Quality of Life Research, 5, pp. 91–100

NHP or SIP—A comparative study in renal insufficiency associated anemia

M. L. Essink-Bot,* P. F. M. Krabbe, H. M. E. van Agt and G. J. Bonsel

Department of Public Health, Erasmus University Rotterdam, The Netherlands (M. L. Essink-Bot, P. F. M. Krabbe, H. M. E. van Agt); Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands (G. J. Bonsel)

In this study we compared the feasibility, internal structure and psychometric characteristics (internal consistency, test–retest reliability, construct validity) of two widely used generic health status measures, i.e. the Nottingham Health Profile (NHP) and the Sickness Impact Profile (SIP) when employed among a sample of patients on renal dialysis (n=63). The NHP was found to be more feasible, i.e. shorter and less difficult, than the SIP. The NHP scales showed somewhat higher levels of internal consistency (mean α=0.67, range=0.39–0.80) than the SIP scales (mean α=0.65, range=0.14–0.82). Test–retest reliability with a 24-hour interval was acceptable for most NHP scales (not available for the SIP in this study). Intercorrelations between the NHP scales were somewhat weaker than those for the SIP, and the expected patterns of scale intercorrelations were largely confirmed. The overall pattern of correlations *between* NHP scales and SIP scales was consistent with expectations, although the correlations were generally rather weak. Correlations between NHP scales and SIP scales and instruments measuring mainly physical functioning (ADL, Karnofsky) were largely as expected. Similarly, correlations between NHP scales and SIP scales and instruments measuring mainly psychological functioning [STAI (anxiety), SDS-Zung (depression)] were also as expected, although here the correlations were weaker for the SIP when compared with the NHP. The Index of Well-being exhibited intra-class correlations >0.3 with one SIP scale and with five out of six NHP scales. Common factor analysis, yielding a two-factor solution with a physical and a mental factor of equal importance, showed the SIP scales to load more on the physical factor, while the NHP scales loaded more on the mental factor. The NHP generally performed better than the SIP in terms of feasibility and internal consistency. Physical functioning is emphasized in the SIP, whereas the emphasis of the NHP lies on mental functioning. The analysis confirmed to some extent the intentions of the constructors of NHP and SIP respectively, i.e. the NHP to be a measure of perceived health and the SIP to be a more functional measure.

Key words: Factor analysis; health status; methodology; Nottingham Health Profile; psychometrics; quality of life; Sickness Impact Profile

Introduction

The assessment of the consequences of disease and treatment on quality of life has gained widespread application. Quality of life in the context of disease and treatment is generally limited to 'health-related quality of life', which is commonly referred to as 'health status'. Health status can be comprehensively operationalized as physical, psychological and social functioning. Examples of applied quantitative health status measurement include the National Health Interview Surveys, research in which the effectiveness of drugs is evaluated, as well as medical technology assessment (MTA) of costly intervention programmes. Data are commonly collected by administering a questionnaire to the subject whose health status is to be measured.

It has become common practice, especially in MTA, to employ a combination of generic instruments with disease and/or domain specific ones. Generic instruments, being comprehensive and non

The authors wish to thank the respondents who participated in this study, Hans Severens MSc and Prof. Eddy KA van Doorslaer PhD for their respective contributions to the present study, as well as the Advisory Board for the Health Research Promotion Programme (Adviesgroep SGO) for supporting financially the Research Programme 'Standardization in Medical Technology Assessment'.

* To whom correspondence should be addressed at Department of Public Health, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands; Tel +10 408.7714; Fax +10 436.6831.

disease-specific, allow for the comparison of results among disease stages, and among different diagnostic categories.

Each of the currently available generic instruments has its own strengths and weaknesses. There is, however, little empirical information available on the relative performance of these instruments. We hope that the present paper will contribute to the existing knowledge base by addressing an empirical comparison of two generic instruments for measuring health status, i.e. the Nottingham Health Profile (NHP) and the Sickness Impact Profile (SIP).

The specific research questions addressed in this study were: (1) How do the NHP and the SIP compare in terms of feasibility? (2) How do the NHP and the SIP compare in terms of reliability? (3) Is there empirical support for the hypothesized structures of the NHP and the SIP in terms of the health status domains being addressed (i.e., construct validity)?

Quantative analyses of patient data were combined with qualitative research of the questionnaires and literature research. For this purpose we could make use of an existing dataset from a group of patients with renal insufficiency who were treated by renal dialysis. The diseases and the intervention have variable consequences for functioning in the physical, psychological and social domains.

Methods

Instruments

The *Nottingham Health Profile* was developed in the 1970s in the United Kingdom as a measure of perceived health for use in population surveys.[1] The NHP (part 1) consists of 38 dichotomous items which are grouped into six scales, labelled respectively Physical Mobility, Energy, Pain, Sleep, Social Isolation and Emotional Reaction. Each scale ranges from 0 (=optimal) to 100. The ultimate score has a profile format. The Dutch adaptation of the NHP used in the current study has been previously tested in several patient populations. Some NHP items are shown in Table 1.

The *Sickness Impact Profile* was developed in the USA between 1972 and 1981 as an instrument to assess the consequences of disease and treatment in functional terms. The 136 items are grouped into twelve scales: sleep and rest, eating, work, home management, recreation and pastime, ambulation, mobility, body care and movement (scores of the latter three may be combined as a physical subscore), social interaction, alertness behavior, emotional behavior, communication (scores of the latter four may be combined as a psychosocial subscore). Apart from a 12-dimensional profile score and the physical and psychosocial subscores, the SIP provides the opportunity to compute a total score. Each score ranges from 0 (=optimal) to 100. In the self-assessment version of the SIP the respondent is requested to tick the statements that apply to him/her in relation to his/her health. The SIP was adapted into Dutch by researchers of the Utrecht Institute for General Practice.[2,3] Some examples of SIP items are shown in Table 2. Data on five additional instruments were used in the investigation of the construct validity of the NHP and the SIP.

The *State-Trait Anxiety Inventory* (STAI) is an American 20-item questionnaire, of which a validated and normed Dutch version is available (ZBV).[4,5] We used the 'state'-part, which measures situational anxiety.[6] The total score ranges from 20 (=no anxiety) to 80.

The *Self-rating Depression Scale* (SDS-Zung) is an American 20-item instrument for measuring depression, with a total score ranging from 25 (=no depressive state) to 100.[7] We used the Dutch version as recommended by the Dutch Psychiatric Society (Vereniging voor Psychiatrie).[8,9]

The *Karnofsky Performance Scale* (or Index) was developed by Karnofsky in 1948 to enable quantification of 'objective' quality of life aspects in the evaluation of drugs against cancer.[10] In the original index, the levels are labelled with figures 0 (=dead), 10 . . . 100 (=optimal). We translated the original USA version

Table 1. Examples of NHP items (Hunt 1986)[21]

NHP Item
I have trouble getting up and down stairs or steps (Physical Mobility)
I'm tired all the time (Energy)
I'm in pain when I walk (Pain)
I'm waking up in the early hours of the morning (Sleep)
The days seem to drag (Emotional Reactions)
I feel that I am a burden to people (Social Isolation)

Table 2. Examples of SIP items (Bergner, 1981)[22]

SIP Items
I sleep or nap during the day (Sleep and Rest SR)
I am eating no food at all, nutrition is taken through tubes or intravenous fluids (Eating E)
I often act irritable toward my work associates (Work W)
I am not doing any of the maintenance or repair work around the house that I usually do (Home management HM)
I am going out for entertainment less (Recreation and pastimes RP)
I walk shorter distances or stop to rest often (Ambulation A)
I stay away from home only for brief period of time (Mobility M)
I am very clumsy in body movements (Body care and movement BCM)
I isolate myself as much as I can from the rest of the family (Social interaction SI)
I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things (Alertness behavior AB)
I act irritable and impatient with myself, for example, talk badly about myself, swear at myself, blame myself for things that happen (Emotional behavior EB)
I am having trouble writing or typing (Communication C)

and adapted it to make it suitable for self-assessment.

Independency with respect to *Activities of Daily Life* (ADL) was assessed by a Dutch instrument asking whether the respondent is able to conduct nine activities independently, and if so, at which effort. The nine activities are listed as: getting in and out of bed, going to the lavatory, washing oneself, dressing, eating and drinking, taking a short walk, taking steps, cycling, shopping and cooking. The summary score ranges from 1–10 (=completely ADL independent).[11]

The *Index of Well-Being* (IWB) is a measure for subjective well-being which was developed for American population surveys with a score range from 2.1–14.7 (= optimal well-being). It was adapted into Dutch.[12]

Patients

We used patients' data from a study to evaluate the effectiveness of erythropietin (EPO) in the treatment of renal insufficiency associated anemia. Questionnaire administration took place around a dialysis session. Before a dialysis session the assessment included completion of a comprehensive questionnaire, which included the NHP but excluded the SIP. The SIP was completed 24 hours later. This second questionnaire also included the NHP in a sample of the patients to investigate test–retest reliability. We did not collect SIP test–retest data because it was considered too burdensome for the patients.

The optimal test–retest interval has to be short enough to preclude a change in health status on the one hand, but long enough to eliminate recollection effects. A change in health status is imaginable between the assessments mentioned above, just preceding dialysis and 24 hours afterwards, respectively. When asked, patients and clinicians generally judged this change as insignificant in relation to the overall health status effects associated with terminal renal insufficiency. Recollection effects can probably be ignored, especially because the NHP was part of a comprehensive questionnaire at the test-assessment.

In the present analyses data were available from 63 patients. Although the study included administration of questionnaires in a longitudinal design, we used data from one administration per patient to prevent introduction of artificial dependence in the data. We had 13 assessments preceding EPO treatment and 50 assessments 1–36 weeks after the start of EPO treatment. The mean age of the respondents was 54 years (sd=16 years, range=21–78 years), 35 (56%) of them were men.

Statistics

Features of score distribution. Mean scores, standard deviations, and the percentages of respondents with the best possible score and the worst possible score, respectively, were computed.

The internal consistency was determined with Cronbach's α-coefficient. An α-coefficient of 0.70 or higher was considered as sufficient for the purpose of group comparisons.[13]

Test–retest reliability was assessed with the intraclass correlation coefficient (ICC). The ICC is a statistic comparable with the conventional Pearson's

correlation coefficient, with level effects between variables being taken additionally into consideration.[14,15] Exact standards for the required magnitude of the reliability coefficient (is the instrument reliable enough?) are difficult to give. A test used for individual judgement should be more reliable than one used for group decisions. Whether a level of test–retest reliability of a test is acceptable for comparisons among groups depends on the size of the group under study: a sample of 1,000 can tolerate a much less reliable instrument than a sample of 10.[16]

The internal structure of the NHP and the SIP was examined with the use of correlation techniques. Matrices of intraclass correlation coefficients (ICCs) between the NHP scales and between the SIP scales, respectively, were computed. For each questionnaire scale, the square root of the mean of the squared ICC between that scale and each of the other scales was computed to summarize the correlation matrix. This statistic was used instead of simply averaging ICCs, in order to retain the interpretation of the squared ICC as the amount of variance shared.

Three approaches were taken to investigate the construct validity of the NHP and the SIP. Firstly, the pattern of ICCs between the scales of the NHP and the SIP were examined. It was hypothesized that those scales that are conceptually related would be strongly correlated, while those scales with less in common would exhibit weaker correlations. Secondly, correlation patterns as observed between the scales of the NHP and the SIP and the STAI, the SDS-Zung, the ADL, the Karnofsky and the IWB were compared with *a priori* hypotheses with respect to these correlation patterns. Thirdly, common factor analysis with varimax rotation was employed to examine the relationships among the elements of the two health status measures and the five additional instruments.

Results

Feasibility

The meaning of the feasibility of questionnaires is not uniformly defined. Some aspects of the NHP and the SIP, considered by the authors to be determinants of 'feasibility' are addressed below.

Item content: The NHP items refer mainly to 'generic' physical and mental actions, including for example walking, standing, bending, sleeping, making contact with others, so that the items are applicable to a broad range of age groups, persons in different phases of their lives, and to both sexes.* The SIP-items refer to a larger extent to activities, including for example tying shoe laces, performing household tasks, lying in bed, performing paid work, visiting friends and caring for children.

Instructions: The SIP instructs respondents to tick the statements which apply to him/her in relation to his/her health. The NHP asks respondents to tick 'yes' if they have the problem stated in each item. The addition of 'in relation to his/her health' contributes to the complexity of the SIP.

Routing: Routing refers to conditional questions following responses to preceding questions. There is no routing in the NHP; all respondents must answer all questions. The inclusion of routing in the SIP for Work items adds to the complexity of the instrument and our data did in fact confirm that the respondents were confused. For example, although only 22 respondents indicated that they performed paid work, the SIP Work-items were answered by 44 respondents. Because of this, the SIP Work dimension was left out of further analyses.

Length: The NHP consists of 38 items. It has been reported that an average of 10 minutes is the completion time for self-assessment. The respondents in the present study needed on average 8 minutes (sd=3 min). The SIP consists of 136 items, with reports of completion time ranging from 20–30 minutes.

Complexity: The reading burden may be indicated by the number of words per item. The NHP-DA consists on average of 8.5 (sd=3.9) words per item, the SIP of 11.7 (sd=6.3). The SIP contains 16 questions comprising more than 20 words, compared with the NHP where this does not occur.

Features of score distribution

Mean scores, standard deviations, and the percentages of the respondents with the maximum possible score and the minimum possible score, respectively, for each instrument are shown in Table 3. The distributions of the scores of the SIP were even more

* An exception to the broad applicability of the NHP was observed when the NHP was employed in another study among patients with spinal cord injury. As these patients were not able to walk at all, most of the items belonging to the dimensions Physical Mobility and Pain were 'not applicable' for them.

Table 3. Features of score distribution, internal consistency (Cronbach's α) and 24-hours test–retest reliability (ICC) of NHP and SIP scales; score distributions of STAI, SDS-Zung, ADL, IWB and Karnofsky. Renal dialysis patients, n=63

	mean	sd	% max*	% min**	α	test–retest
NHP (score 0–100)						
Physical Mobility (8)***	26.3	24.8	29	0	0.80	0.80
Energy (3)	33.0	35.8	43	13	0.69	0.62
Pain (8)	13.3	20.6	46	0	0.76	0.73
Sleep (5)	38.6	34.9	24	10	0.66	0.75
Emotional Reactions (9)	17.6	21.8	38	2	0.74	0.55
Social Isolation (5)	12.9	19.7	60	0	0.39	0.57
SIP (score 0–100)						
Sleep and Rest (7)	16.8	17.1	27	0	0.48	—
Emotional Behavior (9)	6.5	11.0	67	0	0.62	—
Bodycare and Movement (23)	6.7	9.9	38	0	0.81	—
Home Management (10)	21.7	20.5	21	0	0.68	—
Mobility (10)	12.7	14.1	46	0	0.70	—
Social Interaction (20)	9.3	9.7	25	0	0.75	—
Ambulation (12)	15.4	14.7	29	0	0.73	—
Alertness Behavior (10)	11.8	18.5	57	0	0.82	—
Communication (9)	6.0	12.6	71	0	0.77	—
Recreation and Pastimes (8)	29.5	22.8	16	0	0.66	—
Eating (9)	9.4	5.4	13	0	0.14	—
SIP total score	12.2	9.5	16	0	0.95	—
SIP physical score	9.8	10.6	19	0	0.89	—
SIP psychosocial score	8.6	10.2	0	0	0.90	—
ADL (score 10–1)	8.8	1.4	44	0	—	—
STAI (score 20–80)	38.6	11.3	3	0	—	—
SDS-Zung (score 25–100)	40.1	8.2	1	0	—	—
Karnofsky (score 100–0)	72.2	16.4	11	0	—	—
IWB (score 14.7–2.1)	10.4	3.2	0	0	—	—

* % max=percentage of respondents with best possible score (ceiling);
** % min=percentage of respondents with worst possible score (floor);
*** number of items

skewed in the direction of good functioning than those of the NHP.

Internal consistency and test–retest reliability

The internal consistency coefficients (Cronbach's α) for NHP and SIP scales respectively are shown in Table 3.

The scales of the NHP yielded somewhat higher internal consistency estimates (mean α=0.67; range=0.39–0.80) than those of the SIP (mean α=0.65; range=0.14–0.82). The α-coefficients for three of the NHP scales [Social Isolation (0.39), Sleep (0.66) and Energy (0.69)] and for five of the SIP scales [Sleep and rest (0.48), Emotional behavior (0.62), Home management (0.68), Recreation and pastimes (0.66), Eating (0.14!)] fell well below the 0.70 standard recommended for group comparisons. Nineteen SIP-items showed zero variance, which was explainable because they addressed very serious impairment of functioning.

Test–retest reliability estimates (ICCs) for the NHP scales are also shown in Table 3. The precautions mentioned in the *Patients* section are to be borne in mind when interpreting these figures. As could be expected from the item content, test–retest reliability was highest for Physical Mobility^{NHP}. Test–retest reliability was rather low for Social Isolation^{NHP} and Emotional Reaction^{NHP}.

Table 4. Internal structure of NHP and SIP: summary* of ICCs for each scale with the other scales of NHP and SIP respectively

NHP	Physmob	Pain	Energy	Sleep	Soc	Emot						
	0.41	0.38	0.33	0.32	0.35	0.43						Total
												0.37
SIP	SR	EB	BCM	HM	M	SI	A	AB	C	RP	E	Total
	0.39	0.39	0.43	0.43	0.48	0.45	0.43	0.39	0.38	0.32	0.22	0.40

* For example: the figure of 0.41 for NHP Physical Mobility represents the square root of $(((0.49)^2+(0.48)^2+(0.40)^2+(0.31)^2+(0.33)^2)/5)$ (Appendix 1)

Table 5. Correlation (ICCs) of NHP and SIP scales, respectively, with STAI, SDS, ADL Karnofsky and IWB (renal dialysis patients, n=63)

	ADL*	Karnofsky*	STAI*	SDS-Zung*	IWB*
NHP					
Physical Mobility	0.58	0.55	0.35	0.37	0.37
Pain	0.41	0.32	0.25	0.35	0.23
Energy	0.20	0.34	0.48	0.28	0.36
Sleep	0.25	0.30	0.32	0.24	0.39
Emotional Reactions	0.32	0.22	0.28	0.46	0.34
Social Isolation	0.27	0.35	0.48	0.48	0.37
SIP					
Sleep and Rest	0.25	0.27	0.31	0.35	0.21
Emotional Behavior	0.16	0.15	0.22	0.27	0.14
Bodycare and Movement	0.55	0.20	0.13	0.28	0.12
Home Management	0.42	0.40	0.34	0.37	0.32
Mobility	0.57	0.34	0.22	0.29	0.10
Social Interaction	0.20	0.16	0.23	0.34	0.17
Ambulation	0.51	0.32	0.20	0.37	0.20
Alertness Behavior	0.19	0.27	0.30	0.41	0.24
Communication	0.18	0.18	0.08	0.17	0.04
Recreation and Pastimes	0.16	0.35	0.35	0.22	0.33
Eating	0.05	0.07	0.04	0.06	0.02

* rescaled to a 0–100 scale (0=optimal score) in accordance with NHP and SIP scales

Structure

The ICCs for the NHP scales and the SIP scales, respectively, are summarized in Table 4 (complete data shown in Appendix 1). In general, the NHP scales were somewhat less highly intercorrelated than were the SIP scales. As was expected, high ICCs were observed between Social Isolation^{NHP} and Emotional Reaction^{NHP}. The SIP scales grouped in the Physical subscore (Bodycare and Movement, Mobility, Ambulation) showed high intercorrelations. A similar pattern was observed for the SIP scales grouped into the psychosocial subscore (Social Interaction, Alertness Behavior, Emotional Behavior, Communication). Eating^{SIP} correlated low with the other SIP scales.

Construct validity

Firstly, the matrix of ICCs between NHP scales and SIP scales is presented in Appendix 1. We expected higher correlations between 'physical' dimensions (Physical Mobility^{NHP}, Bodycare and Movement^{SIP}, Mobility^{SIP}, Ambulation^{SIP}) and between 'psychosocial' dimensions (Social Isolation^{NHP}, Emotional Reaction^{NHP}, Emotional Behavior^{SIP}, Social Interaction^{SIP}, Alertness Behavior^{SIP}, and Communication^{SIP}); and weaker correlations between physical and psychosocial dimensions. The correlations observed between the NHP and SIP scales were generally rather low. There were some deviations from the expected patterns; for example, low ICCs between Social Isolation^{NHP} and Emotional Behavior^{SIP}, between Social

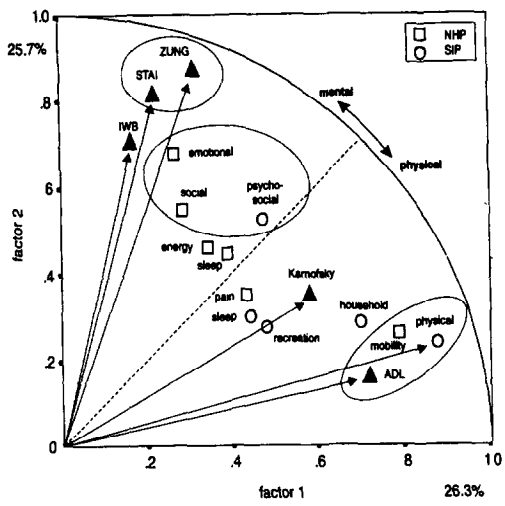
Interactions^{NHP} and Communication^{SIP}, between Emotional Reaction^{NHP} and Communication^{SIP}. The latter two observations are understandable as the items of Communication^{SIP} are of a rather physical nature (e.g., difficulties in speaking).

Secondly, correlation patterns as observed between the scales of NHP and SIP and five instruments with proved validity (STAI, SDS-Zung, ADL, Karnofsky, IWB) were compared to *a priori* hypotheses with respect to these correlation patterns. For example, we expected the highest correlations with ADL and Karnofsky for Physical Mobility^{NHP} and for Pain^{NHP}, and we expected the highest correlations with STAI and SDS-Zung for Social Isolation^{NHP} and Emotional Reaction^{NHP}. We similarly expected the highest correlations with STAI and SDS-Zung for the components of the psychosocial subscore of the SIP (Social Interaction, Alertness behavior, Emotional behavior and Communication), and the highest correlations with ADL and Karnofsky for the components of the physical subscore of the SIP (Bodycare and movement, Mobility, Ambulation).

The association patterns observed between the NHP and the SIP, respectively, and the other five instruments were largely as expected. Exceptions were Communication^{SIP} which correlated weakly with STAI and SDS-Zung, understandable in view of the from the reasoning described above, and Social interactions^{NHP} which also correlated weakly with STAI. The IWB (as a measure for experienced well-being) showed the highest correlations (ICC >0.3) with Recreation and pastimes^{SIP}, Household management^{SIP}, and all NHP dimensions except Pain^{NHP}.

Common factor analysis with varimax rotation of the combined data of NHP (6 scales), SIP (physical subscore, psychosocial subscore, Sleep and rest, Recreation and pastimes, Household management*), ADL, Karnofsky, STAI, SDS-Zung and IWB yielded two factors with eigen values >1.0; see Figure 1. The first factor explained 26.3% of common variance and was interpreted as a physical dimension, the second factor explained 25.7% of common variance and was interpreted as a mental dimension. Scales with high loadings on the physical factor were the Physical subscore of the SIP; Physical Mobility^{NHP}; ADL; Household management^{SIP}; and Karnofsky. Scales with high loadings on the mental factor were SDS-Zung; STAI; IWB; Emotional reaction^{NHP}; Social Isolation^{NHP}; and the SIP psychosocial subscore. The physical scales of NHP and SIP (Physical Mobility^{NHP} and the physical subscore^{SIP}) are closer to each other in Figure 1 than the mental scales (Emotional reaction^{NHP}, Social Isolation^{NHP}, psychosocial subscore^{SIP}). This means that there is more similarity between NHP and SIP in the physical domain than in the mental domain. The IWB loaded very high on the second factor, indicating that well-being as measured with the IWB is largely determined by mental factors in this population.

Figure 1. Factor analysis with varimax rotation of NHP, SIP, ADL, Karnofsky, STAI, SDS-zung and IWB (renal dialysis patients, n=63)



Conclusion and discussion

In this study we have compared the feasibility, structure and psychometric characteristics of two well-known generic health status measures—the NHP and the SIP—when employed in a group of renal dialysis patients. The results are summarized in Table 6.

The NHP can be considered to be generally more feasible than the SIP. The NHP is shorter and less difficult. The observed difference in item contents (relating to actions in the NHP, to activities in the SIP) might cause the SIP to be less universally applicable and more culture-bound than the NHP. For example, the Work items of the SIP have often been observed to be omitted from the questionnaire in elderly populations. It is interesting to note that Part 2 of the NHP, which was not used in the empirical

* Eating^{SIP} was left out of the ultimate factor analysis that is presented here, because it was so different from the other variables (see low correlations with the other variables) that it emerged as a separate 'factor' and interfered too much with the factor analysis.

Table 6. Summary of the empirical comparison of NHP and SIP

	NHP	SIP
Feasibility	generally better	
Internal consistency	acceptable for 5 out of 6 scales	acceptable for 8 out of 11 scales
Test–retest reliability	acceptable	not available
Structure	confirmed	confirmed
Construct validity	more emphasis on mental health, perceived health	more emphasis on physical health, functional health

part of our study and is thus not addressed in this paper, refers to activities as well.

The results for internal consistency were better for the NHP than for the SIP. The internal consistency is (almost) acceptable for five out of six NHP scales, and for eight out of 11 SIP scales. Published data on internal consistency of the NHP scales for the UK version appeared to be unavailable. The study by Erdman *et al.*[17] among 276 Dutch general practice patients showed a mean of 0.78, all αs 0.70 or higher. The lower internal consistency estimates in our study, especially for the Social Isolation scale (0.39), may be due to the different nature of the study population. It supports the fact that psychometric characteristics are population-specific.

Internal consistency estimates for 10 out of 12 USA SIP scales are available for a stratified sample of members of a USA prepaid group practice [n=495; mean α=0.61, range=0.29 (Eating) to 0.82 (Social interaction); eight out of ten α's below 0.70] and a group of 168 noncognitively impaired nursing home patients [mean α=0.72, range=0.60 (eating, sleep and rest) to 0.84 (Body care and movement); three out of ten α's below 0.70].[18] These results and the results of the present study are indicative of a borderline acceptable level of internal consistency of several SIP scales. Internal consistency estimates for the SIP as a whole (136 items) exceed 0.90 for the USA, the Swedish, the Spanish and the Dutch version, but this is partly attributable to the large number of items.

With respect to test–retest reliability, results (4-week intervals) for the UK NHP among 58 arthrosis patients were in the range of 0.77–0.85 (Spearman rank correlation coefficients) and among 93 patients with peripheral vascular disease in the range of 0.75–0.88.[20,21] Test–retest reliability of the Dutch NHP in cardiac patients showed Spearman rank correlations of 0.69–0.84.[17] The somewhat lower test–retest reliability estimates in the present study may be partly attributed to the fact that it is not quite sure that patients' health status remained unchanged between the two assessments: preceding dialysis and 24 hours later. For the USA SIP, 24 hours test–retest reliability coefficient was 0.92 for the total score over 136 items.[22]

Examination of the inter-scale correlations for the NHP and the SIP showed these correlations to be of moderate magnitude, suggesting little redundancy of information generated by the scales of the instruments. For the Dutch NHP, these results replicate the findings of Erdman *et al.*[17]

The ICCs observed between NHP scales and SIP scales were rather low, suggesting that the NHP and SIP to some extent measure different aspects of health status. ICCs observed between the NHP scales and the SIP scales, respectively, and instruments indicating mainly physical functioning (ADL, Karnofsky) and mainly psychological functioning (STAI, SDS-Zung) were largely as expected. However, the psychosocial scales of the SIP correlated more weakly with STAI and SDS-Zung than the psychosocial scales of the NHP. The IWB exhibited ICCs >0.3 with one SIP scale and with five out of six NHP scales. Factor analysis yielded a two-factor solution with a physical and a mental factor of equal importance and showed the SIP scales to load more on the physical factor (with the psychosocial subscore as the only exception). A similar result was obtained by Bruin *et al.*,[23] who performed principal components analysis on 835 SIPs completed by subjects from different diagnostic categories.

The NHP scales, however, loaded more on the mental factor (exceptions: Physical Mobility, Pain). This may be interpreted as the SIP emphasizing physical functioning, whereas the NHP emphasizes mental functioning. The analysis also confirms to some extent the intentions of the constructors of the NHP and the SIP respectively, i.e. that the NHP was intended to be a measure of perceived health while the SIP was intended to be a more functional measure.

The results of the present study add to the developing body of knowledge with respect to performance characteristics of Dutch adaptations of the NHP and the SIP. A cross-culturally adapted health status measure is essentially a new instrument, and

investigation of its characteristics is required.[16] Cross-cultural adaptation of health status measures requires more than 'conceptually equivalent' translation, because of expected cultural differences with respect to health beliefs and response to questionnaires. This is required even among residents of industrialized societies. Jacobs showed that the USA item weights for the SIP items can be validly applied for Dutch SIP data.[24] The French NHP item weights showed some differences if compared with the British ones.[25]

The NHP generally performed better than the SIP in this study. This does not imply that the NHP is generally to be preferred to the SIP in medical evaluation research. Firstly, responsiveness to change over time was not a subject of comparison in the present study. Secondly, performance characteristics of generic instruments for health status are probably population specific. For an instrument to perform well it must do so in terms of feasibility, internal consistency, test–retest reliability, and validity including responsiveness to change over time. An instrument which performs well according to the aforementioned criteria in a population of elderly, rather seriously ill patients with renal insufficiency will not necessarily perform equally well when applied for example to young patients with lung problems. The possibility that an instrument performs equally well in all types of patient groups with varying degrees of illness can be seriously doubted. The case might eventually be that NHP and SIP are each superior in different groups.

Appendix 1. ICCs of NHP scales and SIP scales (n=63)

	Phys-mob	Pain	Energy	Sleep	Social	Emotion	SR	EB	BCM	HM	M	SI	A	AB	C	RP	E
Physmob																	
Pain	0.49																
Energy	0.48	0.25															
Sleep	0.40	0.23	0.33														
Social	0.31	0.38	0.12	0.24													
Emotion	0.33	0.48	0.38	0.38	0.56												
SR	0.27	0.20	0.09	0.22	0.23	0.23											
EB	0.10	0.31	0.11	0.09	0.22	0.38	0.35										
BCM	0.32	0.42	0.10	0.11	0.34	0.24	0.29	0.50									
HM	0.56	0.28	0.29	0.26	0.35	0.35	0.46	0.24	0.34								
M	0.40	0.31	0.18	0.15	0.32	0.33	0.51	0.47	0.60	0.61							
SI	0.17	0.14	0.14	0.11	0.38	0.31	0.51	0.56	0.48	0.38	0.50						
A	0.43	0.36	0.23	0.13	0.21	0.27	0.40	0.38	0.54	0.58	0.61	0.36					
AB	0.25	0.34	0.19	0.11	0.59	0.41	0.28	0.42	0.41	0.43	0.41	0.50	0.38				
C	0.16	0.15	0.05	0.05	0.27	0.25	0.22	0.42	0.51	0.30	0.42	0.44	0.26	0.57			
RP	0.37	0.08	0.33	0.23	0.17	0.26	0.46	0.20	0.14	0.54	0.34	0.26	0.42	0.29	0.17		
E	0.02	0.05	0.02	0.02	0.02	0.07	0.21	0.23	0.21	0.14	0.20	0.40	0.22	0.16	0.14	0.12	

PHYSMOB=NHP Physical Mobility; PAIN=NHP Pain; ENERGY=NHP Energy; SLEEP=NHP Sleep; SOCIAL=NHP Social Isolation; EMOTION=NHP Emotional Reaction
SR=SIP Sleep and rest; EB=SIP Emotional behavior; BCM=SIP Bodycare and movement; HM=SIP Home management; MOB=SIP Mobility; SI=SIP Social interaction; A=SIP Ambulation; AB=SIP Alertness behavior; C=SIP Communication; RP=SIP Recreation and pastimes; E=SIP Eating

References

1. Hunt SM, McEwen J, McKennaSP. Measuring health status: A new tool for clinicians and epidemiologists. *J R Coll Gen Pract* 1985; **35**: 185–188.
2. Luttik A, Jacobs HM, De Witte LP. De Sickness Impact Profile. Vakgroep Huisartsgeneeskunde Rijksuniversiteit Utrecht /Instituut voor Revalidatievraagstukken Rijksuniversiteit Limburg, 1987.
3. Melker RA de, Touw-Otten F, Jacobs HM, Luttik A. De waarde van de 'sickness impact profile' als uitkomstmeting. *Ned Tijdschr Geneeskd* 1990; **134**: 946–948.
4. Ploeg HM van der, Defares PB, Spielberger CD, eds. *Manual of the State-Trait Anxiety Inventory (Handleiding bij de zelfbeoordelingsvragenlijst).* Leiden: Swets and Zeitlinger, 1980.
5. Ploeg HM van der. Validation of de State-trait Anxiety Inventory (Validatie van de zelfbeoordelingsvragenlijst). *Ned T Psychol* 1980; **35**: 243–249.
6. Spielberger CD, Gorschuch RL, Lushene RE, eds. *STAI manual for the state-trait anxiety inventory.* Consulting Psychologists Press, Palo Alto, California 1970.
7. Zung WWK. A self-rating depression scale. *Arch Gen Psych* 1965; **13**: 63–70.
8. Zitman FG, Griez EJL, Hooijer Chr. Standardisation of depression assessment questionnaires (Standaardisaering depressievragenlijsten; in Dutch). *T Psychiatr* 1989; **31**: 114–135.
9. Dijkstra P. The Zung self-rating depression scale (De zelfbeoordelingsschaal voor depressie van Zung). In: van Praag HM, Rooymans HGM, *Stemming en ontstemming.* Amsterdam: de Erven Boon, 1974: 98–120.
10. Karnofsky DA, Abelman WH, Craver LF, Burchenal JH. The use of nitrogen mustards in the palliative treatment of carcinoma. *Cancer* 1948; 643–654.
11. Bonsel GJ, Bot ML, Boterblom A, Veer F van 't. Costs and effects of heart transpalntation (De kosten en effecten van harttransplantatie), part 2A, 2B, 2C: *Quality of life—documentation, interview, results (Kwaliteit van leven voor en na harttransplantatie—documentatie, interview, resultaten).* Rotterdam: Department of Public Health, Erasmus University, 1988.
12. Campbell A, Converse PE, Rodgers WL. *The quality of American life: perceptions, evaluations and satisfactions.* New York: Russell Sage Foundation, 1976.
13. Nunnally JC. *Psychometric theory.* New York: McGraw Hill, 1978.
14. Dunn G. *Design and analysis of reliability studies.* Oxford: Oxford University Press, 1989.
15. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. *Controlled Clinical Trials* 1991; **12**: 142S.
16. Streiner DL, Norman GR. *Health measurement scales.* Oxford: Oxford Medical Publications. 1989.
17. Erdman RAM, Passchier J, Kooijman M, Stronks DL. The Dutch version of the Nottingham Health Profile: investigations of psychometric aspects. *Psych Reports* 1993; **72**: 1027–1035.
18. Rothman ML, Hedrick S, Inui T. The Sickness Impact Profile as a measure of the health status of noncognitively impaired nursing home residents. *Med Care* 1989; **27 (Suppl 3)**, S157–167.
19. de Bruin AF, Witte LP, Stevens FCJ, Diederiks JPM. The usefullness of the sickness Impact Profile as a generic fuctional status measure (in Dutch; English abstract). *T Soc Gezondheidsz* 1992; **70**: 160–170.
20. Hunt SM, McKenna SP, Williams J. Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthrosis. *J Epidemiol Comm Health* 1981; **35**: 297–303.
21. Hunt SM, McEwen J, McKenna SP. *Measuring health status.* London: Croom Helm. 1986: 116.
22. 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.
23. de Bruin AF, Diederiks JPM, de Witte LP, Stevens FCJ, Philipsen H. The development of a short generic version of the Sickness Impact Profile. *J Clin Epidemiol* 1994; **47**: 407–418.
24. Jacobs HM, Luttik A, Touw-Otten FWMM, de Melker RA. De 'sickness impact profile': resultaten van een valideringsonderzoek van de Nederlandse versie. *Ned Tijdschr Geneeskd* 1990; **134**: 1950–1954.
25. Bucquet D, Condon S, Ritchie K. The French version of the Nottingham health profile. A comparison of items weights with those of the source version. *Soc Sci Med* 1990; **30**: 829–835.

(Received 4 June 1995; accepted 17 July 1995)