across groups known to differ, using external information independent of the SF-36. It was hypothesised that SF-36 scores for the old would be lower than those for the young; women would have lower scores than men; people reporting longstanding health conditions would have lower scores than those without any such conditions.[10 25]

All statistical analyses were carried out using the Statistical Package for the Social Sciences (SPSS 7.0 for Windows).

RESULTS

Translation

The Chinese SF-36 translation was equivalent to the original version with a few exceptions. Bowling and playing golf (PF02) were common among Americans and Europeans but not in Chinese. In this version, mopping the floor and practising Tai-Chi were used as complementary examples of moderate activities for clarity because we did not know exactly whether they were culturally equivalent. Translating a mile into its mathematically correct equivalent of 1609 metres expresses a degree of accuracy not intended in the original form. Thus, one mile was translated into 1500 metres. One block was translated into the distance between two street crossings. Some difficulties were also encountered in producing corresponding expressions in Chinese equivalent to full of pep (VT01) and have a lot of energy (VT02). In this Chinese

Table 6 The SF-36 scale scores for the general population of Hangzhou by age and gender group

Age group	PF	RP	BP	GH	VT	SF	RE	MH
18–24								
Men	94.3 (11.1)	91.0 (19.7)	86.6 (13.5)	64.1 (20.2)	58.0 (22.3)	83.2 (15.7)	80.9 (27.9)	52.4 (22.4)
Women	90.2 (11.6)	90.0 (21.2)	84.6 (18.1)	62.4 (17.3)	54.0 (18.8)	85.8 (15.3)	85.9 (28.5)	54.3 (20.5)
25–34								
Men	90.8 (12.7)	91.6 (21.6)	88.0 (16.1)	63.5 (18.2)	55.6 (20.3)	85.2 (16.3)	89.4 (24.9)	57.8 (22.4)
Women	87.6 (14.4)	86.4 (28.0)	87.8 (14.7)	61.7 (18.0)	54.5 (16.8)	86.7 (15.5)	86.0 (30.0)	58.8 (19.7)
35–44								
Men	85.9 (19.4)	84.4 (29.8)	85.4 (17.6)	60.1 (20.7)	55.2 (20.7)	83.4 (17.9)	85.2 (31.7)	60.3 (22.4)
Women	80.8 (20.7)	80.6 (33.6)	81.7 (20.5)	56.1 (20.3)	48.9 (20.8)	83.2 (17.3)	84.2 (32.5)	56.9 (20.8)
45–54								
Men	86.5 (14.9)	83.2 (31.8)	81.5 (19.2)	55.4 (17.5)	51.7 (19.9)	83.4 (16.6)	85.6 (31.2)	58.6 (23.0)
Women	81.1 (17.6)	78.0 (36.9)	75.6 (22.2)	53.0 (21.2)	50.3 (23.4)	82.3 (17.8)	84.6 (32.4)	59.1 (24.2)
55–64								
Men	81.6 (17.3)	82.3 (32.2)	81.8 (20.0)	56.2 (20.1)	55.0 (21.5)	81.8 (17.6)	87.1 (29.8)	65.8 (22.5)
Women	76.8 (18.9)	77.6 (36.0)	74.7 (25.5)	50.9 (18.4)	48.0 (18.8)	83.2 (19.0)	82.8 (35.8)	65.0 (23.2)
⩾65								
Men	73.0 (21.9)	68.2 (43.2)	76.6 (21.9)	52.4 (20.0)	49.7 (21.7)	81.5 (19.4)	76.2 (40.8)	62.8 (24.6)
Women	60.6 (26.9)	68.7 (42.4)	72.8 (26.3)	46.6 (22.1)	46.0 (22.7)	75.3 (23.1)	85.1 (34.6)	61.8 (26.0)

version, VT01 conveyed that one is ready to work physically and spiritually, while VT02 emphasised physical health.

Completeness of data

Of the 1972 eligible subjects, respondents were 1688 (85.6%). The mean age was 46.0 years. The Myer's index was 7.94, suggesting a fairly accurate age reporting. Among the respondents, 859 (50.9%) were male. Education levels: 23 (1.4%) were illiteracy or quasi-illiteracy, 243 (14.4%) had primary school education, 1115 (66.4%) had middle school education, and 299 (17.8%) had college or higher education. Marital status: 175 (10.5%) were unmarried, 1400 (84.4%) were married, 25 (1.5%) were separated or divorced, and 59 (3.6%) were widowed. The mean time to complete the questionnaire was 10 minutes. Altogether 1316 (78.0%) respondents answered all 36 items. On average, 3.8% of responses per item (range 0.3%–6.6%) were missing.

Non-response bias

Non-respondents were older, female, less educated. Of them, 54.3% were 65 years old and over, 64.6% were women, 65.5% were illiteracy or quasi-illiteracy. There were significant differences in age, sex, marital status, education level, occupation, and family patterns between respondents and non-respondents ($p<0.05$). Results of logistic regression models suggested: higher education level, and closer ties of family relationship were predictive of response ($p<0.05$).

Tests of scaling assumptions

The assumption of equal intervals was well supported in the GH and PF scales. Going from the least to the most favourable answer, average empirical scores were 1.0, 3.0, 4.0, 4.5, 5.0 for GH01 item, 1.0, 1.5, 2.5, 3.5, 5.0 for GH02–GH05 items, and 1.0, 2.0, 3.0 for the PF scale. However, the assumption was violated in the VT and MH scales. The positions of the two most undesirable responses were switched. The empirical scoring schemes were 1.4, 1.0, 1.8, 3.6, 4.7, 6.0 for the VT scale and 2.7, 1.0, 1.2, 2.8, 4.2, 6.0 for the MH scale respectively.

The clustering and ordering of item means was the same as that of the source version[22] and other Chinese versions,[17,18] except for items GH01, PF02, PF03. The items for each scale had similar standard deviations except those for the PF, BP, SF scales. Table 1 shows the results of item convergent and discriminant validity tests. Correlations between items and hypothesised scale were 0.4 or above for all except item VT03 and the SF scale. The average scaling success rates were 91.4% (32 of 35) for convergent validity, and 92.5% (259 of 280) for discriminant validity.

Cronbach's α reliability coefficients ranged from 0.72 to 0.88 for six scales, 0.66 for the VT scale and 0.39 for the SF scale that was equal to or below correlations between the SF and the RE, MH scales respectively. The correlation between the MH and the VT scale was 0.52. Table 2 shows comparison of Cronbach's α in studies using different Chinese SF-36 versions.[17,18] The two weeks test-retest reliability coefficients ranged from 0.66 to 0.94.

Factor analysis identified two principal components that could be used to explain 56.3% of the total variance. However, the results were not entirely consistent with the hypothesised model.[9] The PF scale was fairly evenly loaded on the "physical" factor, the factor loading 0.59 is lower than the RP scale. The RE scale was found to have a strong association with the "physical" factor and a weak association with the "mental" factor. The VT scale was found to have a higher loading on the "mental" factor than the MH scale. The SF scale was fairly evenly loaded on the both factors (table 3).

Key points

- With the transition of the disease spectrum, Health related Quality of Life (HRQOL) instruments are becoming necessary tools in the health status measurement and clinical effectiveness assessment. Although many have been developed for Western populations, few are available to the Chinese.
- We report the development of a self administered Chinese (mainland) version of the Short-Form Health Survey (SF-36) and report the results of psychometric testing, reliability, and validity among the general population.
- The Chinese (mainland) version of the SF-36 functioned in general population of Hangzhou similarly to the American population tested.
- The results of studies on application of different versions of the Chinese SF-36 to different Chinese groups were compared.
- To improve the Chinese SF-36 scales, further studies among various Chinese regions and ethnic groups are needed.

As tables 4 and 5 show, all the scale scores for the old were lower than those for the young ($p<0.05$), women had lower scores in all scales than men except the RE scale. The differences were significant ($p<0.05$) in the PF, BP, GH, and VT scales. Table 6 presents the norm reference by age and sex group. The comparison of the SF-36 scale scores for different Chinese populations and the US norms are given in table 7.[9,17,26]

DISCUSSION

The translation process set by the IQOLA Project entails forward translations by at least two translators who were native speakers of the target language, rating of translation quality by two other bilinguals, and back translations by two translators who were native speakers of American-English or British-English.[20] Because native English speakers were unavailable, we did not fully adhere to this strategy. Our study suggested that the Chinese (mainland) version of the SF-36 functioned in the general population of Hangzhou, mainland China similarly to the original American population tested. Apart for the SF scale, seven scales succeeded in convergent and discriminant validity tests. Cronbach's α coefficients of six scales were satisfactory for group comparison. The two weeks test-retest observed moderate to strong association. Factor analysis identified two principal components. Chinese SF-36 could distinguish known groups successfully.

However, there are still a few areas that need further examination. The item PF02 "moderate activities" and PF03 "lifting or carrying groceries" had lower means than their previous item cluster. This may be because "moderate activities" such as bowling and golf are uncommon and considered difficult to perform among Chinese, and the complementary example practising Tai-Chi is popular only with some old Chinese men.

Table 7 Comparison of the SF-36 scale scores for different Chinese populations and the US norms

Sample	PF	RP	BP	GH	VT	SF	RE	MH
Hangzhou	82.2 (19.8)	81.2 (33.6)	81.5 (20.5)	56.7 (20.2)	52.0 (20.9)	83.0 (17.8)	84.4 (32.4)	59.7 (22.7)
American Chinese*	79.4 (23.4)	67.5 (37.3)	62.3 (21.9)	58.8 (22.7)	59.0 (20.3)	75.1 (22.7)	61.2 (43.7)	63.9 (20.4)
Hong Kong†	91.8 (12.9)	82.4 (31.0)	84.0 (21.9)	56.0 (20.2)	60.3 (18.6)	91.2 (16.5)	71.7 (38.4)	72.8 (16.6)
US norm‡	84.2 (23.3)	81.0 (34.0)	75.2 (23.7)	72.0 (20.3)	60.9 (21.0)	83.3 (22.7)	81.3 (33.0)	74.7 (18.0)

*Ren *et al.*[17] †Lam *et al.*[26] ‡Ware *et al.*[9]

The same applied in the US study.[17] "Lifting or carrying groceries" is abstract to Chinese in the mainland. The scaling assumption on equal item variance could not be satisfied in PF, BP, and SF scales. The standard deviations of PF05, PF09, PF10 measuring low levels of functioning were smaller than other items in the same scale because more than 85% of the subjects scored the highest score of 3 on these three items. The standard deviations of items BP02 and SF01 were smaller than items BP01 and SF02 respectively. The same was found in studies in Kong Hong and the US.[17 18] This finding seems to point to the differences in cultural interpretation of items. Deeply ingrained in the Confucian ideology of collectivism, it is socially unacceptable for Chinese to use "sickness" as an excuse to avoid working or socialising with others.

The Chinese (American Chinese) version produced similar findings with respect to reliability, convergent, and discriminant validity tests.[17] Both versions found poor (<0.4) levels of item-scale correlation for the SF scale. The item SF01 was more highly related to the BP, RE. and MH scales, and the item SF02 was more highly related to the VT and MH scales. The item VT03 was highly correlated with the MH scale than the parent scale. Cronbach's α coefficient was below 0.70 for the SF scale. The MH scale was strongly correlated with the VT scale. However, application of the Chinese (Hong Kong) version shared less common factors with these data.[18] Correlations between items and hypothesised scale were 0.4 or above for all except items PF03, PF05, PF09, PF10, and GH01. The scaling success rate for discriminant validity was 100% for all scales except the PF scale. Cronbach's α coefficients were more than the inter-scale correlations for all the scales, but that for the SF scale was still below 0.7. Given the fact that there are apparent regional differences in China in terms of economy, culture, and even language, further researches among various Chinese regions and ethnic groups are needed to improve the Chinese SF-36.

Of the eight scales, the SF scale was least satisfactory in the scaling assumption testing, because of only two items in this scale and lower item homogeneity. Factor analysis revealed two principal components, but there were still some deviations from the hypothesised model. The study of a Chinese (Taiwanese) SF-36 version produced the similar results.[19] Results of a Chinese (Hong Kong) version fit the hypothesised physical/mental health structure better,[18] but application of the same version in a big sample size in Singapore yielded similar pattern of factor correlations comparable to our study.[27] It is suggested that the conceptual framework of the instrument needs to be further improved for cross cultural health status measurement.

ACKNOWLEDGEMENTS

We gratefully acknowledge Mr Shanrong Cai, Miss Weining Ma, and MrsYe Gu for their help in translation, Professor Qianjin Jiang, Professor Jian Wang, and Professor Gengyao Fu for their useful comments on psychometric tests. We also thank the staff of Xiacheng and Gongshu District Health Bureau, for their assistance in the heath survey.

Authors' affiliations

L Li, H M Wang, Department of Social Medicine, School of Medicine, Zhejiang University, China
Y Shen, Department of Health Statistics, School of Medicine, Zhejiang University

Funding: the study was supported by a grant from the Science and Technology Bureau of Zhejiang Province of China. (grant no: 991104209).

Conflicts of interest: none.

REFERENCES

1 **Oman AR**. Epidemiologic transition in the United States. *Population Bulletin* 1977;**32**:2.
2 **Gong YL**. *Social medicine*. Beijing: People's Hygiene Press, 2000.
3 **Patrick DL**, Erickson P. *Health status and health policy: allocating resources to health care*. New York: Oxford University Press, 1993.
4 **Geigle R**, Jones SB. Outcomes measurement: a report from the front. *Inquiry* 1990;**27**:7.
5 **Campen CV**, Sixma H, Friele RD, *et al*. Quality of care and patient satisfaction: A review of measuring instruments. *Med Care Res Rev* 1995;**52**:109–33.
6 **Till JE**, Osoba D, Oater L, *et al*. Research on health-related quality of life: Dissemination into practical applications. *Qual Life Res* 1994;**3**:279–83.
7 **McDowell I**, Newell C. *Measuring health: a guide to rating scales and questionnaires*. New York: Oxford University Press, 1987.
8 **Guillein F**, Bombardier C, Beaton D. Cross-culture adaptation of health-related quality of life measure: Literature review and proposed guidelines. *J Clin Epidemiol* 1993;**46**:1417–32.
9 **Ware JE**, Snow KK, Kosinski M, *et al*. *SF-36 Health survey- manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Center, 1993.
10 **Ware JE**, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): Conceptual framework and item selection. *Med Care* 1992;**30**:473–83.
11 **McHorney CA** , Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey: Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993;**31**:247–63.
12 **McHorney CA**, Ware JE, Lu JFR, *et al*. The MOS 36-Item Short-Form Health Survey (SF-36):III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care* 1994;**32**:40–66.
13 **Aaronson NK**, Acquadro C, Alonso J, *et al*. International quality of life assessment (IQOLA) project. *Qual Life Res* 1992;**1**:349–51.
14 **Perneger TV**, Leplege A, Etter JF, *et al*. Validation of a French-language version of the MOS 36-item short form health survey(SF-36) in young healthy adults**.** *J Clin Epidemiol* 1995;**48**:1051–60.
15 **Bullinger M**. German translation and psychometric testing of the SF-36 health survey: preliminary results from the IQOLA project. International Quality of Life Assessment. *Soc Sci Med* 1995;**41**:1359–66.
16 **Gandek B**, Ware JE. Methods for validating and norming translations of health status questionnaires: the IQOLA project approach. *J Clin Epidemiol* 1998;**51**:953–9.
17 **Ren XS**, Amick B, Zhou L, *et al*. Translation and psychometric evaluation of a Chinese version of the SF-36 health survey in the United States. *J Clin Epidemiol* 1998;**51**:1129–38.
18 **Lam CLK**, Gandek B, Ren XS, *et al*. Tests of scaling assumptions and construct validity of the Chinese(HK) version of the SF-36 health survey. *J Clin Epidemiol* 1998;**51**:1139–47.
19 **Fuh JL**, Wang SJ, Lu SR, *et al*. Psychometric evaluation of a Chinese (Taiwanese) version of the SF-36 health survey amongst middle-aged women from a rural community. *Qual Life Res* 2000;**9**:675–83.
20 **Bullinger M**, Alonso J, Apolone G, *et al*. Translation health status questionnaires and evaluating their quality: The IQOLA Project Approach. *J Clin Epidemiol* 1998;**51**:913–23.
21 **Shryock**, H., Siegel, J. *The methods and materials of demography*. San Diego: Academic Press, 1976.
22 **Ware JE**, Keller SD, Gandek B, *et al*. Evaluating translations of health status questionnaires: Methods from the IQOLA Project. *Int J Technol Assess Health Care* 1995;**11**:525–51.
23 **Likert R**. A technique for the measurement of attitudes. *Arch Psychol* 1932;**140**:5–55.
24 **Jin PH**. *Medical statistical method*. Shanghai: Shanghai Medical University Press, 1993.
25 **McHorney CA**, Kosinski M, Ware JE. Comparisons of the costs and quality of norms for the SF-36 Health Survey collected by mail versus telephone interview: results from a national survey. *Med Care* 1994;**32**:551–67.
26 **Lam CLK**, Lauder IJ, Lam TP, *et al*. Population based norming of the Chinese (HK) version of the SF-36 health survey. *The Hong Kong Practitioner* 1999;**21**:460.
27 **Thumboo J**, Fong KY, Machin D, *et al*. A community-based study of scaling assumptions and construct validity of the English (UK) and Chinese (HK) SF-36 in Singapore. *Qual Life Res* 2001;**10**:175–88.