Validating the SF-36 health survey questionnaire: new outcome measure for primary care

J E Brazier, R Harper, N M B Jones, A O'Cathain, K J Thomas, T Usherwood, L Westlake

Abstract

Objectives—To test the acceptability, validity, and reliability of the short form 36 health survey questionnaire (SF-36) and to compare it with the Nottingham health profile.

Design—Postal survey using a questionnaire booklet together with a letter from the general practitioner. Non-respondents received two reminders at two week intervals. The SF-36 questionnaire was retested on a subsample of respondents two weeks after the first mailing.

Setting—Two general practices in Sheffield.

Patients—1980 patients aged 16-74 years randomly selected from the two practice lists.

Main outcome measures—Scores for each health dimension on the SF-36 questionnaire and the Nottingham health profile. Response to questions on recent use of health services and sociodemographic characteristics.

Results—The response rate for the SF-36 questionnaire was high (83%) and the rate of completion for each dimension was over 95%. Considerable evidence was found for the reliability of the SF-36 (Cronbach's $\alpha > 0{\cdot}85$, reliability coefficient $> 0{\cdot}75$ for all dimensions except social functioning) and for construct validity in terms of distinguishing between groups with expected health differences. The SF-36 was able to detect low levels of ill health in patients who had scored 0 (good health) on the Nottingham health profile.

Conclusions—The SF-36 is a promising new instrument for measuring health perception in a general population. It is easy to use, acceptable to patients, and fulfils stringent criteria of reliability and validity. Its use in other contexts and with different disease groups requires further research.

Medical Care Research Unit and Department of General Practice, University of Sheffield Medical School, Sheffield S10 2RX
J E Brazier, *lecturer in health economics*
R Harper, *research associate*
N M B Jones, *statistician*
A O'Cathain, *research associate*
K J Thomas, *senior research associate*
T Usherwood, *senior lecturer in general practice*
L Westlake, *statistician*

Correspondence to:
Mr Brazier.

BMJ 1992;305:160-4

Introduction

It is important to be able to measure the perception of health of the population to assess the benefit of health care interventions and to target services. However, existing measures of mortality and morbidity in the NHS are too narrow, particularly in general practice, to measure the benefit of interventions aimed at improving a wide range of dimensions including mobility, functioning, mental health, and overall well being. Researchers have developed measures to assess the health of people with specific diseases or disabilities,[1,2] but these are of limited application when studying people with more than one condition or comparing perceived health across different groups. What is required is a measure which is comprehensive and sensitive to the full range of illness. To be of practical use the measure must also be brief and easy to use.

One measure which is sensitive to health differences in a general population has been developed out of the Rand Corporation's health insurance experiment, a comprehensive evaluation of alternative methods of financing health care in the United States.[3] The original general health measure was lengthy, containing 108 items. In an attempt "to develop a general health survey that is comprehensive and psychometrically sound, yet short enough to be practical for use in large scale studies of patients in practice settings,"[4] the authors experimented with several shortened versions. The short form 20 has already been fielded with some success in the medical outcomes study surveys in the United States[5] and in Scotland.[6] However, the substantially revised short form 36 health survey questionnaire (SF-36) has yet to be independently validated in Britain. We examined the reliability and validity of the SF-36 in a British population, and compared it with the Nottingham health profile,[7] which is widely used in Britain.

Methods

The SF-36 questionnaire is a self administered questionnaire containing 36 items which takes about five minutes to complete. It measures health on eight multi-item dimensions, covering functional status, well being, and overall evaluation of health (table I).

TABLE I—*Dimensions of the SF-36 health survey questionnaire*

Area	Dimension	No of questions
Functional status	Physical functioning	10
	Social functioning	2
	Role limitations (physical problems)	4
	Role limitations (emotional problems)	3
Wellbeing	Mental health	5
	Vitality	4
	Pain	2
Overall evaluation of health	General health perception	5
	Health change*	1
Total		36

*This item is not included in the eight dimensions nor is it scored.

Five of these dimensions are similar to those in the Nottingham health profile, but items in the SF-36 questionnaire are claimed to detect positive as well as negative states of health.[4] In six of the eight dimensions patients are asked to rate their responses on three or six-point scales (box) rather than simply responding yes or no as in the Nottingham questionnaire. For each dimension, item scores are coded, summed, and transformed on to a scale from 0 (worst health) to 100 (best health).

We conducted face to face interviews using the original American version of the SF-36 in a general practice surgery and among colleagues to examine its acceptability. As a result the wording of six questions was altered slightly. This anglicised version of the

Samples of questions from the SF-36

The following questions are about activities you might do during a typical day. Does your health limit you in these activities? If so, how much?

	Yes, limited a lot	Yes, limited a little	No, not limited at all
Climbing several flights of stairs	○	○	○
Bending, kneeling, or stooping	○	○	○
Walking half a mile	○	○	○

These questions are about how you feel, how things have been with you during the past month.
How much time during the past month:

	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
Did you feel full of life?	○	○	○	○	○	○
Have you felt downhearted and low?	○	○	○	○	○	○
Has your health limited your social activities (like visiting friends or close relatives)?	○	○	○	○	○	○

SF-36 was incorporated into a booklet, together with the Nottingham health profile and questions on sociodemographic characteristics and recent use of health services. We conducted a pilot postal survey of 120 patients from a general practice list to test the acceptability of mailing the booklet. We obtained a response rate of 40% without reminders, with a good completion rate.

The questionnaire booklet was sent to 1980 people aged 16-74 years randomly selected from two general practice lists in Sheffield. It was accompanied by a letter from the general practitioner, endorsing the aims of the study. Two reminder letters and further booklets were sent to non-respondents at intervals of two weeks.

To examine the retest reliability a copy of the SF-36 questionnaire was sent to 250 randomly selected respondents after two weeks.

STATISTICAL ANALYSIS

The responses to the questionnaire were subjected to recommended tests of reliability and validity.[8,9] These are discussed in detail below.

Internal consistency is the extent to which items within a dimension are correlated with each other. It can be examined by several methods: item to own dimension correlations calculated after correction for overlap; Cronbach's α, a widely used method based on correlations between items; and reliability coefficients for each dimension calculated by two way analysis of variance.[10] We used non-parametric versions of these tests to avoid any distributional assumption.

Test-retest reliability—A correlation coefficient measures the degree of association between the test and retest scores but does not indicate the direction of this association. For example, if everyone consistently scored lower on the retest, the correlation coefficient would be highly positive. To overcome this, Bland and Altman recommended a technique which examines the distribution of differences in scores.[11] The differences are plotted, an overall mean and variance of differences calculated, and 95% confidence intervals constructed around the mean by assuming a normal distribution. The test and retest scores are assumed to be from the same distribution when the differences have a mean of zero and 95% of the differences lie within the 95% confidence limits.

Validity—The validity of a health measure is conceptually difficult to prove without a standard. One method is to examine construct validity, where hypotheses or constructs concerning the expected distribution of health between groups are examined by the measure being validated.[8,9] For example, women, older people, and people in social classes IV and V might be expected to perceive relatively poorer health; people making use of health services might also be expected to have poorer perceived health than non-users. We used Kruskal-Wallis one way analysis of variance to test whether the SF-36 scores differed significantly among these groups. The convergent and discriminant validity of SF-36 was examined by the multitrait multimethod matrix.[12] For convergent validity, the correlation between comparable dimensions on SF-36 and Nottingham health profile—for example, between physical functioning and physical mobility—should be higher than the correlations between less comparable dimensions—for example, physical functioning and social isolation. We tested discriminant validity by comparing item to own scale correlation with item to other scale correlation. The item to own scale correlation should be higher if the categories within the SF-36 questionnaire are valid.

Discriminatory power—The ability of an instrument to discriminate between different levels of ill health is strictly a form of validity testing. We considered it separately because it is a key criterion for any measure of general health in a population. Discriminatory power is indicated by the frequency distributions of scores obtained from the measures, with a less skewed distribution indicating greater discriminatory power. A highly skewed distribution of scores requires use of a binary outcome whereas a wider range of scores enables detection of intermediate health states. However, it should be confirmed that greater discriminatory power is genuine and correctly identifies ill health.

Results

We received completed questionnaires from 1582 of the 1980 patients surveyed, of whom 77 could not be contacted, thus giving a response rate of 83%. Of the 250 patients sent a repeat test, 187 (75%) responded. The proportions of missing data from each dimension were lower (0·5%-4%) for the SF-36 questionnaire than for the Nottingham health profile (4-7%). Because so few data were missing for the SF-36 dimensions and the study sample was large, we did not substitute for missing data. The extent of missing data was significantly associated ($p<0·001$) with increasing age in three of the eight SF-36 dimensions (pain, role limitations due to physical problems, and role limitations due to emotional problems).

CHARACTERISTICS OF SAMPLE

The sociodemographic characteristics and use of health services of the respondents did not differ from those found in the general household survey (1988) for the same age range, except for socioeconomic class, where the study sample included fewer people in class II but more in class III and more employed women. Too few patients from ethnic minorities were available to permit separate analyses. Non-respondents in the main survey (n=297) were significantly more likely to be male and younger in age and less likely to have visited their general practitioner recently ($p<0·005$).

INTERNAL CONSISTENCY

Internal consistency was acceptable. The item to own dimension correlations, after correction for overlap, exceeded 0·5 for all except three of the 33 items. Cronbach's α exceeded the recommended

minimum of 0·85[9] and the reliability coefficients were greater than 0·75 for all dimensions except social functioning (α=0·73, reliability=0·74) (table II). The results for social functioning partly reflect the low number of items (two) in that dimension.

TEST-RETEST RELIABILITY AT TWO WEEKS

The re-test scores were highly correlated with those from the main survey (table II). In the analysis recommended by Bland and Altman[11] the mean of the differences was significantly different from zero for six dimensions but did not exceed one point on the 100 point scale, making it clinically insignificant (table II). For all dimensions 91-98% of cases lay within the 95% confidence interval constructed for a normal distribution.

TABLE II—*Reliability of SF-36 questionnaire in general practice population*

Dimension	Internal consistency: Cronbach's α	Internal consistency: Reliability coefficients	Test-retest reliability (2 week interval): Correlation	Test-retest reliability (2 week interval): Mean difference	Test-retest reliability (2 week interval): % Of cases lying within 95% confidence interval
Physical functioning	0·93	0·93	0·81	0·49	98
Social functioning	0·73	0·74	0·60	0·15	93
Role limitations (physical problems)	0·96	0·88	0·69	0·57*	98
Role limitations (emotional problems)	0·96	0·79	0·63	0·44*	97
Pain	0·85	0·84	0·78	0·70*	95
Mental health	0·95	0·91	0·75	0·71*	91
Vitality	0·96	0·87	0·80	0·39*	96
General health perception	0·95	0·80	0·80	0·40*	96

*Significantly different from zero at 5% level.

VALIDITY

Table III shows the distributions of SF-36 scores by sex, age, social class, and use of health services and for patients with chronic disease. The distribution of scores conformed to what might be expected, thus providing evidence of construct validity. Men perceived themselves to be significantly healthier than women ($p<0·001$), except on the general health dimension. Significant age gradients were found for physical functioning and pain ($p<0·001$), but little or no gradient was found for mental health ($p=0·585$). Health decreased with lower social class across all dimensions ($p<0·05$) except for general health perception. Those patients who had consulted a general practitioner in the previous two weeks had poorer perceived health than those who had not consulted recently. Seventy seven patients for whom the general practitioner had diagnosed one or more chronic physical problems perceived their health as worse on all dimensions ($p<0·001$), except mental health, than a sample of patients without chronic physical problems matched for age, sex, and general practice ($p<0·05$).

The expected relations for convergent and discriminant validity were mostly satisfied (table IV). Correlation coefficients for four comparable dimensions of the SF-36 questionnaire and Nottingham health profile were higher than correlations between non-comparable dimensions. This was not found for the correlation of social functioning with social

TABLE III—*Mean scores on dimensions of SF-36 questionnaire in relation to sociodemographic variables and use of health services*

Variable	n*	Physical functioning	Social functioning	Role limitation (physical)	Role limitation (emotional)	Pain	Mental health	Vitality	General health perception
Age (years):									
16-24	240	94	91	92	84	87	74	68	76
25-34	357	95	89	90	84	84	73	63	77
35-44	298	89	87	81	81	78	70	58	72
45-54	267	84	87	83	82	77	72	59	70
55-64	230	74	84	72	80	73	74	59	65
65-74	103	60	80	59	73	67	73	57	58
Sex:									
Male	675	88	90	86	86	81	77	65	72
Female	829	85	84	80	78	77	69	57	71
Socioeconomic class:									
I	38	93	91	87	85	78	75	64	75
II	98	91	90	88	86	81	76	63	75
III non-manual	584	88	86	82	79	80	71	60	73
III manual	302	85	90	84	84	79	77	64	70
IV	277	85	87	81	83	77	72	59	70
V	53	80	79	67	72	72	68	55	65
Students	51	94	95	96	86	85	78	73	77
Chronic physical problems:									
Yes	77	66	74	58	74	59	69	50	53
No	77	78	86	77	74	76	71	57	66
General practitioner consultation in previous 2 weeks:									
Yes	290	81	76	67	73	68	66	52	63
No	1208	88	89	86	84	82	74	63	73
Outpatient attendance in previous 3 months:									
Yes	212	74	75	63	72	64	67	53	59
No	1280	88	89	86	83	82	73	62	73

*n Is the minimum number of respondents completing one dimension. The number of respondents varied for each dimension.

TABLE IV—*Multitrait multimethod matrix of correlation coefficients for SF-36 questionnaire and Nottingham health profile*

	SF-36					Nottingham health profile				
	Physical functioning	Social functioning	Pain	Mental health	Vitality	Physical mobility	Social isolation	Pain	Emotional reactions	Energy
SF-36:										
Physical functioning	0·93*									
Social functioning	0·38	0·74*								
Pain	0·48	0·46	0·84*							
Mental health	0·24	0·56	0·31	0·91*						
Vitality	0·44	0·57	0·48	0·69	0·87*					
Nottingham health profile:										
Physical morbidity	−0·52†	−0·35	−0·45	−0·19	−0·36	0·78*				
Social isolation	−0·20	−0·41†	−0·18	−0·47	−0·36	0·18	0·74*			
Pain	−0·47	−0·35	−0·55†	−0·21	−0·33	0·63	0·17	0·87*		
Emotional reactions	−0·18	−0·53	−0·28	−0·67†	−0·55	0·20	0·49	0·21	0·83*	
Energy	−0·37	−0·51	−0·37	−0·47	−0·68†	0·36	0·38	0·34	0·54	0·68*

*Reliability coefficient. †Correlation coefficients are negative because the two scales run in the opposite direction.

isolation, where the constituent questions seemed to address different aspects of social well being.

DISCRIMINATORY POWER

Comparison of the frequency distribution of SF-36 scores and scores on the comparable dimensions of the Nottingham questionnaire (figs 1 and 2) showed that the SF-36 scores were less skewed. The median scores for all Nottingham health profile dimensions were zero (good health) but were less than 100 (poorer health) on five of the eight dimensions of the SF-36.

Table V shows the patients who scored zero on the Nottingham questionnaire (good health) divided according to those who scored 100 (good health) and those who scored less than 100 (poorer health) on the SF-36 questionnaire (table V). The poorer health group had a higher proportion of women, had an older

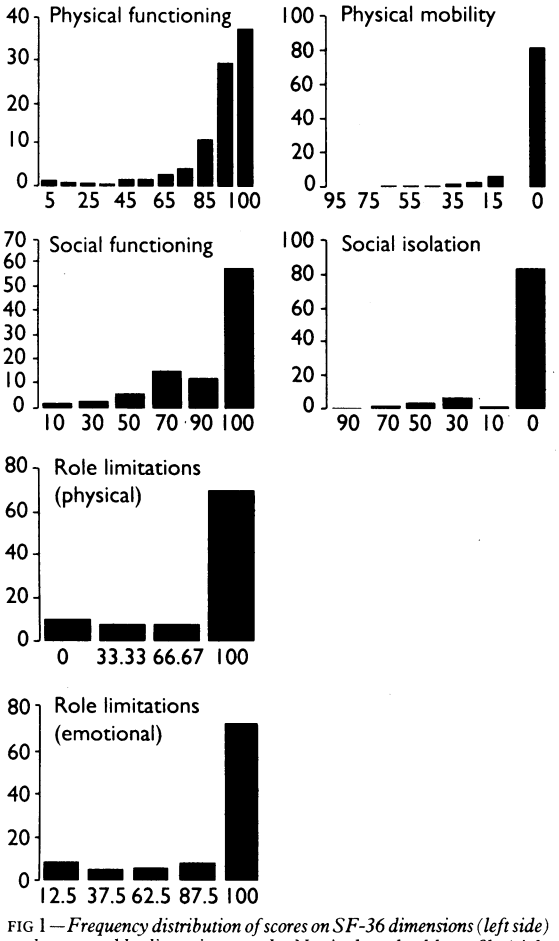

FIG 1—*Frequency distribution of scores on SF-36 dimensions (left side) and comparable dimensions on the Nottingham health profile (right side): functional status*

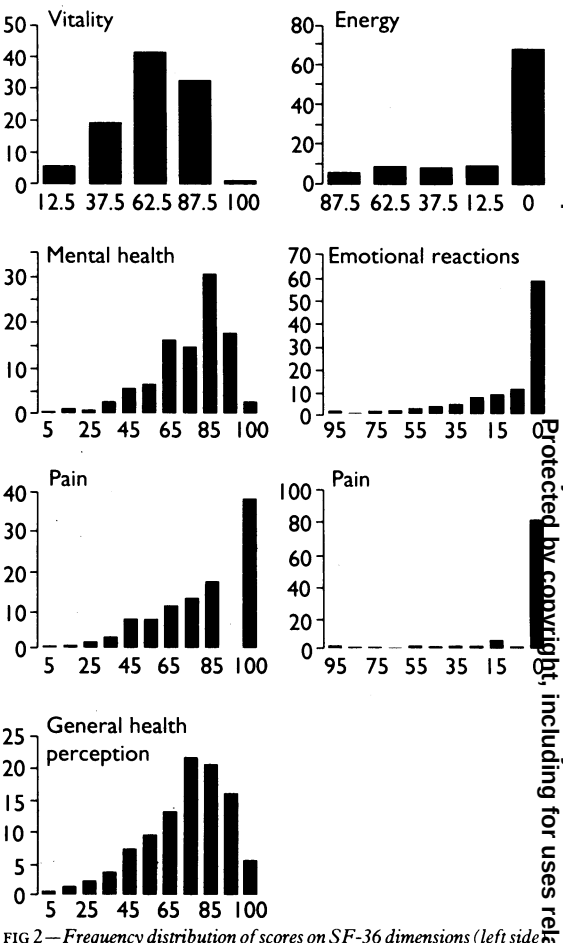

FIG 2—*Frequency distribution of scores on SF-36 dimensions (left side) and comparable dimensions on the Nottingham health profile (right side): well being and overall health*

mean age, and contained a higher percentage of patients not in full time employment than the good health group. Patients in the poorer health group were more likely to have consulted a general practitioner or used outpatient services. These results were significant for physical functioning, social functioning, and pain ($p<0.05$). The numbers of patients scoring 100 in the remaining two comparable dimensions (mental health and vitality) were too few for significance to be shown.

Discussion

In attempting to be comprehensive, existing general health questionnaires such as the sickness impact profile may be too long or require interviews, or both. In primary care or community settings the contact time with patients is often short, and thus to be practical and

TABLE V—*Analysis of results for patients scoring zero on Nottingham health profile: comparison of those in good health (SF-36=100) with those scoring in poorer health (SF-36<100) in relation to sociodemographic characteristics and use of health services*

Dimension score	No of patients	Mean age (years)	Sex (% female)	% Not full time employed	% Visiting general practitioner in previous 2 weeks	% Attending outpatients in previous 3 months	% Inpatients in past year
Physical functioning:							
100	551	30	50·3	38·0	15	8	9
<100	657	44***	57·1*	51·7***	21*	13*	10
Social functioning:							
100	832	39	47·6	44·0	13	9	8
<100	399	42	62·9***	54·5**	29***	19***	13
Pain:							
100	567	35	49·0	41·3	11	7	8
<100	653	38*	58·3**	49·2**	23***	14***	10
Mental health:							
100	36	40	36·1	45·7	11	11	3
<100	816	41	48·8	44·2	14	12	10
Vitality:							
100	22	28·5	22·7	31·8	10	10	5
<100	999	38*	50·1*	43·9	15	11	9

*p<0·05, **p<0·01, ***p<0·001, by χ^2 test except for age (by Mann-Whitney U test)

acceptable to the population the questionnaire must be brief, easy to use, and preferably self administered. These features are also important for researchers, who may want to add a generic health measure to a disease specific questionnaire. The SF-36 questionnaire seemed to meet these criteria, taking just five minutes to complete. We achieved a response rate of 83%, and despite its presentation being more complex than that of the Nottingham questionnaire there were fewer missing data. This quantitative evidence, and the favourable impression for face to face interviews, suggests that the SF-36 questionnaire is an acceptable measure of the health of a general population.

Our findings supported the developers' claims of internal consistency for the SF-36 questionnaire.[4] The test-retest reliability of the SF-36 questionnaire has not been examined before, and since an instrument with a high discriminatory power may be unreliable[8] it was reassuring to find that test-retest reliability was excellent. The maximum mean difference in dimension scores was 0·80, which implies that a person with a test score of 70 might score 71 on retesting. This difference is of no practical significance.

The evidence for the construct validity of the SF-36 was substantial. The expected distribution of scores was observed by sociodemographic characteristics, general practitioner consultation, use of hospital services, and a group of patients with chronic physical problems.

COMPARISON WITH NOTTINGHAM QUESTIONNAIRE

In Britain many researchers,[1] and more recently the NHS,[13] have used the Nottingham health profile to study aspects of health including rheumatoid arthritis,[14] migraine,[14] hypertension,[15] heart transplantation,[16] renal lithotripsy,[17] and cholecystectomy.[18] It has also been successfully applied in other countries.[19 20] The questionnaire takes just a few minutes to complete and is acceptable to the general population.[7] However, it has been criticised for tapping the extreme end of ill health and therefore being unsuitable for examining improvements in health in a general population.[1 21] Our results strongly support this criticism—most of the general population sampled registered a zero score on the Nottingham dimensions, producing highly skewed distributions. The distributions of SF-36 scores were less skewed and showed a substantially higher prevalence of perceived health problems, particularly with regard to mental health and vitality.

By dividing patients who scored zero (good health) on the Nottingham profile into those who scored 100 (good health) or less than 100 (poorer health) on the SF-36 questionnaire we were able to identify people with perceived health problems who were missed by the Nottingham profile. The SF-36 questionnaire therefore seems preferable to the Nottingham profile for measuring the health of a population with relatively minor conditions, such as in general practice or the community.

APPLICABILITY

The King's Fund is supporting several validation studies looking at different patient groups to determine whether the questionnaire is suitable for studying specific groups as well as the general population. Indications from unpublished work in the United States suggest that the SF-36 questionnaire could be used to study a wide range of serious conditions. However, the higher level of missing data for the 65-74 year old age group in our study suggests that further research is required before it is widely applied to elderly patients. Measures such as the SF-36, which produce a profile of scores, can be criticised as unsuitable for comparisons between treatments that may improve the dimension scores differentially. For this purpose a single index of health is preferable and it is not yet known whether SF-36 scores can be used to generate a valid single index. Existing measures which purport to provide single indices, such as the York quality of life measure, have also yet to be validated.[22]

We thank our colleagues in the Department of General Practice, Dr John Poyser, and Dr Helen Joesbury. The study was supported by a grant from the Medical Research Council. The Medical Care Research Unit is funded by the Department of Health and Trent Regional Health Authority. The opinions in this article are those of the authors.

1 Wilkin D, Hallam L, Doggett MA. *Measures of need and outcome for primary health care*. Oxford: Oxford Medical Press, 1992.
2 Bowling A. *Measuring health: a review of quality of life measurement scales*. Milton Keynes: Open University Press, 1991.
3 Ware JE, Brook RH, Williams KN, Stewart AL, Davies-Avery A. *Conceptualisation and measurement of health for adults in the health insurance study. Vol 1. Model of health and methodology*. Santa Monica, California: Rand Corporation, 1980. (Publication No R-1987/1-HEW.)
4 Ware John E, Sherbourne CD. The SF-36 short-form health status survey. 1. Conceptual framework and item selection. *Med Care* (in press).
5 Stewart AL, Hays RD, Ware JE. The MOS short form general health survey. *Med Care*, 1988;**26**:724-35.
6 Anderson JStC, Sullivan F, Usherwood TP. The medical outcomes study instrument (MOSI)—use of a new health status measure in Britain. *Fam Pract* 1990;7:205-18.
7 Hunt S, McKenna SP, McEwen J. *The Nottingham health profile user's manual*. Manchester: Galen Research and Consultancy, 1989.
8 Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press, 1989.
9 McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. New York: Oxford University Press, 1987.
10 Kerlinger FN. *Foundations of behavioural research*. New York: Holt, Rinehart, and Winston, 1973.
11 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307-10.
12 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959;**56**:81-105.
13 *Final report: the CASPE/Freeman outcome study*. London: CASPE Research, 1991.
14 Jenkinson C, Fitzpatrick R. The Nottingham health profile: an analysis of its sensitivity in differentiating illness groups. *Soc Sci Med* 1988;**27**:1411-4.
15 De Lame PA, Droussin AM, Thomson M, Verhaest L, Wallace S. The effects of endopril on hypertension and quality of life. A large multi-center study in Belgium. *Acta Cardiologica*, 1989;**44**:289-302.
16 O'Brien BJ, Banner NR, Gibson S, Yacoub M. The Nottingham health profile as a measure of quality of life following combined heart and lung transplantation. *J Epidemiol Community Health* 1988;**42**:232-4.
17 May N, Petruckevitch A, Snowdon C. Patients quality of life following extracorporeal shock wave lithotripsy and percutaneous nepholithotomy for renal calculi. *Int J Technol Assess Health Care* 1990;**6**:631-40.
18 Milner PC, Nicholl JP, Westlake L, Williams BT, Birch S, Ross B, *et al*. The evaluation of lithotripsy as a treatment for gallstones: a randomised controlled trial approach in England. *Journal of Lithotripsy and Stone Disease* 1989;**1**:122-32.
19 Wiklund I, Romanus B, Hunt SM. Reliability of the Swedish version of the Nottingham health profile. *Int Disabil Stud* 1988;**10**:159-63.
20 Baum FE, Cooke RD. Community-health needs assessment: use of the Nottingham health profile in an Australian study. *Med J Aust* 1989;**150**: 581-90.
21 Kind P, Carr-Hill R. The Nottingham health profile: a useful tool for epidemiologists? *Soc Sci Med* 1987;**25**:905-10.
22 Carr-Hill R, Morris J. Current practice in obtaining the "Q" in QALYs: a cautionary note. *BMJ* 1991;**303**:699-701.

(Accepted 16 June 1992)