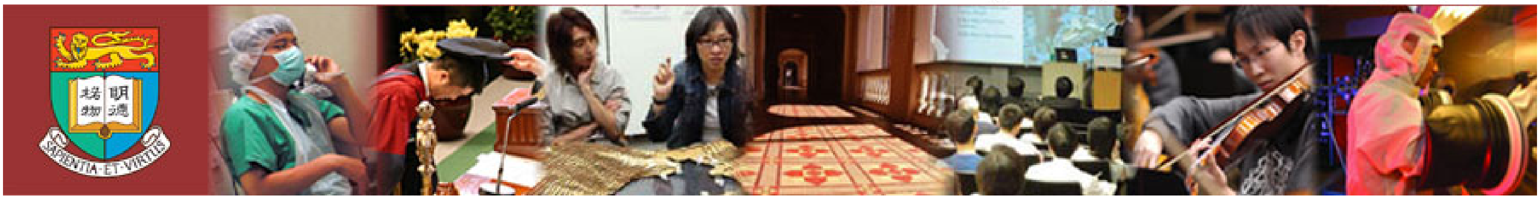

Tests of Scaling Assumptions and Construct Validity of the Chinese (HK) Version of the SF-36 Health Survey

Cindy L K Lam, FRCGP, FHKAM (Family Medicine)[1]
Associate Professor, Family Medicine Unit, the University of Hong Kong

Barbara Gandek, MS [2]
Senior Project Director, IQOLA Project, New England Medical Center, Boston, USA

Xinhua Steve Ren, Ph.D.[3]
Senior Research Scientist,
Center for Health Quality, Outcomes, and Economic Research,
Edith Nourse Rogers Memorial Veterans Hospital, Bedford, USA
Assistant Professor, Boston University School of Public Health, Boston,USA

M.S. Chan, Ph.D [4]
Associate Professor, Department of Chinese, the University of Hong Kong

1. General Practice Unit, 3rd Floor, Ap Lei Chau Clinic, 161 Main Street, Ap Lei Chau, Hong Kong SAR, China. Fax: (852) 2814 7475, E mail: clklam@hku.hk

2. The Health Institute, New England Medical Center, Box #345, 750 Washington Street, Boston, Massachusetts 02111, U.S.A.

3. Center for Health Quality, Outcomes and Economic Research, Edith Nourse Rogers Memorial Veterans Hospital, 200 Springs Road, Bedford, MA 01730, U.S.A.

4. Department of Chinese, the University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

Tests of Scaling Assumptions and Construct Validity of the Chinese (HK) Version of the SF-36 Health Survey

Abstract

Few health-related quality of life (HRQOL) survey instruments are available to the Chinese although many have been developed for Western populations. This paper describes the testing of the acceptability, conceptual equivalence, scaling assumptions and construct validity of a Chinese (HK [Hong Kong]) version of the MOS SF-36 Health Survey. A Chinese (HK) SF-36 survey form was developed by an iterative translation process. It was administered to 236 Chinese subjects who also rated the understanding, difficulty, relevance, and acceptability of each question. The scores were tested against the original scaling assumptions. The SF-36 profile of our subjects was compared to US results for conceptual equivalence. Most subjects did not have any problem in understanding and answering the SF-36. Item means were generally clustered as hypothesized. All but a few items satisfied all scaling assumptions. The shape of the eight scale SF-36 profile was similar to that of American patients, suggesting conceptual equivalence. We conclude that the Chinese (HK) version of the SF-36 Health Survey has achieved conceptual equivalence and satisfied the psychometric scaling assumptions well enough to warrant further use and testing, using the standard scoring algorithms.

Keywords: SF-36 Health Survey, Health-related quality of life, Chinese, Cross-cultural, Construct validity

Introduction

Rapid economic development in Hong Kong has resulted in an increase in life expectancy and aging of the population. The average life expectancy is 81 years for women and 76 years for men in Hong Kong [1]. The proportion of the population who are 65 years or older has doubled from the 4.5% in 1971 to 9% in 1996. Chronic disabling diseases such as stroke and arthritis have become the major health problems, although they are often not lethal, they can affect the quality of life and place a substantial burden on the health care system. Traditional indicators such as mortality and objective clinical parameters are no longer sufficient to assess the effect of illness and the outcome of treatments. They have to be supplemented with self-rated health-related quality of life (HRQOL) measures [2-4]. HRQOL has been found to be valid and sensitive in predicting mortality in the elderly [5], detecting functional impairment [6], and determining consultation rates [7]. There is an increasing demand for a valid and acceptable HRQOL measure for the people in Hong Kong.

Although many HRQOL measures have been developed in the last two decades in Western countries [8], few are applicable to the people in Hong Kong. The major obstacle is the cultural and language difference between the populations of Hong Kong and Western countries. Ninety-six percent of the population of Hong Kong is Chinese, their written language is Chinese and the daily spoken language is Cantonese. Only 30% of the population can speak English [1]. Translation into Chinese and testing for cross-cultural validity are required before a HRQOL survey form can be applied the people in Hong Kong.

The MOS 36-Item Short Form Health Survey (SF-36) developed by Ware et al. in the United States (U.S.) is gaining international popularity [9,10]. It consists of 36 items grouped under 11 questions. The scores of the 36 items are summated into eight multi-item scales: physical functioning, limitations due to physical health problems (role-physical), bodily pain, general health, vitality, social functioning, limitations due to emotional health problems (role-emotional), and mental health, and one single-item scale on health transition. Higher scores represent better health status. The SF-36 captures most of the important concepts of HRQOL [4,11] and has been shown to be useful in general and clinical populations [9].

At the time of writing this paper, the SF-36 has been translated and tested in more than 40 countries, and normed in 12 countries. Ren et al developed and tested a Chinese version of the SF-36 on Chinese Americans [12], but its acceptability or validity on Chinese people living in Asia is not known.

The aim of our study was to test the acceptability, conceptual equivalence, scaling assumptions and construct validity of a Chinese (HK) version of the SF-36. If these tests are met and other tests of validity confirmed, the Chinese (HK) SF-36 would have potential application to Chinese people living in Hong Kong as well as on those who have migrated from Hong Kong to other countries. We hope that our work will stimulate a wider use of HRQOL as a health and outcome indicator in the care of nearly one quarter of the world's population who is Chinese.

Method

The IQOLA Project has developed a three-stage method for cross-cultural adaptation of the MOS SF-36 Health Survey [10]. The first is translation of the original survey into the native language and evaluation to ensure conceptual equivalence and respondent acceptance, to produce a form that can be used in data collection. The second stage is formal psychometric tests of the assumptions underlying item scoring and construction of multi-item scales, to ensure that the scoring algorithms can be applied to the population concerned. The third stage is the validation and norming studies that provide a basis for interpretation [10]. The first two stages are prerequisites to the third and are essential before the population concerned can use the instrument. The Chinese (HK) SF-36 was subject to the first two stages of evaluation in this study.

Development of the Chinese (HK) SF-36

Two professional English-Chinese translators who were native speakers of Chinese translated the original US-English SF-36 into written Chinese independently. A panel consisting of the translators and the three bi-lingual authors (Lam, Ren, and Chan) reviewed the translations and compared them with the Chinese translation developed by Ren et al [12] to form the first draft of the Chinese (HK) SF-36. Another translator who was blind to the original SF-36 back translated this first draft to English. The back-translation was assessed for equivalence to the original by the IQOLA Project Director (Gandek). Discrepancies between the original US-English form and the first draft Chinese (HK) translation were reviewed by the panel and a second version of the Chinese (HK) SF-36 was developed. The latter was back translated by another translator to

English, and the IQOLA Project Director again rated the back-translation. The second version of the Chinese (HK) SF-36 was assessed by four Chinese linguistic experts, one each from Hong Kong, Mainland China, Taiwan, and Singapore, for semantic equivalence to the original SF-36, clarity and grammatical accuracy. The second version of the Chinese (HK) SF-36 was used in stage two of the study.

Study Sample

Two convenience samples of Chinese people aged over 14 years were used in order to include subjects from a wide range of age, educational, and social groups. The first sample was drawn from one of every three patients attending a Government funded family medicine clinic in Hong Kong from December 14, 1995 to February 16, 1996. The second sample consisted of University students randomly sampled from membership directories of the Social Service Group and the Catholic Society of the University of Hong Kong, these two student groups were chosen because they included students from different faculties and all years of studies. The Chinese (HK) SF-36 was administered by an interviewer in Cantonese to each subject who then indicated whether he/she understood each question, found it difficult to answer, thought it relevant to him/her, and minded answering the question. Information on age, sex, educational level, and social class by occupation of the head of family of the respondent was also obtained [13].

Data Analysis

The item means were clustered and compared to a hypothesized order that had been confirmed in previous studies [10]. It was hypothesized that it was less likely for people to be able to achieve higher than lower levels of function or to endorse positive than negative health states. An item that measures a higher level of function, e.g. vigorous activities (PF1), should have a lower mean than one that measures a lower level of function, e.g. bathing or dressing (PF10); an item that measures positive health, e.g. I am a happy person (MH5), should have a lower mean than one that measures negative health, e.g. feeling downhearted and blue (MH4). Items within a scale measuring similar levels of function were put into one cluster and those measuring different levels of function were put into different clusters. The relative order of the item-cluster means should follow the hypothesized order by the levels of function they measure [10]. Items within the same cluster should have similar means and no ordering was hypothesized. If each translated item of the Chinese (HK) SF-36 defined the same level of health as in the original form the item means should cluster in the same order as hypothesized.

We used the Multitrait Analysis Program-Revised (MAP-R) to test whether the scores satisfied the scaling assumptions of the original SF-36 [14]. The SF-36 scale scores were constructed using the method of summated ratings, or Likert-type scale construction, based on five assumptions [10,15,16]: 1. Items measuring the same concept should have approximately equal variances (standard deviations), to avoid the need for standardization - this is a test of equal item variance; 2. An item should be substantially linearly related to the underlying concept being measured (item-scale correlation should be 0.4 or above) - this is a test of item internal consistency; 3. Items in a given scale should contain about the same proportion of information about a concept, therefore, there should be roughly equal item-scale correlation within a scale (equivalent item-scale correlation); 4. An item should correlate higher with its hypothesized scale than with scales measuring other concepts (item discriminant validity). The statistical significance of the difference between the item-hypothesized scale and item-competing scale correlation was tested by the Steiger's t-test for dependent correlation [17]; and 5. Scale scores should be reproducible (reliability) and interpretable (inter-scale correlation). Internal reliability of scale scores was measured by the Cronbach's alpha coefficient. Nunnally has suggested 0.7 as the minimum reliability coefficient for group comparison [18]. Correlation between scales should be less than their internal reliability coefficients (Cronbach's alpha) if each scale measures a unique concept.

After confirmation of scaling assumptions, summation of the raw scale scores and calculation of the transformed scale scores according to the standard formula described in the SF-36 Health Survey Manual [9] were done by the SPSS for Windows program. The scale means and the relative relationship (shape) of the eight SF-36 profile were compared to those of the U.S. general population and U.S. patients with chronic diseases [9,15]. The shape of the profiles should be similar if there is conceptual equivalence between the Chinese (HK) SF-36 and the original form [10].

Previous factor analysis studies in the U.S. and Western European countries have confirmed the construct validity of the SF-36 in relation to its hypothesized two principal components of health (physical and mental) [19,20]. The physical functioning (PF), role-physical (RP), and bodily pain (BP) scales have a strong association with the physical component and a weak association with the mental component. Mental health

(MH), role-emotional (RE), and social functioning (SF) have a strong association with the mental component and a weak association with the physical component. Vitality (VT) and general health (GH) scales have a moderate to substantial association with both physical and mental components. To test whether our results fit the hypothesized physical/mental structure of the original SF-36, two principal components were extracted from the correlation among the eight scales of the sample and rotated to orthogonal simple structure using SAS [19,20].

Results

Sample

Two hundred and thirty-six Chinese subjects (185 clinic patients, and 51 University students) were surveyed. There were 184 (78%) females and 52 (22%) males. The mean age of the subjects was 43 years (S.D. 18.3, range 15-93). The social class distribution was 3 (1.3%) professionals (I), 12 (5.1%) associate professionals (II), 115 (48.7%) skilled workers (III), 46 (19.5%) semi-skilled workers (IV), and 60 (25.4%) unskilled workers (V). Fifty-two subjects (22%) had no formal education, 58 (24.6%) had primary (1-6 years) education, 69 (29.2%) had secondary (7-13 years) education, and 57 (24.2%) had tertiary (>13 years) education.

Translation Equivalence and Acceptability of the Chinese (HK) SF-36

The back-translation of the Chinese (HK) SF-36 was equivalent to the original SF-36 for all the questions and responses with a few exceptions. In the Chinese (HK) SF-36, practicing Tai-Chi was used instead of playing golf as an example of moderate exercise; the distance of walking was expressed in both number of blocks and meters for clarity. The literal translation of ‘moderately’ in Chinese means ‘middle’ which is not the same as the original meaning in English; we translated it to the Chinese term that meant ‘somewhat’, which was closer in meaning to the original. The linguistic experts agreed that the Chinese (HK) SF-36 was equivalent in meaning to the original form and the translation was easy to understand.

All subjects answered all the questions of the Chinese (HK) SF-36 and there were no missing scores. Table one shows the number (%) of subjects who indicated problems for each of the questions. Very few people did not understand, had difficulty with or minded answering any of the questions. A number of subjects said that the questions on social functioning, physical functioning, role-physical or bodily pain were not relevant to them, the main reason given was because they had very few social or physical activities, or bodily pain.

Psychometric Analysis of the SF-36 Scores

Table 2 lists the mean scores and standard deviations (SD) of the 36 items grouped under their scales in the hypothesized item-cluster order, from low to high mean scores. The clustering and ordering of the item means of our subjects were the same as that hypothesized, except for items PF10, GH3, and RE3. The items within each scale had similar standard deviations (scaling assumption one on equal item variances) except those of the PF scale. The standard deviations of PF3, PF5, PF9, and PF10 were relatively small because over 95% of the subjects scored the highest score of 3 on these items.

Table 3 shows the Pearson item-scale correlation between each item and scale. The correlation between an item and its hypothesized scale was 0.4 or above (scaling assumption 2 on item internal consistency) for all except PF3, PF5, PF9, PF10, and GH1. The item-scale correlations within the same scale were generally similar (scaling assumption 3) except for items PF3, PF5, PF9, PF10, and RE3. The scaling success rate on discriminant validity (assumption 4) was 100% for all scales except the PF scale, which had only 92.5% scaling success because item PF10 had a lower correlation with its hypothesized scale than with six of the eight competing scales.

Table 4 shows that the Cronbach's alpha coefficients of internal reliability (R_{TT}) were above the standard of 0.7 for all except the SF scale whose alpha was only 0.65. The inter-scale correlations were less than the scale internal reliability coefficient for all the scales, showing that each scale measured a unique concept, relative to other scales. The results satisfied scaling assumption 5 on reliable and interpretable scale scores. Since the data, with a few minor exceptions, satisfied the scaling assumptions, the scale scores were calculated using the standard SF-36 algorithms [9].

Table 5 shows the hypothesized associations and the rotated factor loadings between the eight scales and the two (physical/mental) health components. Our results fit the hypothesized physical/mental health structure well, although the GH scale had a lower loading on the mental component than that hypothesized.

Table 6 compares the scale scores of our clinic patients (n=185) with those of U.S. patients with chronic diseases and U.S. general population [9,15]. The Hong Kong scores spread over the full range, except those of the physical functioning (PF) and general health (GH) scales. There were few floor (subjects who had the lowest possible score) or ceiling (subjects who had the highest possible score) effects in the GH, VT and MH scales. The PF, RP, BP, SF and RE scales had more ceiling than floor effects. The Hong Kong RP, BP and VT means lay between those of American norm and patients, the

Hong Kong GH, RE and MH means were the lower and the Hong Kong PF and SF means were the higher than the U.S. means. Figure one shows that the means and shape of the eight scales of the SF-36 profile of the Hong Kong patients (HK-Pat) were similar to those of the U.S. patients (USA-Pat) [15].

Discussion

We used two sampling frames in our study in order to include subjects from both sexes, a wide range of ages, different social classes and all educational levels in the testing of the Chinese (HK) SF-36. A much larger sample size would be needed if subjects were randomly selected from the general population, in order to include people from extreme age and social groups. We believe that our results on the acceptability, conceptual equivalence, construct validity and psychometric properties of the Chinese (HK) SF-36 can be generalized to other Chinese people in Hong Kong because our sample consisted of subjects with a wide range of demographic characteristics. We excluded the student sample from the comparison with the U.S. results because the students were generally much younger than the other samples and age could have a significant effect on the SF-36 scores [9].

The relatively low relevance ratings on the items on social activities could be explained by the findings of an earlier study in that many Chinese did not think social activities being important [21]. Although it was possible that social activities had a different meaning to our subjects from that intended to be measured by the SF-36, it was unlikely because the same study showed that many Chinese in Hong Kong were able to describe social activities appropriately.

The deviations of the means of items PF10 and GH3 from the hypothesized order of item-clusters called for a re-examination of the concepts of these two items. PF10 (bathing or dressing) had a lower mean than PF9 (walking one block) but the difference was only

0.01. Both PF9 and PF10 measure low levels of functioning that are not expected to be a problem for ambulatory subjects. It might be more appropriate to cluster PF9 and PF10 together. The deviation of GH3 was also observed in the U.S. population [9]. GH3 measures health relative to that of other people which is conceptually more similar to GH2 than GH1 and GH5. The latter two items measure absolute health. Thus, it was not surprising that the mean of GH3 was more similar to that of GH2 than those of GH1 and GH5.

RE3 (didn't do work as carefully: yes/no) had a lower mean and item-scale correlation than the other two RE items. The difference could be the result of a difference in work attitude between the Chinese and Americans. On the other hand, RE3 had a double negative in its question and answer, which might have been misinterpreted by some subjects. As a result of this concern, we have revised the translation of this item to 'did work not as carefully as before: yes/no' for further studies to see if the response to this item will change.

Four PF items (PF3, PF5, PF9 and PF10) did not satisfy some of the scaling assumptions underlying the scoring algorithms of the SF-36 because their scores were skewed. Ninety-six to one hundred percent of subjects scored the maximum score of 3 on these items because they measure low levels of functioning. The little variation in the scores resulted in small standard deviations in these PF items and therefore they could not satisfy the scaling assumption on equal item variance. The skew of the scores in these items also affected their item internal consistency and equal item-scale correlation. In addition, the item-scale correlation between item GH1 and the GH scale was .39, just below the standard of 0.4 for item internal consistency. This was because of skewing towards the lower scores for this item. Many Chinese people are reluctant to say that their health is excellent because they believe God may become jealous and punish them for boasting about their health.

The scores of our subjects fit the hypothesized two physical/mental component structure of the original SF-36 survey form. This means that the SF-36 Physical and Mental Health Summary Scales might be applicable to the Chinese [20]. Further studies on a larger representative sample of the general population are required to confirm their validity and to standardize the weights for the calculation of the physical and mental summary scores [20]. If the SF-36 Physical and Mental Summary Scales were shown to be applicable, we could proceed to the testing of the validity of measuring the two summary scores with the shorter SF-12 survey form [20,22].

The high ceiling effects of the PF, RP, BP, SF and RE scales were expected since they measured limitations and disability, which many of our ambulatory subjects did not have. On the other hand, the GH, VT and MH scales are ‘bipolar’ in that they measure a wide range of health status from the very poor to the very good and the majority of subjects, even without any limitations or disability, would score in the middle range [9]. There were few floor effects, except with the RE scale, suggesting that the Chinese (HK) SF-36 would be able to detect any deterioration in the health status of these subjects.

Since our sample was small, no adjustment was made for the comparison of our results to those of the US samples. The shape of the SF-36 profile of our patients was very similar to that of the U.S. patients, suggesting conceptual equivalence [10,23]. Most of the Hong Kong means were better than those of American patients, probably because our patients were younger (mean age = 48.7 years) than the U.S. patients (mean age =54 years), and some of our patients did not have any chronic diseases. The Hong Kong GH, RE and MH means were lower than those of the U.S. samples. The low GH mean was the result of the low mean scores of items GH1 (general health) and GH5 (my health is excellent) for reasons discussed earlier. The lower RE and MH means were also observed among Chinese people living in the U. S. by Ren et al. [12]. This could indicate a worse mental health status among the Chinese, but this could also be related to how good mental health is defined in the Chinese culture. The mental state most commonly wanted by the Chinese in Hong Kong is feeling peaceful and contented [21]. It is considered a virtue in the Chinese culture if ‘one is the first person to be worried but the last to be happy ‘. Further cognitive testing and qualitative studies with the Chinese would be valuable to the understanding of their interpretation of the meaning and expectation on general and mental health.

Conclusions

The Chinese (HK) version of the SF-36 was found to be equivalent in concepts to the original US-English SF-36. It was well understood and accepted by Chinese people with different demographic characteristics in Hong Kong. Our results confirmed the scaling assumptions and construct validity of the Chinese (HK) SF-36. Therefore, the Chinese (HK) SF-36 scale can be scored using the standard scoring algorithms [9].

The Chinese (HK) SF-36 has completed the first two stages of the IQOLA Project evaluation process, the form is ready for use on the Chinese in Hong Kong. We should proceed to the final stage of the cross-cultural validation process that includes testing and norming of the Chinese (HK) SF-36 on the general population to obtain a reference basis for interpretation and standardization. If our results are replicated and the third stage (validation and norming) is successful, then we will have a valid and standardized HRQOL measure to assess the burden of illnesses and the effectiveness of treatments for our population. This will help to make health services in Hong Kong more patient-centered and cost-effective.

The SF-36 is becoming a standard HRQOL measure in health surveys, service evaluation and clinical trials in the U.S. and other countries. We hope our study will encourage the adaptation and testing of the SF-36 on Chinese populations in other parts of the world, so that it can eventually be applicable to nearly a quarter of the world's population. The breakdown of the language barrier will enable Chinese subjects to take part in national and international HRQOL studies.

References

1. Census and Statistics Department. Hong Kong 1991 Population Census - Main Report. Hong Kong: Government Printer, 1993.
2. Lohr KN. Outcome measurement: concepts and questions. Inquiry 1988; 25:37-50.
3. Greenfield S, Nelson EC. Recent developments and future issues in the use of health status assessment measures in clinical settings. Med Care 1992; 30(Supp): MS23-41.
4. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. JAMA 1995; 273: 59-65.
5. Idler EL, Kasl SV, Lemke JH. Self-evaluated health and mortality among the elderly in New Haven, Connecticut, and Iowa and Washington counties, Iowa, 1982-1986. Am J Epidemiol 1990;131: 91-103.
6. Lam CLK. Health outcome of stroke patients in Hong Kong. Huisarts en Wetenschap 1995; 38: 129-31.
7. Lam CLK, Tse MHW. A study of patients' subjective perception of their health status. HK Pract. 1988; 10: 3291-4.
8. Bowling A. Measuring health - A Review of Quality of Life Measurement Scales. Philadelphia: OUP; 1991

9. Ware JE, Snow KK, Kosinski M,Gandek B. SF-36 Health Survey - Manual and Interpretation Guide. Boston, MA: The Health Institute, the New England Medical Center, 1993.

10. Ware JE, Keller SD, Gandek B, Brazier JE, Sullivan M. Evaluating translations of health status questionnaires: Methods from the IQOLA Project. Int J Tech Ass Health Care 1995; 11: 525-51.

11. Ware JE, Sherbourne CD. The MOS 36-Item Short Form Health Survey (SF-36): I. Conceptual framework and item selection. Med Care 1992; 30: 473-83.

12. Ren XS, Amick B, Zhou L, Gandek B, Ware JE. Chinese version of the SF-36 Health Survey: a review of translation and report on the psychometric testing results . J Clin Epidemiol 1998;

13. General Registrar Office. Registrar General's Classification of Occupation. London: General Registrar's Office; 1966.

14. Hays RD, Hayashi T, Carson S, Ware JE. User's Guide for the Multitrait Analysis Program (MAP). Santa Monica CA: RAND Corporation; 1988: RAND Note N-2786-RC.

15. McHorney CA, Ware JE, Lu R, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Medical Care 1994; 32: 40-66.

16. Likert R. A technique for the measurement of attitudes. Arch of Psychol 1932; 140: 5-55.

17. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol Bull 1980; 87: 245-51.

18. Nunnally JC. Psychometric Theory, 3rd edition. New York: McGraw Hill; 1994.

19. McHorney CA, Ware JE, 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;31:247-63.

20. Ware JE, Kosinski M, Keller SD. SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: The Health Institute, New England Medical Center; 1994.

21. Lam CLK. Do the COOP/WONCA Charts measure important functions? - The Chinese perspective. in HKCGP, 14th WONCA World Conference Book of Abstracts. Hong Kong: HKCGP; 1995: p.45.

22. Ware JE, Kosinski M, Keller SD. SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. Boston, MA: The Health Institute, New England Medical Centre; 1995.

23. Herdman M, Fox-Rushby J, Badia X. 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res 1997; 6: 237-47.

Acknowledgments

This study was funded by the Committee of Research and Conference Grant, the University of Hong Kong. We are grateful to Dr. William Lin, Miss Ee San Chen and Mr. Shaoming Lian for reviewing the translation. We would like to thank our research assistants, Cyrina Chan and Ada Au, for their help in data collection and entry. Thanks also go to the staff of the IQOLA Project, the Health Institute, New England Medical Center, Boston, USA, for their assistance in data analysis.

Conflict of Interest: None

Table 1: Number (%) of Subjects Who Had Problems with the Questions of the Chinese (HK) SF-36 (n=236)

Question Number	Do not Understand	Difficult	Not Relevant	Mind Answering
1 (General Health)	2 (0.8)	3 (1.3)	7 (3.0)	2 (0.8)
2 (Health Transition)	2 (0.8)	1 (0.4)	9 (3.8)	2 (0.8)
3 (Physical Activities)	0	1 (0.4)	26 (11)	1 (0.4)
4 (Role-physical Function)	1 (0.4)	0	27 (11.4)	1 (0.4)
5 (Role-emotional Function)	2 (0.8)	2 (0.8)	12 (5.1)	2 (0.8)
6 (Extent of Social Activities)	1 (0.4)	0	35 (14.8)	1 (0.4)
7 (Degree of Bodily Pain)	2 (0.8)	1 (0.4)	21 (8.9)	1 (0.4)
8 (Limitation from Bodily Pain)	1 (0.4)	1 (0.4)	19 (8.1)	2 (0.8)
9 (Vitality/ Mental Health)	2 (0.8)	2 (0.8)	13 (5.5)	2 (0.8)
10 (Frequency of Social Activities)	0	1 (0.4)	31 (13.1)	2 (0.8)
11 (General Health)	5 (2.1)	3 (1.3)	12 (5.1)	3 (1.3)

Table 2: Distribution of Mean (SD) Item Scores in the Hypothesized Item-Cluster Order

SF-36 Items (question number)		**Mean Item Score (SD)**	
Clustered in the hypothesized order		**Hong Kong (n=236)**	**American Norm (n=2227)** [10]
PF	1 (3a) Vigorous activities	2.04 (.81)	2.17
	4 (3d) Climbing several flights	2.69 (.63)	2.54
	6 (3f) Bending, kneeling or stooping	2.70 (.64)	2.59
	7 (3g) Walking more than 1 Km	2.58 (.75)	2.55
	2 (3b) Moderate activities	2.76 (.56)	2.65
	3 (3c) Lifting/carrying groceries	2.95 (.23)	2.72
	8 (3h)Walking several blocks	2.88 (.39)	2.69
	5 (3e) Climbing one flight	2.98 (.14)	2.78
	9 (3i) Walking one block	3.00 (.14)	2.82
	10 (3j) Bathing or dressing	2.99 (.13)	2.88
RP	2 (4b) Accomplished less	1.58 (.49)	1.73
	1 (4a) Cut down time on work	1.65 (.48)	1.83
	3 (4c) Limited in kind of work	1.63 (.48)	1.78
	4 (4d) Difficulty performing work	1.58 (.49)	1.77
BP	1 (7) Intensity of bodily pain	4.37 (1.55)	4.78
	2 (8) Extent pain interfered with work	4.67 (1.34)	4.58
GH	1 (1) Your health is: excellent.....poor	2.60 (.99)	3.77
	3 (11b) As healthy as anybody	3.38 (1.06)	3.80
	5 (11d) Health is excellent	2.81 (1.09)	3.72
	4 (11c) Expect health to get worse	3.35 (1.11)	3.66
	2 (11a) Seem to get sick a little easier	3.53 (1.08)	4.19
VT	1 (9a) Feel full of pep	3.11 (1.47)	3.82
	2 (9e) Have a lot of energy	3.34 (1.38)	3.82
	3 (9g) Feel worn out	4.45 (1.39)	4.34
	4 (9i) Feel tired	4.16 (1.34)	4.02
SF	2 (10) Frequency social act. interfered	4.33 (.99)	4.25
	1 (6) Extent social act. interfered	4.46 (.92)	4.35
RE	2 (5b) Accomplish less	1.51 (.50)	1.75
	1 (5a) Cut down amount of time on work	1.56 (.50)	1.84
	3 (5c) Didn’t do work as carefully	1.42 (.49)	1.82
MH	3 (9d) Felt calm & peaceful	3.95 (1.38)	4.06
	5 (9h) Been a happy person	3.99 (1.41)	4.43
	1 (9b) Been very nervous	4.44 (1.38)	4.85
	2 (9c) Felt down in the dumps	4.85 (1.17)	5.33
	4 (9f) Felt downhearted & blue	4.64 (1.19)	4.98
HT	(2) Health compared to 1 year ago	2.76 (.84)	3.14

Table 3: Pearson Item-Scale Correlations[a] (Significance of Difference[b])

Item	PF	RP	BP	GH	VT	SF	RE	MH	HT
PF1	.55*	.32 (2)	.22 (2)	.27 (2)	.14 (2)	.09 (2)	-.01 (2)	.01 (2)	.17 (2)
PF2	.60*	.26 (2)	.23 (2)	.26 (2)	.13 (2)	.13 (2)	.10 (2)	.12 (2)	.11 (2)
PF3	.35*	.15 (2)	.18 (2)	.08 (2)	.05 (2)	.07 (2)	.02 (2)	.03 (2)	.07 (2)
PF4	.56*	.26 (2)	.29 (2)	.27 (2)	.16 (2)	.11 (2)	.08 (2)	.13 (2)	.16 (2)
PF5	.31*	.13 (2)	.11 (2)	.05 (2)	.03 (2)	.09 (2)	.06 (2)	.05 (2)	-.01 (2)
PF6	.61*	.31 (2)	.27 (2)	.23 (2)	.19 (2)	.08 (2)	.02 (2)	.04 (2)	.17 (2)
PF7	.69*	.29 (2)	.22 (2)	.26 (2)	.30 (2)	.13 (2)	.06 (2)	.05 (2)	.14 (2)
PF8	.63*	.21 (2)	.17 (2)	.10 (2)	.16 (2)	.03 (2)	.04 (2)	-.02 (2)	.09 (2)
PF9	.00*	.00 (1)	.00 (1)	.00 (1)	.00 (1)	.00 (1)	.00 (1)	.00 (1)	.00 (1)
PF10	.06*	.02 (1)	.10 (-1)	.12 (-1)	.09 (-1)	.03 (1)	.08 (-1)	.12 (-1)	.06 (-1)
RP1	.21 (2)	.61*	.27 (2)	.29 (2)	.27 (2)	.30 (2)	.24 (2)	.14 (2)	.12 (2)
RP2	.32 (2)	.73*	.29 (2)	.28 (2)	.30 (2)	.28 (2)	.35 (2)	.18 (2)	.19 (2)
RP3	.35 (2)	.64*	.36 (2)	.22 (2)	.25 (2)	.35 (2)	.43 (2)	.19 (2)	.14 (2)
RP4	.37 (2)	.64*	.31 (2)	.28 (2)	.30 (2)	.21 (2)	.24 (2)	.07 (2)	.19 (2)
BP1	.32 (2)	.30 (2)	.77*	.26 (2)	.26 (2)	.24 (2)	.11 (2)	.17 (2)	.19 (2)
BP2	.30 (2)	.42 (2)	.77*	.24 (2)	.26 (2)	.40 (2)	.24 (2)	.18 (2)	.14 (2)
GH1	.32 (1)	.31 (1)	.23 (2)	.39*	.27 (1)	.01 (2)	.05 (2)	.07 (2)	.19 (2)
GH2	.29 (2)	.24 (2)	.15 (2)	.45*	.32 (2)	.09 (2)	-.01 (2)	.18 (2)	.21 (2)
GH3	.18 (2)	.16 (2)	.19 (2)	.55*	.31 (2)	.15 (2)	.04 (2)	.19 (2)	.13 (2)
GH4	.15 (2)	.27 (2)	.15 (2)	.44*	.23 (2)	.19 (2)	.06 (2)	.14 (2)	.34 (2)
GH5	.19 (2)	.15 (2)	.20 (2)	.54*	.36 (2)	.10 (2)	.09 (2)	.25 (2)	.20 (2)
VT1	.17 (2)	.22 (2)	.11 (2)	.27 (2)	.47*	.12 (2)	.14 (2)	.24 (2)	.23 (2)
VT2	.16 (2)	.23 (2)	.16 (2)	.30 (2)	.51*	.22 (2)	.22 (2)	.32 (2)	.15 (2)
VT3	.17 (2)	.29 (2)	.28 (2)	.35 (2)	.58*	.31 (2)	.20 (2)	.43 (2)	.30 (2)
VT4	.25 (2)	.31 (2)	.29 (2)	.40 (2)	.60*	.34 (2)	.26 (2)	.42 (2)	.26 (2)
SF1	.07 (2)	.26 (2)	.30 (2)	.16 (2)	.27 (2)	.48*	.33 (2)	.40 (1)	.13 (2)
SF2	.17 (2)	.34 (2)	.27 (2)	.12 (2)	.29 (2)	.48*	.35 (2)	.39 (1)	.05 (2)
RE1	-.02 (2)	.27 (2)	.23 (2)	.03 (2)	.17 (2)	.40 (2)	.67*	.37 (2)	.10 (2)
RE2	.15 (2)	.38 (2)	.13 (2)	.05 (2)	.24 (2)	.30 (2)	.67*	.31 (2)	.14 (2)
RE3	.04 (2)	.30 (2)	.10 (2)	.11 (2)	.26 (2)	.28 (2)	.49*	.22 (2)	.14 (2)
MH1	.01 (2)	.21 (2)	.18 (2)	.21 (2)	.37 (2)	.35 (2)	.26 (2)	.53*	.06 (2)
MH2	.08 (2)	.22 (2)	.21 (2)	.22 (2)	.40 (2)	.46 (2)	.41 (2)	.69*	.24 (2)
MH3	.03 (2)	.05 (2)	.10 (2)	.08 (2)	.23 (2)	.22 (2)	.23 (2)	.43*	.02 (2)
MH4	.11 (2)	.16 (2)	.17 (2)	.15 (2)	.38 (2)	.40 (2)	.28 (2)	.65*	.22 (2)
MH5	.08 (2)	.01 (2)	.02 (2)	.22 (2)	.33 (1)	.25 (2)	.16 (2)	.45*	.06 (2)
HT	.20	.20	.17	.31	.32	.10	.15	.16	---*

* Correlation between an item and its hypothesized scale.

a: Item-scale correlation corrected for overlap for coefficients followed by *. Standard error = 0.07.

b: Item-hypothesized scale correlation is (significantly higher = 2; higher = 1; lower = -1; significantly lower = -2) than item-competing scale correlation, as determined by Steiger's t-test (17).

Table 4: Internal Reliability Coefficients (R_{TT}) and Inter-scale Correlations

		Scale-Scale Correlation								
Scale	**R_{TT}**	**PF**	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**	**HT**
PF	.78									
RP	.83	.38								
BP	.87	.33	.38							
GH	.71	.33	.33	.27						
VT	.74	.25	.34	.28	.43					
SF	.65	.14	.35	.33	.16	.33				
RE	.77	.07	.39	.18	.07	.27	.40			
MH	.77	.08	.18	.18	.24	.47	.45	.36		
HT	---	.20	.20	.17	.31	.32	.10	.15	.16	

Table 5: Correlation (r) between Scale Scores and Rotated Principal Components

	Hypothesized Association		Sample (n=236)	
	Physical Component	Mental Component	Physical Component	Mental Component
PF	●	O	.78	-.08
RP	●	O	.63	.34
BP	●	O	.61	.23
GH	ω	ω	.70	.10
VT	ω	ω	.48	.51
SF	ω	●	.20	.74
RE	O	●	.05	.74
MH	O	●	.10	.77

● Strong association (r ≥ .70)

ω Moderate to substantial association (.30 < r < .70)

O Weak association (r ≤ .30)

Table 6: Comparison between SF-36 Scale Scores of Hong Kong and U.S. Subjects [9,15]

Scale	Transformed Score Mean (SD)	% Floor	% Ceiling
PF			
Hong Kong (n=185)	85.95 (15.44)	0.0	21.6
US patients (n=3445)	73.21 (26.59)	0.8	19.2
US norm (n=2474)	84.15 (23.28)	0.8	38.8
RP			
Hong Kong (n=185)	57.97 (39.54)	20.5	36.2
US patients (n=3445)	56.69 (40.52)	24.3	36.7
US norm (n=2474)	80.96 (34.00)	10.3	70.9
BP			
Hong Kong (n=185)	70.42 (27.76)	2.2	34.6
US patients (n=3445)	68.67 (25.10)	0.9	17.8
US norm (n=2474)	75.15 (23.69)	0.6	31.9
GH			
Hong Kong (n=185)	51.46 (18.29)	1.6	0.0
US patients (n=3445)	60.04 (21.25)	0.2	1.4
US norm (n=2474)	71.95 (20.34)	0.0	7.4
VT			
Hong Kong (n=185)	55.76 (20.94)	2.2	1.1
US patients (n=3445)	53.51 (22.05)	1.1	0.9
US norm (n=2474)	60.86 (20.96)	0.52	1.5
SF			
Hong Kong (n=185)	85.27 (20.17)	0.5	54.1
US patients (n=3445)	80.53 (24.33)	0.9	46.3
US norm (n=2474)	83.28 (22.69)	0.6	52.3
RE			
Hong Kong (n=185)	52.79 (40.75)	29.7	32.4
US patients (n=3445)	68.83 (39.66)	18.1	56.1
US norm (n=2474)	81.26 (33.04)	9.6	71.0
MH			
Hong Kong (n=185)	68.65 (19.09)	1.1	3.2
US patients (n=3445)	71.44 (21.10)	0.1	4.4
US norm (n=2474)	74.74 (18.05)	0.0	3.9