Translation and calibration of a Chinese version of the Sickness Impact Profile for use in Hong Kong

TG Short, MY Rowbottom, JPF Lau, GWY Lai, TA Buckley, TE Oh

Because of the lack of suitable generic health status measures in Hong Kong that reflect the value structure and culture of the society, we have translated and calibrated the Sickness Impact Profile, which was originally developed in the United States. Translation consisted of a sequence of forward translations into Chinese, back translations into English and, when there was a loss of meaning or conceptual equivalence, retranslation into Chinese. Sixty Hong Kong Chinese people, including health professionals, patients, and members of the public were then recruited to determine item and dimension weights for the Hong Kong population. Individual ratings were averaged to obtain a consensus value for each item. Within-category reliability analysis for scores varied from 0.70 to 0.92 (Cronbach's alpha values) and overall internal consistency was 0.98. There were only small differences between this version and the original American English version in the ratings given to individual questions (Pearson's product moment correlation coefficient, r=0.80). We have developed a Chinese translation of the Sickness Impact Profile which can now be used to evaluate health outcomes in Hong Kong and to compare outcomes with studies in other populations where the Sickness Impact Profile was used.

HKMJ 1998;4:375-81

Key words: Health status; Health surveys; Sickness Impact Profile; Translations

Introduction

An evaluation of the effectiveness of a health service requires measures of health status that reflect the value structure and culture of the society in which the service is based. There is a lack of such health status measures suitable for use in Hong Kong. To conduct patient outcome studies at the Prince of Wales Hospital we have translated the Sickness Impact Profile (SIP) into Chinese and calibrated it for use in a Hong Kong Chinese patient population.

The SIP was originally developed in the United States in the 1970s.[1-5] As it is a generic health status measure, it is suitable for use in a wide range of patients. The SIP consists of 136 yes/no-type questions grouped into the following 12 broad categories: sleep and rest, emotional behaviour, body care and movement, home management, mobility, social interaction, ambulation, alertness behaviour, communication, work, recreation and pastimes, and eating. Examples of questions from the questionnaire are given in the Box and Tables 1, 2, and 3. The score can vary from 0 to 100, with a higher score meaning a more severe degree of disability. A general American population has been found to have a mean SIP score of 3.5.[6] Some relevant scores from past studies of patients with stable, chronic diseases include the following: angina, 8; rheumatoid arthritis, 16; non–oxygen-dependent chronic obstructive airways disease (COAD), 17; and oxygen-dependent COAD, 24.[6] The SIP is currently regarded as the best measure of quality of life and functional health status,[7] and was chosen by us because it has been used extensively overseas.

The questions were chosen from an initial bank of 300 items describing sickness-related behavioural changes.[2] The questionnaire has been tested for its construct validity, clinical validity, reliability, sensitivity, and administrative feasibility. It has been used for patients with a broad range of conditions and found to be a reliable indicator of health status, a sensitive indicator of changes in health status, and a valid measure of the impact of different diseases on an

Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, Hong Kong:
Department of Anaesthesia and Intensive Care
TG Short, MD, FANZCA
MY Rowbottom, RN
TA Buckley, MB, ChB, FHKAM (Anaesthesiology)
TE Oh, FANZCA, FRACP (Edin)
Centre for Clinical Trials and Epidemiological Research
JPF Lau, PhD
GWY Lai, MSc

Correspondence to: Dr TG Short

Examples of questions from each category of the American English version of the SIP, with Chinese translations*

Examples of translations from each of the 12 categories

Sleep and rest:	I spend much of the day lying down in order to rest 日間大部份時間我都臥床休息。
Emotional behaviour:	I have attempted suicide 我曾經企圖自殺。
Body care and movement:	I do not have control of my bowels 我不能控制大便。
Home management:	I am not doing any of the regular daily work around the house that I would usually do 我以前常做的家務，現在已完全沒有做。
Mobility:	I am not now using public transport 我現在沒有使用公共交通工具。
Social interaction:	I am not going out to visit people at all 我現在完全沒有外出探訪親友。
Ambulation:	I do not walk at all 我現在完全沒有步行。
Alertness behaviour:	I forget a lot—for example, things that happened recently, where I put things, appointments 我的記性很差。
Communication:	I am having trouble writing or typing 我書寫或打字有困難。
Work:	I am working shorter hours 我工作的時間較前短。
Recreation and pastimes:	I am not doing any of my usual physical recreation activities 我完全沒有做我以前常做的體能活動及娛樂。
Eating:	I am eating special or different food—for example, soft food, bland diet, low salt, low fat, low sugar 我現在只能吃特別餐，例如：較軟的食物，糖尿餐等等。

* The full version of the Sickness Impact Profile is available from the authors

individual's health status.[6-11] Disadvantages include the fact that it is a long questionnaire (taking approximately 30 minutes to complete) and it is not sensitive to very small degrees of disability in essentially healthy patients. The SIP is considered to be an absolute measure of health status—that is, a patient's health status is compared to a standard determined by the society in which the patient lives. To establish the standard, individual items in the questionnaire are weighted by a panel of local judges as to their importance. An alternative approach is to use a questionnaire that asks for comparisons with a patient's previous health or self-perception of current health. An example of such a questionnaire is the RAND 36-Item Health Survey (SF-36).[8] This approach can be more sensitive to measuring a change when it is possible to measure a patient's health status before making an intervention, but it has the disadvantage of not clearly showing the significance of any measured changes in health status.

To fulfil the requirements of an effective generic health status measure, patients must be questioned in their own language. The SIP has been translated into many languages and calibrated for use in different cultures so that cross-cultural comparisons of health status and outcome could be made.[6,7,9-11]

Subjects and methods

The translation procedure

The procedures used to revise a Hong Kong Chinese version were those recommended by the International Advisory Committee to the SIP.[12] These guidelines included a description of the methodology to be followed and the number of subjects needed, to create a valid translation of the questionnaire. Approval for making our translation and copies of the approved procedures were obtained from the holders of the copyright to the original version (Johns Hopkins University,

Table 1. Outlier items (on a 15-point scale), where there was a low level of agreement between judges as to the importance of the items to the Sickness Impact Profile

Item	Outlier items	Mean (standard deviation)
SI-19	I am not doing the things I usually do to take care of my children or family	8.9 (4.1)
C-1	I am having trouble writing or typing	7.2 (4.2)
C-5	I don't write except to sign my name	7.9 (4.2)
W-1	I am not working at all	10.5 (4.0)

Table 2. Comparison of the most dysfunctional and least dysfunctional items (on a 15-point scale) in each category together with the American score for these items

Item	Most dysfunctional feature	HK score	US score
SR-1	I spend much of the day lying down in order to rest	10.1	8.3
EB-4	I have attempted suicide	11.6	13.2
BCM-21	I do not have control of my bowels	12.3	12.8
HM-3	I am not doing any of the regular daily work around the house that I would usually do	7.2	8.6
M-2	I stay within one room	10.1	10.6
SI-15	I have frequent outbursts of anger at family members, eg...	9.1	11.9
A-6	I do not walk at all	10.9	10.5
AB-6	I sometimes behave as if I am confused or disorientated in time or place, eg...	9.8	11.3
C-2	I communicate mostly by gestures, eg...	9.4	10.2
W-1	I am not working at all	10.5	36.1
RP-4	I am not doing any of my usual inactive recreation or pastime activities, eg...	7.7	8.4
E-9	I am eating no food at all; nutrition is taken through tubes or intravenous fluids	12.5	13.3
	Least dysfunctional feature		
SR-6	I sleep less at night, eg...	6.4	6.1
EB-2	I laugh or cry suddenly	8.0	6.8
BCM-12	I change position frequently	5.4	3.0
HM-2	I am doing less of the regular daily work around the house than I would usually do	4.9	4.4
M-8	I am not going into town	5.0	4.8
SI-7	I am cutting down the length of visits with friends	5.0	4.3
A-12	I walk more slowly	4.3	3.5
AB-2	I have more minor accidents, eg...	7.3	7.5
C-4	I often lose control of my voice when I talk, eg...	6.3	8.3
W-5	I am working shorter hours	6.7	4.3
RP-6	I am doing fewer community activities	4.2	3.3
E-1	I am eating much less than usual	3.9	3.7

Table 3. Comparison of the American and Chinese versions of the Sickness Impact Profile: outlier items, and residuals (observed-predicted) for final scores

Item	Outliers	HK score	US score	Residual
EB-5	I act nervous or restless	8.2	4.6	3.6
BCM-2	I do not move into or out of bed or chair by myself but am moved by a person or mechanical aid	8.6	12.1	-3.5
M-5	I am not now using public transportation	6.7	4.1	2.6
SI-2	I am not going out to visit people at all	6.6	10.1	-3.5
SI-8	I am avoiding social visits from others	5.4	8.0	-2.6
SI-15	I have frequent outbursts of anger at family members, eg...	9.1	11.9	-2.8
SI-18	I refuse contact with family members, eg...	8.4	11.5	-3.1
SI-20	I am not joking with family members as I usually do	7.3	4.3	3.0
AB-9	I make more mistakes than usual	9.0	6.4	2.6
W-1	I am not working at all	10.5	36.1	-25.6
W-2	I am doing part of my job at home	7.4	3.7	3.7
W-8	I am working at my usual job but with some changes, eg...	8.3	3.4	4.9
W-9	I do not do my job as carefully and accurately as usual	9.0	6.2	2.8
E-3	I am eating special or different food, eg...	7.4	4.3	3.1

Baltimore, US). In Hong Kong, Cantonese is the spoken language and Mandarin (Putonghua) Chinese is the written language. The translation exercise was designed to produce a formal Chinese version of the SIP that would be appropriate for use in Hong Kong.

Initially, two independent forward translations were made; items were translated so that the original concept was retained. Both translators were native Cantonese speakers resident in Hong Kong and who had qualifications in English-to-Chinese translation. The two translations were assigned quality ratings for clarity, common language, and conceptual adequacy by two other independent translators who also spoke Cantonese as a first language. The translations were pooled and a single first translated version created. This version was then back-translated from Chinese into English by two native American English translators who also spoke fluent Cantonese. The back translations were then assigned quality ratings and compared with the original SIP by native English speakers.

Whenever there was a loss of meaning or conceptual equivalence in the back translation, the items were discussed and re-translated into Chinese, so that a second translated version was created, which consisted of the 136 items of interest.

The second translated version was given to a heterogeneous group of 20 individuals including patients, health professionals, and members of the public to identify items that were considered unclear, irrelevant, upsetting, or otherwise problematic. Then, following a further revision, a final Chinese translation was made.

The pilot test

The Chinese translation of the SIP was then tested to determine item and dimension weights for the Hong Kong Chinese population. Sixty people—20 doctors, nurses, and allied health professionals, 20 patients, and 20 members of the public—were recruited to be the judges. All questionnaires were administered by two interviewers who worked together initially, to ensure that they used a similar style when administering the questionnaire. Judges were asked to rate the severity of the dysfunction in each item without regard to the cause and without reference to the way they had distributed items in other categories.

At first, items within each of the 12 categories were scaled on an 11-point equal interval scale with the extremes marked 'least dysfunctional' and 'most dysfunctional' to grade the relative value of the items within each dimension. On completion of each category, judges were asked to review where they placed the items to ensure that all items were correctly placed in accordance with their view of the degree of dysfunction that each item represented. Then the items with the highest and the lowest score within each category were grouped and scaled separately by the same judge on a 15-point equal interval scale to establish the relative importance of items within each category. This two-stage procedure was used in the creation of the original questionnaire.[4,5]

The test-retest reliability of the Hong Kong Chinese version of the SIP was assessed by administering the questionnaire twice, 24 hours apart, to the same 20 patients residing in a home for the young chronically disabled. The patients were of both sexes, aged from 20 to 65 years, and all had significant degrees of disability. The two interviewers were randomly assigned to patients to carry out the task; administration of the SIP was performed by the same interviewer on both occasions. Ten of these patients were part of the previously mentioned group who determined item weights for the questionnaire

Statistical analysis

Statistical testing of the resulting questionnaire was performed according to standard procedures for the design of health surveys and was also similar to the testing procedures used in the original American version of the SIP.[1-5,11,13] A two-step direct scaling procedure was used to obtain consensus values for the importance of each item to the overall SIP. Firstly, the 60 individual ratings obtained for each item were averaged. These averaged values were then scaled, using the weights obtained in the second part of the judging procedure for the highest and lowest scoring items within each category, to give a final value. This two-step process ensured that all items in the questionnaire were given a weight in the overall questionnaire appropriate to their importance in indicating disability. Internal consistency reliability was assessed using the linear correlation between answers to the questions within each category. This measure assumes that there will be a correlation between answers to different questions about the same concept. Outlying items, where there was poor agreement between judges as to their scale values, were identified and scrutinised.

To assess internal consistency, Cronbach's alpha values were calculated for each of the 12 categories using the scale values obtained in the first stage. This estimate calculates the degree of equivalence between answers to sets of similar questions. A reasonable degree of correlation (greater than 0.7) ensures that the questions are measuring the same concept. An overall estimate of the Hong Kong Chinese version of the SIP using the Cronbach's alpha test was also computed to assess the internal consistency of each category in the questionnaire. This estimate of internal consistency was based on the scale values obtained on the 15-point scale to rank the highest and lowest scoring items within each category. Test-retest reliability was assessed by determining Pearson's product moment correlation coefficient (r). A paired, two-tailed Student's *t* test was performed to detect any difference between the values given to items by health professionals, patients, and potential patient groups.

The total score that results from the scaling procedure is not fixed. To be able to make direct comparisons between the scale values of the American English and Hong Kong Chinese versions of the SIP, the total SIP scores for the Hong Kong SIP were rescaled so that the total score was the same as for

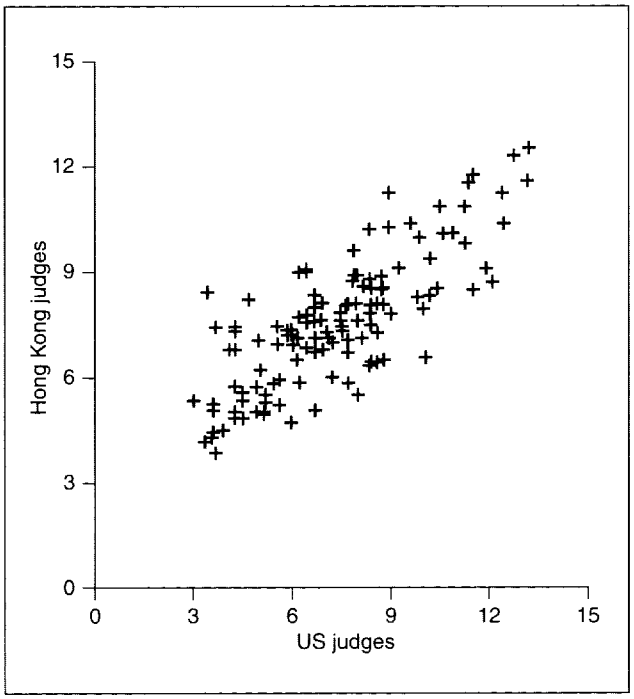

Fig. Comparison of the scale values assigned to 125 items in the Sickness Impact Profile by the Hong Kong and American judges
Pearson's product moment correlation coefficient, r=0.80
Equation of line of best fit: y=0.61x+3.063

the American version but the individual weights assigned to each question were those of the Hong Kong version. The association between the weights assigned to each item in the Chinese and American versions were then assessed by computing the correlation coefficient between the two sets of scale values. Outlier items were defined as those items where the difference between the two scores was more than two standard deviations from the mean difference.

Results

Examples of the original American English questions and their Hong Kong Chinese equivalents taken from each category of the SIP are shown in the Box. Values for the reliability analysis (Cronbach's alpha values) for within-category scorings are listed in Table 4; alpha values varied from 0.70 to 0.92. There were no scores below 0.70, indicating that internal consistency of the questions within each category was maintained in the translation. The estimates of overall internal consistency, based on the scale values for each category obtained on the 15-point scale, ranged from 0.82 to 0.92. Overall internal consistency of the SIP was 0.98. Again, overall reliability of the questionnaire was maintained in the translated version. Outlier items, which reflect poor agreement between judges on the score for the individual items are listed in Table 1; only four items in the profile fell into this category. The number of outlying items was regarded as satisfactory, given the large number of questions in the questionnaire and the satisfactory scores for Cronbach's alpha testing.

The items ranked as most dysfunctional and least dysfunctional within each category, together with their American score, are listed in Table 2. The results highlight the relative importance placed on questions in the questionnaire by the judges. Test-retest coefficients for the 20 individuals who completed the scaling exercise twice varied from 0.56 to 1.00 (mean, 0.75). This result is similar to the test-retest coefficient found for the original questionnaire.[4] There was a small, yet statistically significant difference between

Table 4. Reliability analysis for within-category scorings between observers, within-category correlations with the American scorings, and percentage contribution to the overall score for each category in the Hong Kong and American versions of the Sickness Impact Profile

Item	Reliability of Hong Kong version*	Correlation with US scores†	% contribution	
			Hong Kong	US
Sleep and rest	0.72	0.84	5.4	4.9
Emotional behaviour (EB)	0.70	0.81	7.7	6.9
Body care and movement (BCM)	0.92	0.86	19.3	19.5
Home management	0.86	0.79	6.0	6.5
Mobility (M)	0.86	0.88	7.0	7.0
Social interaction (SI)	0.90	0.70	12.8	14.1
Ambulation (A)	0.87	0.98	8.7	8.2
Alertness behaviour (AB)	0.85	0.54	8.2	7.6
Communication (C)	0.86	0.78	6.8	7.0
Work	0.92	0.78	6.8	7.6
Recreation and pastimes	0.82	0.95	4.1	4.1
Eating	0.84	0.94	7.2	6.9
Physical dimension (A+M+BCM)			35.0	34.7
Psychosocial dimension (SI+C+AB+EB)			35.5	35.6

* Cronbach's alpha values
† Pearson's product moment correlation coefficients

the mean scale values obtained for health professionals and patients versus potential patients. The mean scale values were 8.3 and 7.9, respectively, giving a difference of 0.4 (95% confidence interval, 0.21-0.59; P<0.001).

A comparison between scores obtained for each item in the Chinese and American versions of the SIP is shown in the Figure and the contribution of each category to the overall scores in Table 4. The overall correlation (r) between the two versions was 0.80. A comparison with the American version revealed only small differences in the contribution of individual categories to the overall SIP score. For instance, questions concerning emotional behaviour had higher scores in the Hong Kong version, while questions concerning social interaction and work had higher ratings in the American version. However, within each category there were large differences in the scores assigned to individual items. Scores from the American version for the most and least important questions within each category are shown in Table 2. Outlier items, defined as a difference of greater than two standard deviations from the mean difference in scores, for the American and Chinese versions are outlined in Table 3. These items demonstrate cultural differences between the two societies, but the many questions in the survey and the high degree of correlation in overall scores for each dimension indicate that only small differences in the SIP score would be observed in patients using the two different scales. It was concluded that the translation exercise resulted in a Chinese version of the SIP that has similar metric properties to the original American questionnaire, in terms of measuring functional health status.

Discussion

The SIP was originally formulated as a generic health status measure for use in health surveys, programme planning, policy formation, and in monitoring patient progress in terms of sickness.[3] Translation of the SIP into Chinese and calibration for use in the Hong Kong Chinese population will help doctors assess the health status of patients and to make valid comparisons between the health status of patients in the local population and those from other countries where the SIP has been used.

The internal consistency and the test-retest reliability of the Chinese version was found to be similar to other versions of the SIP.[4,7,11] There were only four items in the questionnaire where agreement between judges as to their importance was found to be poor. These items were retained, however, to maintain the integrity of the questionnaire. One limitation of the current adaptation is that only items in the original SIP have been translated, so the possibility that other sickness-related behaviours unique to the Hong Kong patient population exist has not been explored. However, attempts to elicit additional sickness-related behaviours in other cultures have not found any that were not covered by the original questionnaire.[10] This is partly explained by the large number of questions in the SIP, which means that most aspects of health function universal to humankind have been covered. But considerable differences in weight placed on some items between the American and Hong Kong versions means there are important differences in perception of health between the two cultures. The differences in weight placed on individual questions also indicates that ample opportunity for these differences to be expressed is given by the questionnaire.

By retaining the integrity of the original questionnaire, the construct validity, reliability, and sensitivity to change of the SIP should also have been maintained in our version. Although the clinical validity of the SIP in its Chinese form has not yet been tested, it is very unlikely it will be altered by the translation and cultural adaptation process. We are currently in the process of testing its clinical validity in association with clinical outcome studies in a local patient population.[14]

The calibration exercise provides an opportunity to observe cross-cultural differences regarding the importance of dysfunctions related to health. Unfortunately, with the exception of the original American version, there is only limited documentation in the literature of the actual weightings obtained with other translations. Unlike the American and Chicano Spanish versions, only small differences were found for the importance of different categories of dysfunction to the overall SIP score.[4,11] When examining individual items, however, some important differences emerged, particularly in relation to attitudes toward social interaction and work (Tables 2 and 3). Given the small impact of individual questions on the overall SIP score, it is unlikely that major differences in sickness-related behaviour—and therefore, the overall SIP score—would be observed for patients with similar medical problems within the two cultures.

Hong Kong is an affluent, westernised, urban society that has a well-developed western medical system in place. While the written language is Mandarin Chinese, the spoken language is a dialect—Cantonese.

Consequently, the translation into Chinese should be appropriate for use in other regions of China but it cannot be assumed that the weightings obtained in Hong Kong would also apply. It would be necessary to repeat the calibration exercise before using the SIP elsewhere in China.

Conclusion

We have developed a formal Chinese translation and Hong Kong Chinese cultural weighting of the SIP which is now available for evaluating health outcomes in the Hong Kong Chinese population. By translating the SIP into Chinese and calibrating it to the local population we will be able to make valid comparisons with SIP data obtained from clinical outcome studies in other populations in other countries. Further field testing should be conducted in association with outcome studies to demonstrate the clinical validity of this version in Hong Kong.

Acknowledgements

A Hong Kong Universities Grants Committee Earmarked Grant for Research, No. 221400490, supported this project. The following people also assisted with this project: Drs HY So, J Lee, P Kuo, PT Chui, and Ms E Mak.

References

1. Gilson BS, Gilson JS, Bergner M, et al. The sickness impact profile. Development of an outcome measure of health care. Am J Public Health 1975;65:1304-10.
2. Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS. The sickness impact profile: validation of a health status measure. Med Care 1976;14:57-67.
3. Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. The sickness impact profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv 1976;6:393-415.
4. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787-805.
5. Carter WB, Bobbitt RA, Bergner M, Gilson BS. Validation of an interval scaling: the sickness impact profile. Health Serv Res 1976;3:516-28.
6. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med Care 1989, 27(3 Suppl):217S-232S.
7. de Bruin AF, de Witte LP, Stevens F, Diederiks JP. Sickness Impact Profile: the state of the art of a generic functional status measure. Soc Sci Med 1992;35:1003-14.
8. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. Health Econ 1993,2:217-27.
9. Jacobs HM, Luttik A, Touw-Otten En FW, de Melker RA. The sickness impact profile; results of an evaluation study of the Dutch version [in Dutch]. Ned Tijdschr Geneeskd 1990; 134:1950-4.
10. Patrick DL, Sittampalam Y, Somerville SM, Carter WB, Bergner M. A cross-cultural comparison of health status values. Am J Public Health 1985;75:1402-7.
11. Gilson BS, Erickson D, Chavez CT, Bobbitt RA, Bergner M, Carter WB. A Chicano version of the sickness impact profile (SIP): a health care evaluation instrument crosses the linguistic barrier. Cult, Med Psychiatry 1980;4:137-50.
12. Wu AW. Proposed guidelines for translation/cultural adaptation of the Sickness Impact Profile, Brussels, February 5, 1994. Baltimore: School of Hygiene and Public Health, Johns Hopkins University, 1994.
13. Aday LA. Designing and conducting health surveys. San Francisco: Jossey-Bass Publishers, 1989.
14. Short TG, Buckley TA, Rowbottom MY, Wong E, Oh TE. Long-term outcome and functional health status following intensive care in Hong Kong. Crit Care Med 1998. In press.