British Journal of Health Psychology (1997), 2, 285–300 *Printed in Great Britain*

Psychometric and clinical validity of the SF-36 General Health Survey in the Whitehall II Study

Ron Roberts*

Department of Psychology, University of Westminster, Regent Street, London W1R 8AL, UK

Harry Hemingway and Michael Marmot

International Centre for Health and Society, Department of Epidemiology and Public Health, University College London, 1–19 Torrington Place, London WC1E 6BT, UK

Objectives. To assess the internal consistency, test–retest reliability and factorial structure of the SF-36. To examine the discriminant validity of the instrument in relation to measures of physical (angina, diabetes) and psychological health (chronic GHQ, alcohol problems as defined by CAGE score).

Design. A cross-sectional survey design is used.

Methods. A questionnaire containing the SF-36 was administered to a cohort of 10 308 civil servants aged 35–55 at baseline as part of the Whitehall II study of occupational and life-style influences upon health. Administration of the SF-36 was repeated on a subset of 289 participants four weeks later. Clinical groups were distinguished by self-report measures of health.

Results. Data show high internal consistency (alpha .75–.85). Test–reset reliability was poor for role limitations due to physical problems (.38), though acceptable for all other scales (range .60–.89). Orthogonal and oblique factor analyses confirm a two-factor structure corresponding to dimensions of physical and psychological health. Validation of the scales against criterion groups defined on the basis of self-reported health indicate that physical functioning, social functioning and general mental health have good discriminant validity.

Conclusions. The study provides further evidence that the SF-36 may be used to investigate a range of physical and psychological problems even in relatively healthy populations. However some of the scales do not exhibit high levels of reliability and need careful consideration before being used. When physical and mental component summary scores are used, it is suggested that these are best derived from oblique factor solutions.

Recent years have witnessed an explosion in the development of methods for assessing various aspects of quality of life (Bowling, 1991, 1995; McDowell & Newell, 1987). These have included measures concerning both specific (Duncan & Sander, 1991; Meenan, 1982; Tursky, Jammer & Friedman, 1982) and more general facets of health (Hunt, McEwen & McKenna, 1985; Krupinski, 1980; Patrick & Deyo, 1989). The instruments

*Requests for reprints.

used encompass the assessment of both psychological and physical function. In the psychological domain this covers cognitive impairments (Copeland, 1990; Wechsler, 1986), mood states such as depression or anxiety (Beck, Rial & Rickels, 1974; McNair, Lorr & Droppleman, 1992; Zigmond & Snaith, 1983), and social and role performance (Corney & Clare, 1985). In the physical domain a host of instruments exist for assessing the physical functioning of patients with particular diseases, physical problems or disabilities (Ferrans & Powers, 1985; Schag, Ganz, Kahn & Peterson, 1992). In addition, many tools recognize the inseparable nature of physical and psychological morbidity and assess functioning from a multidimensional perspective (Nouri & Lincoln, 1987).

A number of criticisms have been levelled at quality of life measures. First of all it has been argued that item contents and scales have failed to reflect the diversity of people's impressions as to what constitutes being 'well', and instead have displayed an overconcern with morbidity and other negative aspects of health (WHO, 1987). Contrary to this, Hunt & McCleoud (1987), for example, emphasize the varying natures of what they loosely term health, fitness and well-being—an approach more in keeping with the WHO understanding of health as a state of complete social, physical and psychological well-being. A further complaint has addressed the question of who actually performs the ratings of psychological and behavioural function. Discrepancies between physicians' and patients' assessments have been found in a wide variety of circumstances, from severity of withdrawal from opiates (Turkington & Drummond, 1988) to quality of life following treatment for hypertension (Jachuk, Brierly, Jackuk & Wilcox, 1982). A further practical problem has concerned the length of the measuring instrument—shorter instruments have often meant sacrificing breadth, while more comprehensive instruments require too great an investment in time from respondents.

In the current climate of spiralling medical costs, theoretical issues regarding what are the most appropriate indicators of health, illness and well-being have great significance not only for evaluating the impact of medical care (Tarlov *et al.*, 1989) but also for use in predicting future demand for health care (Fitzpatrick *et al.*, 1992). There is also the question as to how they should be utilized and by whom. Indeed the debate revolving around the 'outcomes movement' is largely predicated upon the assumption of suitable evaluation technology being available. Such indicators of health therefore need to be sufficiently broad that they capture the diversity of health experience and sufficiently comprehensive that they are able to distinguish between respondents, whose quality of life differs with respect to positive as well as negative health states (Ware, 1990). The SF-36 (Ware & Sherbourne, 1992) emerged, after several earlier incarnations (Stewart, Hays & Ware, 1992) from the Rand Health Insurance battery of tests used in the Medical Outcomes Studies (Anderson, Sullivan & Usherwood, 1990; McHorney, Ware, Rogers, Raczek & Rachel, 1992; Tarlov *et al.*, 1989) as a candidate to provide measures of general health for use in both clinical and patient populations. Although relatively few studies have been published to date, it is becoming recognized as the instrument of choice not only in populations of patients but also as a general research tool in epidemiological work (Bowling, 1995). This in the UK at least stems from its apparent superiority over the Nottingham Health Profile (Hunt *et al.*, 1985) in its sensitivity to variations in health status within the populations in which it has been employed. Although there is no complete overlap between the variables covered by the Nottingham Health Profile and SF-36 and some heated debate has ensued over their relative merits (Hunt & McKenna,

1992; Brazier *et al.*, 1992*a*) it seems likely that the trend towards greater use of the SF-36 will accelerate. Work published to date has on the whole suggested good internal consistency, test–retest reliability (Brazier *et al.*, 1992*b*), item-scale discriminant validity and discriminant clinical validity (Ware, Snow, Kosinski & Gandek, 1993), although doubts have been expressed about the psychometric properties of some of the scales when used with particular clinical groups (Jenkinson, Peto & Coulter, 1996). Settings have varied from primary care (Brazier *et al.*, 1992*b*) and community postal surveys (Garrat, Ruta, Abdalla, Buckingham & Rutter, 1993) in the UK to elderly rural communities in the US (Hill & Harries, 1994).

The Whitehall II study of British Civil Servants (Marmot *et al.*, 1991), was set up to investigate the influences of occupation, life-style and biological factors upon health. One of the aims of the study was to utilize the scale scores generated by the SF-36 as outcome measures against which specific hypotheses regarding the causes of ill-health could be examined. Use of the SF-36 within a healthy population of workers such as are found in the Civil Service (Ferrie, Shipley, Marmot, Stansfeld & Davey Smith, 1995) should provide important information on the properties of the SF-36 as well as explore the ability of the tool to detect variations in psychological, social and physical functioning. Data describing variations in scale scores by gender, age and socio-economic position within this cohort has already been reported elsewhere (Hemingway, Nicholson, Stafford, Roberts & Marmot, 1997). The utility and importance of these findings depend on the assumed validity and reliability of the SF-36. The present paper will therefore evaluate the psychometric properties of the SF-36 within this Civil Service population. Part of this will involve exploration of the factorial structure of the SF-36, an aspect of the questionnaire which has not only received little critical assessment but has also formed the basis for a further short 12-item health survey (Ware, Kosinski & Keller, 1996). In addition work will address the capability of the instrument to discriminate between groups of people with a range of known physical, behavioural and psychological problems, thus extending the range of functional domains in which it has been applied.

Method

Participants and design

A cohort of 10 308 (6895 males and 3413 females) Civil Servants aged from 35 to 55 were originally recruited into the study between 1985 and 1988 (Marmot *et al.*, 1991). From our original list a response rate of 73 per cent was obtained, which, after allowing for those who had moved and were no longer eligible at the time of entry to the study, is likely to underestimate the true response rate by around 4 per cent. Between 1991 and 1993 further data collection was undertaken. Respondents who had participated in earlier phases of the study were recontacted and asked to attend for a medical screening examination and to complete a self-administered questionnaire. In this they were asked to provide details on a range of social and demographic variables, a variety of health status measures, health behaviours and a number of measures of psychological functioning (further details are given below). Of those recontacted, 8375 responded (5786 males and 2589 women), yielding a response rate of 81.1 per cent. Of this number 8213 (98.3 per cent) yielded complete scores on all the SF-36 scales. A subset of 289 participants were used for investigating test–retest reliability of the SF-36.

Questionnaire items

The standard UK version of the SF-36 was included in the questionnaire used. The SF-36 can be scored as eight subscales. Each scale score is constructed from a varying number of items. These are physical

functioning (10 items), social functioning (two items), role limitations due to physical problems (four items), role limitation due to emotional problems (three items), vitality (four items), bodily pain (two items), general health perceptions (five items), and general mental health (five items). One item concerns change in health and is not scored as a separate dimension. In addition the 30-item version of the General Health Questionnaire (Goldberg & Williams, 1988) was included in the questionnaire. Standard criteria for caseness are used after the chronic method of scoring (Goodchild & Duncan-Jones, 1985) is employed (i.e. score >12). The four CAGE screening questions for problem drinking were also utilized (Mayfield, McLeod & Hall, 1974), as was a self-reported measure of sickness absence from work. Use of these variables in the study is described below. Two different chronic illness conditions were used to investigate the validity of the SF-36: angina as assessed using the Rose Questionnaire (Rose, McCartney & Reid, 1977) and self-reported diabetes.

Analysis

1. To establish the internal consistency and reliability of the eight scales, data were subjected to multitrait scaling analysis with a version of the Multitrait Analysis Program (MAP), a Fortran program available for PC compatibles running under MSDOS (Hays, Hayashi, Carson & Ware, 1988). This software was originally developed for the Medical Outcomes Study and has the advantage of providing in a single run descriptive statistics for items and scales, internal consistency statistics (Cronbach's alpha) for each scale, tests for item convergent and discriminant validity, as well as inter-scale correlations, both in their raw form and adjusted for the number of items within the scale. Data are accepted in the form of standard ASCII files. Internal consistency scores less than unity indicate the margin of error in different possible versions of the same measurement. As such they may indicate one means by which the scale scores are imperfect measures of true scores.

Testing for convergent and discriminant validity involves three steps; whether items have equivalent variances, whether items within a scale are substantially related ($r > .40$) to the total score computed from other items in the scale (convergent validity) and whether items correlate more highly with other items from within its own scale than items from other scales (divergent validity). The following rules are recommended for ease of use and interpretation (IRCHCA, 1991). Scaling successes are indicated by whether the correlation between an item and its own scale is at least two standard errors higher than its correlation with other scales. Estimates of 'possible' scaling errors are determined by whether a correlation with other scales falls within two standard errors of the correlation with its own scale. 'Probable' scaling errors are deemed to have occurred when the correlation with other scales is greater than or equal to two standard errors of the correlation with its own scale.

2. Test–retest reliabilities with 95 per cent confidence intervals were calculated for each scale using linear correlation on a subset of 289 participants over a four-week period. In addition, the mean difference between the first and repeat measurement is calculated, along with the percentage of differences lying within 1.96 standard deviations of the mean. If the scale and the health it is measuring remains stable over the period of assessment then the mean should be zero.

3. Further psychometric testing of validity involved principal component and factor analyses of scale scores to assess their validity as measures of physical health and psychological well-being. Orthogonal and oblique factorial solutions are computed. This was performed using SAS proc FACTOR (SAS Institute, 1990). Kaiser's measure of sampling adequacy (the ratio of the squared correlations to the sum of squared correlations plus sum of squared partial correlations) is computed. The value tends towards one if the partial correlations are small—a prerequisite for a reliable factorial solution (Tabachnick & Fidell, 1989).

4. A number of clinical tests of the validity of the SF-36 scales were carried out.

(A) General linear models (SAS proc GLM) were used to assess the extent to which each of the scales discriminated between a variety of self-reported measures of physical health and psychological well-being. General linear models permit the use of both continuous and categorical variables within a linear regression framework. In SAS recoding of categorical variables possessing more than two levels is handled internally by the program without explicit construction of dummy variables by the user. In the present analysis a constant sample size across scales is used for each contrast by only including participants with no missing items on all the scales. Thus for each contrast, F ratios will indicate the degree of relevance of a scale to a particular criterion. The greater the F ratio the greater the information provided by a scale about the criterion, relative to error variance. The criterion comparisons used are (i) presence/absence of angina, (ii) presence/absence of

diabetes, (iii) severity of alcohol use—maximum CAGE score vs. other, (iv) GHQ caseness vs. not and (v) sickness absence greater than one week reported in past year vs. no reported sickness absence in past year. The numbers of people in the complete sample are: for angina $N = 249$, for diabetes $N = 65$, for GHQ caseness $N = 1946$, maximum CAGE score $N = 39$ (no score = 5754), for sickness absence greater than seven days $N = 3017$ (no sickness absence $N = 1602$).

(B) Multivariate models were developed to test the propositions that particular SF-36 scales distinguish between groups of people who differ in the type and severity of their health state. General linear models, adjusted for age and sex were constructed to estimate mean differences between contrasting groups for each of the eight scales. As above, a constant sample size across scales is used for each contrast by only including participants with no missing items on all the scales. Five mutually exclusive comparison groups were used. Two medical groups were used, comprising first of all people who report having angina, do not report diabetes and do not satisfy criteria for GHQ caseness ($N = 167$). The second medical group similarly comprised people with diabetes only ($N = 46$). A psychiatric only group consisted of people with no chronic illness (angina or diabetes) but who satisfy criteria on the GHQ. Two medical plus psychiatric groups were formed. The first consisting of those reporting both angina and satisfying GHQ caseness, the second of people reporting diabetes and satisfying GHQ caseness. Too few people reported both angina and diabetes for a serious medical group.

Comparisons were made between these five groups to enable tests of the validity of SF-36 scales. Comparisons examined the extent to which the scales can distinguish between groups with different chronic physical conditions (angina vs. diabetes), between medical and psychiatric conditions (angina vs. GHQ and diabetes vs. GHQ) and between medical and medical plus psychiatric conditions (angina vs. diabetes plus GHQ and diabetes vs. angina plus GHQ). As well as testing the validity of specific SF-36 scales in some of these comparisons (e.g. the general mental health scale should have high relative validity in distinguishing physical from psychological disorders), these comparisons provided information on the extent to which the SF-36 may yield wider disease-specific criteria for the above conditions.

In assessing clinical validity, the relative validity (RV) of a scale is calculated by the ratio of the variance of that scale relative to the scale with the greatest variance on the measured clinical construct.

Results

Internal consistency and reliability

Responses to the items within each of the scales tend to follow a similar distribution—although within the physical functioning scale responses to the first items are more evenly spread across the three alternatives—whilst responses to other items are skewed towards scores which indicate no limitations on physical functioning. This first item indicates that a large number of respondents ($N = 1586$, 19.3 per cent) report being limited a lot in performing vigorous physical activities, whilst another 3223 (39.2 per cent) report some limitations. As these figures are not adjusted for age, caution is required in their interpretation.

Item-scale correlations ranged from .31 to .72 (median = .65). Two possible scaling errors were found. An item in the general mental health scale correlated equally with the vitality scale ($r = .56$) and its own scale, whilst for one item in the physical functioning scale the magnitude of its correlation with general health perceptions ($r = .40$) was almost identical to that with its own scale ($r = .41$).

Table 1 shows the raw correlation coefficients for relationships between scales. These show a pattern of results similar to those reported by Brazier *et al.* (1992*b*), and which therefore suggest that the dimensions exhibit consistent relationships with one another across different populations.

Test–retest correlations are shown with 95 per cent confidence intervals in Table 2. All

Table 1. Correlations between SF-36 scales* ($N = 8213$)

	PF	SF	GHP	RLEP	RLPP	V	GMH	*p*
Physical functioning (PF)	**.85**							
Social functioning (SF)	.32	**.81**						
General health perceptions (GHP)	.40	.38	**.76**					
Role limitations emotional (RLEP)	.14	.53	.26	**.77**				
Role limitations physical (RLPP)	.40	.53	.38	.29	**.84**			
Vitality (V)	.31	.45	.54	.40	.36	**.84**		
General mental health (GMH)	.16	.49	.40	.53	.24	.64	**.79**	
Pain (P)	.45	.41	.38	.18	.48	.34	.23	**.75**

*All correlations $p < .0001$.
Note. Figures in bold indicate internal consistency.

Table 2. Test–retest reliability coefficients for SF-36 scales ($N = 289$, retest interval = 1 month)

	R	95 per cent confidence intervals	Mean difference (1st–2nd)	Percentage of participants within 1.96 SD of mean
Physical functioning	.60	0.51–0.70	−1.99	96.2
Social functioning	.60	0.51–0.69	−0.17	94.5
General health perceptions	.90	0.84–0.95	−0.56	88.6
Role limitations emotional	.60	0.51–0.70	−1.85	93.1
Role limitations physical	.38	0.27–0.49	0.17	89.6
Vitality	.81	0.74–0.88	−1.66	95.8
General mental health	.83	0.76–0.89	−0.79	93.1
Pain	.66	0.51–0.69	−1.41	92.0

Mean reliability coefficient = .67.

the scales tended to show a small improvement over the test–retest interval, though this was significant only for physical functioning and vitality ($p < .05$).

Principal component and factor analysis

Sampling adequacy for the analysis of the eight scales was high; MSA = 0.828 (Norman & Streiner, 1994). As expected principal component analysis yielded a common general health dimension, accounting for 46.18 per cent of the variance, with correlations with all scales ranging from .56 (physical functioning) to .77 (social functioning). The second component increased the explained variance to 61.95 per cent. The percentage variation in each scale explained by the two-factor solution ranged from just under 50 per cent (general health perceptions) to 75 per cent (general mental health). As can be seen in Table 3 varimax rotation of these components yielded a factor structure in which general mental health, role limitations due to emotional problems, social functioning and vitality correlated most highly with the first component, whilst physical functioning, bodily pain and role limitations due to physical problems did so with the second. General health perceptions loaded higher on the second component though also loaded moderately on

the first. Of the eight scales, general mental health was found to provide the best measure of the mental health component, with role limitations due to emotional problems lying closely behind. For the physical health component, both the bodily pain and physical functioning scales provided the best relative measure.

Oblique factor analysis (promax rotation) likewise yielded two correlated factors ($r = .41$) which can also be interpreted as alluding to physical and mental health. Besides the scales stated above, the rotated mental health factor now also contained high loadings from role limitations due to physical problems (.40) and general health perceptions (.54). In a similar manner the physical health factor also contained high loadings from social functioning (.58) and vitality (.49). Factor scores generated by the two different rotations correlated .97 ($p < .0001$) for the physical health factor and .98 ($p < .0001$) for the mental health factor.

Table 3. Factor loadings of SF-36 scales ($N = 8213$)

	Method of rotation				
	Orthogonal (Varimax)			Oblique (Promax)*	
Dimension	1[a]	2[b]	h^2	1[a]	2[b]
Physical functioning	.035	.785	61.7	.200	.755
Role limitations (physical)	.258	.715	57.8	.404	.753
Bodily pain	.132	.786	63.5	.296	.796
General health perceptions	.426	.564	49.9	.536	.641
General mental health	.865	.090	75.6	.865	.269
Role limitations (emotional)	.790	.052	62.6	.783	.217
Social functioning	.627	.459	60.3	.711	.580
Vitality	.717	.345	63.3	.775	.491

[a]Mental health factor.
[b]Physical health factor.
h^2 Percentage of total variance of each scale explained by the two extracted components.
*Correlation between factors = .41.

Clinical validity

For angina all scales distinguished significantly between those who report the condition and those who do not—with the physical functioning scale being the most discriminating. Other scales have low validity relative to this with the exception of general health perceptions (76 per cent). This latter scale provided the most discriminating response to diabetes. These two scales also detected significant differences between people with diabetes and those without, together with the physical functioning scale (RV = 39 per cent). Relative validity for the remaining scales fell below 10 per cent. As would be expected relative validity of the general mental health scale is highest for GHQ caseness. Role limitations due to emotional problems which might also be expected to have high relative validity for this variable has an RV of only 41 per cent, lower than for vitality (46 per cent). As was found with angina, significant differences were observed on all scales for GHQ status.

Two of the scales proved to be almost equally adept at discriminating between people

obtaining a maximum or zero CAGE score: social functioning (RV = 100 per cent) and role limitations due to emotional problems (RV = 93 per cent). Only role limitations due to physical problems and vitality failed to produce significant discrimination in this area. Relative validity concerning sickness absence was greatest for general health perceptions, though also high for pain (88 per cent) and social functioning (77 per cent). Moderate levels of relative validity were also found for physical functioning (61 per cent), role limitations due to physical problems (58 per cent) and vitality (54 per cent). General mental health and role limitations due to emotional problems, though possessing low relative validity, did in common with the other scales produce highly significant differences ($p < .0001$) between the two sickness absence groups.

Means and standard errors for groups differing in medical and psychological functioning are described in Table 4. Pairwise mean differences, F ratios, relative validity estimates and significance levels for the further clinical comparisons are shown in Table 5. From these it can be shown that although social functioning, general health perceptions and vitality have high relative validity (RV > .90) for discriminating between respondents with angina and diabetes no scale discriminated between the two conditions at a statistically significant level ($p > .15$ in all cases). This should perhaps not be considered too surprising when one considers that a correlation between the relative validity estimates of the scales across both conditions indicates that they are highly related ($r = .76$, $N = 8$, $p = .028$).

For both comparisons between medical and psychiatric groups further evidence was provided for the validity of the general mental health scale (RV = 1.00 in both cases). In addition a number of the scales showed significant discriminatory capacity. Four scales showed significant differences between the angina and GHQ groups. These were social functioning, role limitations due to emotional problems, vitality ($p < .0001$) and role limitations due to physical problems ($p < .01$). With the exception of the latter, these same scales also discriminated between the diabetes and GHQ groups ($p < .0001$ for social functioning and role limitations due to emotional problems; $p < .01$ for vitality).

Finally comparisons between medical and medical plus psychiatric groups again found the general mental health scale to give the highest relative validity estimates. Comparing diabetes to angina plus GHQ caseness found adjusted scores on all scales differentiated between the two groups ($p < .0001$ for all scales except bodily pain for which $p < .01$). In addition to general mental health only vitality (RV = .65) had a relative validity estimate above .50. For the contrast between the angina only group and the diabetes plus GHQ group all scales other than general mental health had relative validity estimates below .50. Significant differences were found on social functioning, general health perceptions and vitality ($p < .01$) as well as for role limitations due to emotional problems and role limitations due to physical problems ($p < .0001$).

Discussion

The work presented here shows the SF-36 scales to have high internal consistency and to exhibit a pattern of relationships across the scales which substantially mirror those found in previous studies. Similarly, the structure to emerge from both orthogonal and oblique factor analyses provides support for the physical and psychological dimensions to health which form the theoretical underpinning for the construction of the instrument. In

Table 4. Means and standard errors for groups differing in medical and psychological functioning (Group 1 excludes GHQc caseness diabetes) (Group 2 excludes GHQc, angina) (Group 3 excludes angina, diabetes) (Group 4 excludes diabetes) (Group 5 excludes angina)

	Medical (angina)		Medical (diabetes)		Psychiatric (chronic GHQ)		Psychiatric + medical 1 (GHQ + angina)		Psychiatric + medical 2 (GHQ + diabetes)	
	Group 1 $N = 167$ Mean age = 52.5		Group 2 $N = 46$ Mean age = 51.8		Group 3 $N = 1853$ Mean age = 48.8		Group 4 $N = 77$ Mean age = 50.8		Group 5 $N = 14$ Mean age = 53.3	
Scale	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Physical functioning	81.05	1.27	86.09	2.22	88.64	0.34	73.57	2.13	80.71	5.54
Social functioning	92.96	1.03	92.11	2.41	84.13	0.45	74.03	2.68	84.82	6.03
General health perceptions	64.33	1.46	66.54	2.74	66.78	0.41	55.53	2.20	51.14	5.64
Role limitations emotional	95.81	1.10	97.82	1.22	81.44	0.60	77.49	3.28	90.47	5.44
Role limitations physical	90.87	1.49	92.93	2.42	89.77	0.47	79.22	3.21	83.92	7.68
Vitality	60.87	1.15	65.54	2.78	52.10	0.40	43.51	2.06	47.14	5.80
General mental health	79.95	0.80	82.17	1.41	64.75	0.34	60.94	1.51	57.42	4.46
Pain	82.68	1.20	85.00	2.75	82.86	0.43	74.58	2.23	88.14	3.38

Table 5. Summary of clinical validity comparisons

	Group 1 vs. Group 2			Group 1 vs. Group 3			Group 2 vs. Group 3			Group 1 vs. Group 5			Group 2 vs. Group 4		
	Angina vs. Diabetes			Angina vs. GHQ			Diabetes vs. GHQ			Angina vs. (GHQ + Diabetes)			Diabetes vs. (GHQ + Angina)		
Scale	Mean diff.	*F*	RV	Mean diff.	*F*	RV	Mean diff.	*F*	RV	Mean diff.	*F*	RV	Mean diff.	*F*	RV
Physical functioning	−2.75	1.10	.59	−3.52	7.83**	.04	−0.73	0.1 0	.00	4.44	1.36	.02	13.11	15.32***	.07
Social functioning	3.55	1.82	.97	13.09	51.3***	.29	9.35	7.59**	.14	14.57	13.63**	.18	21.92	25.12***	.32
General health perceptions	−1.57	1.71	.91	−0.36	0.06	.00	0.47	0.03	.00	15.07	13.06**	.17	15.04	8.82***	.11
Role limitations emotional	1.98	1.42	.76	21.95	62.7***	.35	19.58	14.39**	.27	29.27	35.82***	.46	33.67	26.81***	.34
Role limitations physical	0.35	0.01	.01	6.30	6.92**	.04	5.26	1.44	.03	26.32	17.97***	.23	24.78	15.20***	.19
Vitality	−3.59	1.88	1.00	10.17	50.38***	.28	13.21	24.99***	.48	12.23	13.39**	.17	22.95	51.38***	.65
General mental health	−0.14	0.01	.01	16.45	179.9***	1.00	16.45	52.41***	1.00	21.93	77.77***	1.00	21.53	78.71***	1.00
Pain	0.89	0.10	.10	3.23	3.95*	.02	2.05	0.46	.01	2.73	0.59	.01	12.24	9.35**	.12

*$p < .05$; **$p < .01$; ***$p < .0001$.

addition the clinical validity tests have yielded results which suggest that the SF-36 may favourably be used in a number of settings where its validity had not been previously demonstrated. In considering the broad sweep of these findings however a number of further questions do arise.

The results from the principal component analysis whilst replicating those reported by the SF-36's authors (Ware *et al.*, 1993) can be argued to be counter-intuitive. The form of this statistical analysis necessarily implies that the dimensions are orthogonal. A substantial body of research in health psychology however provides evidence that physical and psychological functioning are strongly interrelated. Accordingly both Hyland (1992) and Cohen & Rodriguez (1995) have discussed a number of putative routes (biological, behavioural, cognitive and social) by which mental and physical health may become linked, emphasizing the bidirectional nature of these pathways. It must therefore be considered surprising that an orthogonal factor solution has been interpreted as supporting the validity of the SF-36. It is also recognized by many commentators (Kline, 1994; Tabachnick & Fidell, 1989) that orthogonal factor resolutions may present a distorted picture of influences which in real life may be related. Because of this the current authors favour an oblique factorial solution. The analysis presented here though confirming the proposed underlying mental and physical health constructs, will produce differences in factor scores. As development of the SF-12 (Ware *et al.*, 1996) stems from an orthogonal factor analysis of the SF-36, and is designed to measure the physical and mental components of health we would suggest that its further development would be better predicated upon a more accurate analysis of the relationship between these constructs. Correlations between factor scores generated by orthogonal and oblique analyses are extremely high and when physical and mental health scores are used independently there is likely to be little practical significance to which method is used to generate them. Where they are both used in the same analysis to explore other behavioural outcomes (e.g as confounding variables in a regression analysis) any putative results are likely to be distorted unless the factor scores were generated by methods which allow them to be correlated.

As mentioned the magnitude of the inter-scale correlations found here are similar to those obtained in a UK primary care setting (Brazier *et al.*, 1992*b*), although the relationship between vitality and general mental health in this cohort was found to be even higher ($r = .79$). Hunt & McKenna (1992) have cited the strength of this relationship as casting doubts on the validity of the SF-36. The present authors nevertheless consider that a moderate to high relationship between these variables should not be unexpected, given the low levels of motivation and general activity which are known to co-occur with depression (Alloy, Acocella & Bootzin, 1996). That vitality functioned as a more valid measure of the psychological health dimension than social functioning raises questions as to how high scores on this scale should be interpreted. Though the correlations between the four items which constitute the vitality scale and the general mental health scale (range .48–.61), are significantly less than the correlations with their own scale (range .63–.71) they are nevertheless moderately high. Stansfeld, Roberts & Foot (1997) found vitality to be related more strongly to the emotional reactions subscale of the Nottingham Health Profile than to either pain or physical abilities. Together these results suggest that what is measured as vitality on the SF-36 is more closely related to psychological than physical capacity—and as such may

be allied to motivation—indicating either high levels of motivation or the effect of motivation on perceived energy levels. The nature of the relationship between vitality and general mental health might be further clarified by data regarding sleep problems—an area of omission from the SF-36 and a manifest weakness of it.

Data from the clinical validity tests provide supporting evidence for the validity of several of the scales in specific areas of physical ill-health and poor psychological well-being—notably the physical functioning scale in discriminating between those with or without angina; the social functioning scale for those with alcohol problems; general mental health not only for those with psychiatric morbidity as assessed by the GHQ, but also in discriminating medical from psychological conditions; general health perceptions for discriminating people with diabetes from the rest of the Whitehall II cohort, as well as for those reporting differing periods of sickness absence. The role limitations due to emotional problems scale was almost as powerful as social functioning when addressing those with alcohol problems, which is as one would expect given the destructive effects which problem drinking is known to exert upon social life (Royal College of Psychiatrists, 1987). The remaining scales, role limitations due to physical problems, bodily pain and vitality, although not amongst the most powerful in the circumstances investigated, all proved capable of indicating significantly different levels of functioning in particular instances. This does not however by itself provide overwhelming evidence that the validity of these scales has been established.

One implication of the findings is that scales with the greatest validity in a given clinical or occupational setting may appropriately be used to evaluate outcomes in response to clinical or social interventions. Another is that the scales could be used as predictors for future positive health, morbidity and mortality and assessed for how effectively they predict service utilization. Scales with discriminant ability in specific areas even if not possessing the highest relative validity may still be employed in monitoring areas of functioning which without a multidimensional assessment of health might otherwise be overlooked. The data do suggest that a more prudent use of the SF-36 would be to choose scales on prior theoretical or psychometric grounds rather than simply using all eight. Indiscriminate use without prior theoretical justification should necessitate adjustment of critical p values by correction for the number of scales employed. To this end it is vital that researchers establish a comprehensive understanding of the discriminant and predictive ability of the instrument within specified populations.

The failure to locate significant differences on any of the SF-36 dimensions between the two chronic medical conditions we investigated (angina and diabetes) can be interpreted in several ways. One possibility is that chronic diseases affect diverse areas of psychosocial and physical functioning in broadly similar ways. While this may be true, the failure to detect differences probably arises as much from insufficient breadth in the SF-36 parameters. Known omissions in the range of functions assessed by the SF-36 include sleep, cognitive and sexual functioning; areas where one might expect to find differences between the above conditions. Symptoms of Type II diabetes for example include sexual dysfunction in both males and females (Taylor, 1995). Similarly Stansfeld *et al.* (1997) have argued that the social functioning scale of the SF-36 focuses more on social interaction outside work rather than in it. If angina and diabetes have different consequences for the performance of work roles then the picture revealed by these results may be an artifact. Some preliminary work in this cohort suggests that diabetes is linked

to poorer occupational mobility, whilst this is not true for angina (Roberts & Brunner, 1996). That dimensions of the SF-36 do not possess sufficient power to discriminate between the two conditions within the limits imposed by the current size of each group is also possible. However given the small magnitude of the existing differences between the groups, any real differences which might emerge as the cohort ages and respective numbers increase are unlikely to be clinically significant.

An area of particular concern for users of the SF-36 arises from the test–retest reliabilities for the SF-36 scales reported here, which were generally lower than those considered desirable (Norman & Streiner, 1989) and indeed lower than those reported by Brazier *et al.* (1992*b*). As the retest interval in the current study was four weeks and not two as in the Brazier *et al.* study, the existence of differences between them should not be considered surprising. In addition the two samples differ in their composition—particularly with respect to age and sex. Hemingway *et al.* (1996) have reported the SF-36 is sensitive to differences in these parameters. Of particular interest however is the low test–retest correlation ($r = .38$) for role limitations due to physical problems. This result may genuinely reflect low reliability for this scale although it is possible that it is sensitive to fluctuations in perceived role limitations stemming from transient physical problems. However if this possibility is correct it does suggest that a considerable degree of short-term physical problems would need to have existed in the cohort at the time of sampling to account for the low figure. McHorney, Ware, Lu & Sherbourne (1994) however point to floor effects in the role limitations scales, which if anything would suggest a low sensitivity to change in what is being measured. A further difficulty arises from the nature of the Whitehall II cohort itself, which, as has been remarked, is of above average health compared to the general UK population. Further analysis is required to ascertain if this result is repeatable or if it points to a serious limitation in the reliability of the scale over short retest intervals. It may be that the exact meaning of reliability in a particular population will need to be determined with respect to that population. As others (Bowling, 1995) have commented relatively low reliability may be the price one pays for high discriminatory power—however the relatively low magnitude of some of the reliability coefficients, only three are greater than .80 (general health perceptions, general mental health and vitality), needs to be carefully considered before using the scales. Repeated use of the SF-36 in future phases of the Whitehall II study should help to clarify these issues and as the cohort ages provide further information on its ability to predict and detect differences in psychological and physical well-being.

An important issue when considering the merits and limitations of the SF-36 concerns how generalizable the present results are. This is particularly important given that the Whitehall II population differs from any normative population sample in terms of age, sex, socio-economic status and health. McHorney *et al.* (1994) examined use of the SF-36 in 24 different groups differing in socio-economic characteristics, diagnosis and disease severity. Reliability coefficients obtained from these groups ranged from .64 to .94, which suggests the instruments may be satisfactorily used across a diverse range of people. Whether the lower reliabilities reported here are a characteristic of particularly healthy populations is unknown. Further evidence is required before any firm conclusions can be drawn. In conclusion we would contend that the current study has provided further evidence that the SF-36 may be fruitfully used to investigate a range of physical and psychological problems even in a relatively healthy population. Questions remain

about the psychometric properties of particular scales under particular circumstances, and the assumption of independence between physical and psychological dimensions held to underly the development of the instrument should be reconsidered.

Acknowledgements

We would like to thank the Agency for Health Care Policy Research (5 RO1 HS06516) and the New England Medical Center for funding the work on which this paper was undertaken. In addition we would like to thank all participating Civil Servants and their Welfare, Personnel and Establishment Officers, The Civil Service Occupational Health Service, Dr George Sorrie, Dr Adrian Semmence and the Council of Civil Service Unions. We acknowledge with thanks the assistance of funding from the Medical Research Council, the Health and Safety Executive, National Heart Lung and Blood Institute (2 RO1 HL36310), Ontario Workers' Compensation Institute, and the Henry J. Kaiser Family Foundation. We would also like to thank Eric Brunner, Rob Canner, Sara Foot, Merry Cross and the late Sol Levine. Particular thanks are also extended to Oonagh Fleming.

References

Alloy, L. B., Acocella, J. & Bootzin, R. R. (1996). *Abnormal Psychology: Current Perspectives*, 7th ed. Maidenhead, Berks: McGraw-Hill.

Anderson, J. St C., Sullivan, F. & Usherwood, T. P. (1990). The Medical Outcomes Study Instrument (MOSI)—Use of a new health status measure in Britain. *Family Practice*, 7(3), 205–218.

Beck, A. T., Rial, W. Y. & Rickels, K. (1974). Short form of depression inventory: Cross-validation. *Psychological Reports*, 34, 1184–1186.

Bowling, A. (1991). *Measuring Health: A Review of Quality of Life Measurement Scales*. Buckingham: Open University Press.

Bowling, A. (1995). *Measuring Disease. A Review of Disease Specific Quality of Life Measurement Scales*. Buckingham: Open University Press.

Brazier, J. E., Harper, R., Jones, N. M. B., O'Cathain, Thomas, K. J., Usherwood, T. & Westlake, L. (1992*a*). Validating the SF36. *British Medical Journal*, 305, 646.

Brazier, J. E., Harper, R., Jones, N. M. B., O'Cathain, Thomas, K. J., Usherwood, T. & Westlake, L. (1992*b*). Validating the SF36 Health Survey Questionnaire: New outcome measures for primary care. *British Medical Journal*, 305, 160–164.

Cohen, S. & Rodriguez, M. S. (1995). Pathways linking affective disorders and physical disorders. *Health Psychology*, 14(5), 374–380.

Copeland, J. R. M. (1990). Suitable instruments for detecting dementia in community samples. *Age and Ageing*, 19, 81–83.

Corney, R. H. & Clare, A. (1985). The construction, development and testing of a self-report questionnaire to identify social problems. *Psychological Medicine*, 15, 637–649.

Duncan, J. S. & Sander, J. W. A. S. (1991). The Chalfont Seizure Severity Scale. *Journal of Neurology, Neurosurgery and Psychiatry*, 54, 873–876.

Ferrans, C. E. & Powers, M. J. (1985). Quality of Life Index: Development and psychometric properties. *Advances in Nursing Science*, 8, 15–24.

Ferrie, J., Shipley, M., Marmot, M. G., Stansfeld, S. & Davey Smith, G. (1995). Health effects of anticipated job change and non-employment: Longitudinal data from the Whitehall II study. *British Medical Journal*, 311, 1264–1269.

Fitzpatrick, R., Fletcher, A., Gore, S., Jones, D., Spiegelhalter, D. & Cox, D. (1992). Quality of life measures in health care. I: Applications and issues in assessment. *British Medical Journal*, 305, 1074–1077.

Garratt, A. M., Ruta, D. A., Abdalla, M. I., Buckingham, J. K. & Rutter, I. T. (1993). The SF-36 health survey questionnaire: An outcome measure suitable for routine use within the NHS? *British Medical Journal*, 306, 1440–1444.

Goldberg, D. P. & Williams, P. (1988). *A User's Guide to the General Health Questionnaire*. Windsor: NFER-NELSON.

Goodchild, M. E. & Duncan-Jones, P. (1985). Chronicity and the General Health Questionnaire. *British Journal of Psychiatry*, 146, 55–61.

Hays, R. D., Hayashi, T., Carson, C. & Ware, J. E. (1988). *User's Guide for the Multitrait Analysis Program (MAP)*. Santa Monica, CA: Rand Corporation.

Hemingway, H., Nicholson, A., Stafford, M., Roberts, R. & Marmot, M. (1997). The impact of socio-economic status on health functioning as assessed by the SF-36: The Whitehall II study. *American Journal of Public Health* (in press).

Hill, S. & Harries, U. (1994). Assessing the outcome of health care for the older person in community settings: Should we use the SF-36? *Outcomes Briefing, UK Clearing House for the Assessment of Health Outcomes*, 4, 26–27.

Hunt, S. & McCleoud, M. (1987). Health and behavioural change: Some lay perspectives. *Community Medicine*, 9, 68–76.

Hunt, S. & McKenna, S. P. (1992). Validating the SF-36. *British Medical Journal*, 305, 645.

Hunt, S. M., McEwen, J. & McKenna, S. P. (1985). Measuring health status: A new tool for clinicians and epidemiologists. *Journal of the Royal College of General Practitioners*, 35, 185–188.

Hyland, M. E. (1992). A reformulation of quality of life for medical science. *Quality of Life Research*, 1, 267–272.

International Resource Center for Health Care Assessment (1991). *Memorandom on NEWMAP: Revised Multitrait Analysis Program Software (MAP-R)*. Boston: The Health Institute. New England Medical Center.

Jachuk, S. J., Brierly, H., Jachuk, S. & Wilcox, P. M. (1982). The effect of hypertensive drugs on the quality of life. *Journal of the Royal College of General Practitioners*, 32, 103–105.

Jenkinson, C., Peto, V. & Coulter, A. (1996). Making sense of ambiguity: Evaluation of internal reliability and face validity of the SF-36 questionnaire in women presenting with menorrhagia. *Quality in Health Care*, 5, 9–12.

Kline, P. (1994). *An Easy Guide to Factor Analysis*. London: Routledge.

Krupinski, J. (1980). Health and quality of life. *Social Science and Medicine*, 14A, 203–211.

McDowell, I. & Newell, C. (1987). *Measuring Health: A Guide to Rating Scales and Questionnaires*. Oxford: Oxford University Press.

McHorney, C. A., Ware, J. E., Lu, J. R. & Sherbourne, C. D. (1994). The MOS 36-item Short Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care*, 32, 40–66.

McHorney, C. A., Ware, J. E., Rogers, W., Raczek, A. E. & Rachel, J. F. (1992). The validity and relative precision of MOS short and long-form health status scales and Dartmouth COOP charts. *Medical Care*, 30, Supplement, 235–265.

McNair, D. M., Lorr, M. & Droppleman, L. F. (1992). *EdITS Manual for the Profile of Mood States (POMS)*. San Diego, CA: EdITS/Educational and Industrial Testing Service.

Marmot, M., Davey-Smith, G., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E. & Feeney, A. (1991). Health inequalities among British Civil Servants: The Whitehall II study. *Lancet*, 337, 1387–1393.

Mayfield, D., McLeod, G. & Hall, P. (1974). The CAGE questionnaire: Validation of a new alcoholism screening instrument. *American Journal of Psychiatry*, 131, 1121–1123.

Meenan, R. F. (1982). The AIMS Approach to health status measurement: Conceptual background and measurement properties. *Journal of Rheumatology*, 9, 785–788.

Norman, G. R. & Streiner, D. L. (1989). Health measurement scales: A practical guide to their development and use. Oxford: Oxford University Press.

Norman, G. R. & Streiner, D. L. (1994). *Biostatistics: The Bare Essentials*. St. Louis: Mosby.

Nouri, F. M. & Lincoln, N. B. (1987). An extended activities of daily living scale for stroke patients. *Clinical Rehabilitation*. 1, 301–305.

Patrick D. L. & Deyo, R. A. (1989). Generic and disease specific measures in assessing health status and quality of life. *Medical Care*, 27, Supplement, 217–232.

Roberts, R. & Brunner, E. (1996). Diabetes, health selection and occupational mobility. 10th European Health Psychology Society Conference, Dublin.

Rose, G., McCartney, P. & Reid, D. D. (1977). Self-administration of a questionnaire on chest pain: Intermittent claudication. *British Journal of Preventive and Social Medicine*, 31, 42–48.

Royal College of Psychiatrists (1987). *Alcohol: Our Favourite Drug*. London: Tavistock.

SAS Institute (1990). *SAS/STAT User's Guide*. SAS Institute.

Schag, C. A. C., Ganz, P. A., Kahn, B. & Peterson, L. (1992). Assessing the needs and quality of life of patients with HIV infection: Development of the HIV Overview of Problem Situations Evaluation System (HOPES). *Quality of Life Research*, 1, 397–413.

Stansfeld, S., Roberts, R. & Foot, S. (1997). Assessing the validity of the SF-36 general health survey. *Quality of Life Research* (in press).

Stewart, A. L., Hays, R. D. & Ware, J. E. (1992). The MOS Short-form General Health Survey. *Medical Care*, 26(7), 724–735.

Tabachnick, B. G. & Fidell, L. S. (1989). *Using Multivariate Statistics*. London: Harper Collins.

Tarlov, A. R., Ware, J. E., Greenfield, S., Nelson, E. C., Perrin, E. & Zubkoff, M. (1989). The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *Journal of the American Medical Association*, 262(7), 925–930.

Taylor, S. E. (1995). *Health Psychology*, 3rd ed. Maidenhead, Berks: McGraw-Hill.

Turkington, D. & Drummond, D. C. (1988). How should opiate withdrawal be measured? *Drug and Alcohol Dependence*, 24(2), 151–153.

Tursky, B., Jammer, J. D. & Friedman, R. (1982). The pain perception profile: A psychophysical approach to the assessment of pain report. *Behaviour Therapy*, 13, 376–394.

Ware, J. E. (1990). Measuring patient function and well-being: Some lessons from the medical outcomes study. In K. A. Heithoff & K. N. Lohr (Eds), *Effectiveness and Outcomes in Health Care*. Washington, DC: National Academy Press.

Ware, J. E. & Sherbourne, C. D. (1992). The SF36 Short Form Health Status Survey: 1. Conceptual framework and item selection. *Medical Care*, 30(6), 473–483.

Ware, J. E., Snow, K. K., Kosinski, M. & Gandek, B. (1993). *SF-36 Health Survey: Manual and Interpretation Guide*. Boston, MA: The Health Institute, New England Medical Center.

Ware, J., Kosinski, M. & Keller, S. D. (1996). A 12-item Short Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, 34(3), 220–233.

Wechsler, D. (1986). Wechsler Adult Intelligence Scale: Revised UK edition (WAIS-R UK). Sidcup: Psychological Corporation.

World Health Organization (1987). Measurement in Health Promotion and Protection. WHO Regional Publications, European Series No.22.

Zigmond, A. S. & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, 67, 361–370.

Received 3 September 1996; revised version received 14 April 1997