British Journal of Rheumatology 1998;**37:**862–869

COMPARISON OF THE MOS SHORT FORM-12 (SF12) HEALTH STATUS QUESTIONNAIRE WITH THE SF36 IN PATIENTS WITH RHEUMATOID ARTHRITIS

N. P. HURST, D. A. RUTA* and P. KIND†

*Economics & Health Outcomes Unit, Department of Rheumatology, Western General Hospitals Trust, Crewe Road, Edinburgh EH4 2XU, *Department of Epidemiology & Public Health, Ninewells Medical School, University of Dundee, Dundee and †Centre for Health Economics, University of York, York*

SUMMARY

Objective. To compare the performance of the MOS SF12 health survey (SF12) with the SF36 in a sample of 233 patients with rheumatoid arthritis (RA) stratified by functional class.

Methods. The SF12 and SF36 physical and mental component summary scales (PCS and MCS) were compared for test–retest reliability [intra-class correlation coefficient (RC) and repeatability], construct validity and responsiveness [standardized response mean (SRM)] to self-reported change in health.

Results. Overall, despite its brevity, the SF12 is comparable to the SF36 with only some loss of performance. The SF12-PCS is slightly less reliable (RC = 0.75) and responsive to improvements in health (SRM = 0.52) than the SF36-PCS (RC = 0.81; SRM = 0.61). The SF12-PCS correlates strongly with the SF36-PCS ($R = 0.94$), SF36 physical function subscale ($R = 0.77$) and modified Stanford Health Assessment Questionnaire (MHAQ) ($R = 0.71$), but only weakly with the SF36 mental health subscale ($R = 0.22$). SF12-PCS discriminated well between Steinbrocker functional classes; patients in functional classes 1–4, respectively, have SF12-PCS scores 1σ, 2σ, 2.4σ and 2.7σ below the population norm (ANOVA, $F = 35.8$, $P < 0.000$). The SF12-MCS is relatively unresponsive to reported improvement in RA (SRM = 0.31), but is reliable (RC = 0.71) and correlates well with the SF36-MCS ($R = 0.71$). SF12-MCS correlates more closely than the SF36-MCS with the SF36 mental health subscale ($R = 0.86$) and Hospital Anxiety and Depression (HAD) scale ($R = 0.76$). In ANOVA models, only the HAD ($R^2 = 57\%$) score contributes significantly to variance in SF12-MCS ($F = 254.8$; $P < 0.000$), but both the HAD ($R^2 = 24\%$) and MHAQ ($R^2 = 10\%$) scores contribute to variance in the SF36-MCS ($F = 50.9$; $P < 0.000$). Thus, the SF12-MCS has better construct validity for mental health than SF36-MCS in RA subjects. Missing responses to items were high amongst patients in functional class 4 (34%).

Conclusion. The SF12 is a reliable, valid and responsive measure of health status in the majority of RA patients, and meets standards required for comparing groups of patients. Its application in the most severely disabled subjects is uncertain.

KEY WORDS: Health status, Outcome, Rheumatoid arthritis, SF12, Validity, Responsiveness, Reliability.

THERE is growing interest in the use of generic health status questionnaires to provide broad measures of health which can be used either to provide normative population data [1], to compare the impact of different diseases and conditions on health [2], or to monitor the health of both individuals and groups over time [3]. We have been investigating the performance of two such instruments, EuroQol (EQ-5D) and the MOS Short Form-36 (SF36), in a sample of patients with rheumatoid arthritis (RA) stratified by functional class. The initial results and validation against condition-specific measures of arthritis have been reported elsewhere [4, 5].

More recently, a new shorter version of SF36, the SF12, has been described which utilizes only 12 items drawn from each of the eight subscales of the SF36 (Table I). The performance of the SF12 has been reported to be comparable to that of the SF36 [6] while having the advantage of being easier and quicker to complete.

Data from the SF36 may be presented as a profile describing health in each of eight separate dimensions. In addition, factor analysis has been used to show that the SF36 may be reduced to two dimensions—a physical component and mental component summary score (SF36-PCS and SF36-MCS) [7]—which are reported either as *T*-scores or *z*-scores to enable comparison with published norms. These summary scores facilitate hypothesis testing in clinical trials and reduce the risk, associated with multiple statistical comparisons between subscales, of significant findings arising by chance. The SF12 is also reported as either a physical (SF12-PCS) or mental component summary scale (SF12-MCS); the regression weightings used for scoring the SF12 come from a general population [8].

In this study, we have abstracted the SF12 item responses embedded in the SF36 and compared the performance of the complete SF36 with the shorter SF12. This approach gives results similar to those obtained when the SF12 and SF36 are administered separately [8].

METHODS

Patient population

The methods and study population have been reported elsewhere [4, 5]. In brief, a sample size of 240 RA patients [9] was selected on the basis that a relationship between any two measurements would be detected at the 5% significance level if their true correlation was >0.2, with an 80% power, and that a 20% drop-out rate would occur. The sample was stratified by functional class [10] to obtain a broad

Submitted 15 January 1998; revised version accepted 9 March 1998.

Correspondence to: N. Hurst.

cross-section of disease severity. To achieve this, recruitment of consecutive patients into each functional class continued until 60 patients had been entered in each class.

After completing the SF36 questionnaire, Hospital Anxiety and Depression (HAD) scale [11] and modified Stanford Health Assessment Questionnaire (MHAQ) [12], patients were asked to report co-existing medical conditions and drug side-effects. At follow-up 3 months later, patients again completed the questionnaires and were asked: 'Compared to three months ago is your arthritis better, the same or worse?'.

The study received institutional ethical approval and all patients gave written consent.

Scoring of SF36 and SF12

Published factor score coefficients were applied to calculate a *T*-score for the SF36-PCS and SF36-MCS [1, 7]. In *T*-score notation, scores are transformed such that the normal population mean = 50 with a standard deviation (σ) = 10. The SF12 was scored using published regression weights and scoring rules—in particular, if any SF12 item was missing, the SF12 summary scores were recorded as missing [8].

Assessing reliability

Reliability was assessed using test–retest methods. The difference in scores between two administrations of the questionnaire, in those patients reporting that their arthritis and overall health had remained the same over 3 months, was calculated for the SF36 and SF12. Results for the SF36 using test–retest over 2 weeks have been reported elsewhere [5]. Reliability is reported both as 'repeatability' [13] and as a reliability coefficient [14].

Repeatability, i.e. the size of score differences detectable with 95% confidence with repeated measurements in an individual patient, is given by (2.77 × within-subject standard deviation). If a difference in scores of this magnitude is found between repeated measurements on the same individual, it will represent a true difference on 95% of occasions [13].

TABLE I
SF-12 health survey questionnaire scales

	Number of items
I Functional status	
(a) Physical functioning	2
(b) Social functioning	1
(c) Role limitations attributable to physical problems	2
(d) Role limitations attributable to emotional problems	2
II Well-being	
(a) Mental health	2
(b) Energy and fatigue	1
(c) Pain	1
III Overall evaluation of health	
(a) General health perception	1
Total	12

In addition, calculation of a reliability coefficient (RC) [14] also permits direct comparison with published estimates of SF36 and SF12 reliability in conditions other than RA. The RC, which is derived from analysis of variance, is defined as RC = σ^2 pat/(σ^2 pat + σ^2 error) where σ^2 pat is the estimated variance due to patients and σ^2 error is the estimated error variance. Values of RC thus vary from 1 (perfectly reliable) to 0 (totally unreliable). A coefficient exceeding 0.5 is considered acceptable when a measure is used to compare groups of patients [15], although coefficients exceeding 0.9 have been recommended when making comparisons between individual patients or assessing change in scores in an individual over time [16]. The 95% CIs for the RC were calculated as described previously [14].

Assessing validity

The construct validity of the SF12 as a measure of health status in patients with RA was assessed in several ways. First, SF12-PCS and SF12-MCS scores were correlated with the SF36-PCS and SF36-MCS, each of the eight SF36 subscales, the HAD and MHAQ scales. Analysis of variance (ANOVA) and linear regression models were then used to examine the relationship between summary scores and severity of RA as measured by functional class and MHAQ, along with other important covariates such as age, duration of disease, reported co-morbidity and drug side-effects.

Assessing responsiveness to change

The standardized response mean (SRM), which is a measure of 'signal to noise', is defined as the ratio of mean change (δ) to the standard deviation (σ) of the change scores (i.e. δ/σ change) in the population of patients reporting change [17]. An SRM was calculated for SF12-PCS and SF12-MCS in the group of patients reporting improvement in arthritis over a 3 month period. These are compared directly with SRMs for SF36 subscales and SF36-PCS and SF36-MCS previously reported for the same group of patients [5]. The 95% CI for the SRM were calculated using the 95% CI for the mean change in score.

RESULTS

Patient characteristics

The age and disease duration (Table II) have been reported before [4, 5], and are reproduced here for convenience.

Validity of SF12

SF12-PCS. The overall distributions of SF12-PCS and SF36-PCS scores are very similar (Fig. 1). There is also a strong correlation between the SF12-PCS and the SF36-PCS ($R = 0.94$) (Fig. 2a and Table IIIa). The SF12-PCS also correlates strongly, as does the SF36-PCS, with each of the physical subscales of SF36 and the MHAQ, and conversely correlates weakly with the SF36 mental health subscales and the HAD scale (Table III). Change in health status was measured between baseline and 3 months; there was a strong

TABLE II
Patient characteristics

Functional class	*n*	Age		Duration of RA	
		yr (σ)	(range)	yr (σ)	(range)
I	60	49 (14)	(24–77)	5 (7)	(0.15–30)
II	63	53 (15)	(21–80)	11 (12)	(0.2–65)
III	60	59 (12)	(26–87)	16 (11)	(1–40)
IV	50	65 (11)	(39–86)	23 (14)	(4–58)
Males	45	58 (13)	(26–79)	9 (8)	(0.2–29)
Females	188	55 (15)	(21–87)	14 (13)	(0.2–65)
Total	233	56 (14)	(21–87)	13 (13)	(0.2–65)

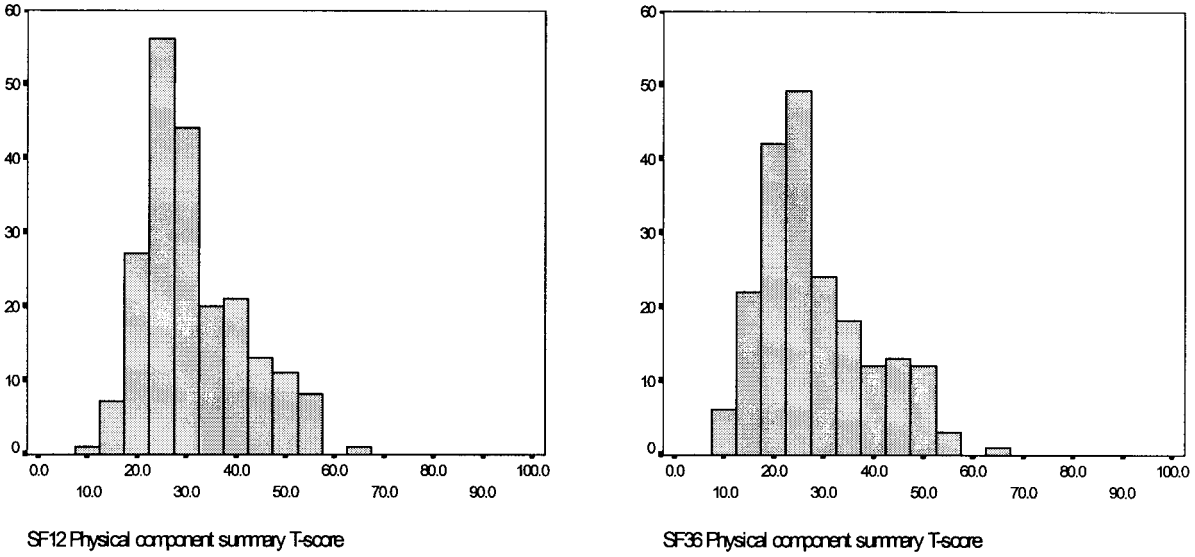

FIG. 1.—Distribution of *T*-scores for SF12 and SF36 physical component summary scales.

correlation between the change scores for SF12-PCS and SF36-PCS ($R = 0.87$; $P < 0.01$) (Fig. 2b).

There is a marked decline in SF12-PCS and SF36-PCS with worsening functional class (Table IV). Comparison of the *F*-statistics shows that SF36-PCS is better than SF12-PCS in discriminating between functional class (relative validity = 0.73). The relationship between MHAQ and either SF12-PCS or SF36-PCS, and the effect of other covariates—age, duration of RA, HAD score, and the presence or absence of co-morbidity or drug side-effects—was explored using ANOVA. This showed that SF12-PCS and SF36-PCS are both significantly related to MHAQ, and age is the only covariate to contribute significantly to the models. In linear regression models, MHAQ and age predict 51% of the variance in SF12-PCS ($F = 99.3$, $P < 0.000$) and 60% of the variance in SF36-PCS ($F = 145$, $P < 0.000$).

SF12-MCS. Only a moderate correlation was seen between the SF12-MCS and the SF36-MCS ($R = 0.71$) (Fig. 2a, Table IIIb). However, SF12-MCS correlates closely with the SF36 mental health subscale and HAD scale, and weakly with the subscales measuring physical attributes (Table IIIb).

The relationship between HAD scores and either SF12-MCS or SF36-MCS, and the effect of other covariates—age, duration of RA, MHAQ score, and the presence or absence of co-morbidity or drug side-effects—was again explored using ANOVA. SF12-MCS was closely related to HAD scores, but SF36-MCS was closely related to both HAD and MHAQ scores. None of the other covariates contributed significantly to the models. In linear regression models, HAD predicts 57% of the variance in SF12-MCS ($F = 254.8$, $P < 0.000$), while the HAD and MHAQ scores predict 24 and 10%, respectively, of the variance in SF36-MCS ($F = 50.9$, $P < 0.000$).

A moderate correlation was seen between change scores for SF12-MCS and SF36-MCS ($R = 0.66$; $P < 0.01$) (Fig. 2b).

Responsiveness

The mean change scores and standardized response means, measured in patients reporting improvement in their arthritis over 3 months, for SF12-PCS and SF12-MCS are comparable to those for SF36-PCS and SF36-MCS, respectively (Table Va).

Reliability

Reliability coefficients and estimates of 'repeatability', measured in patients reporting no change in

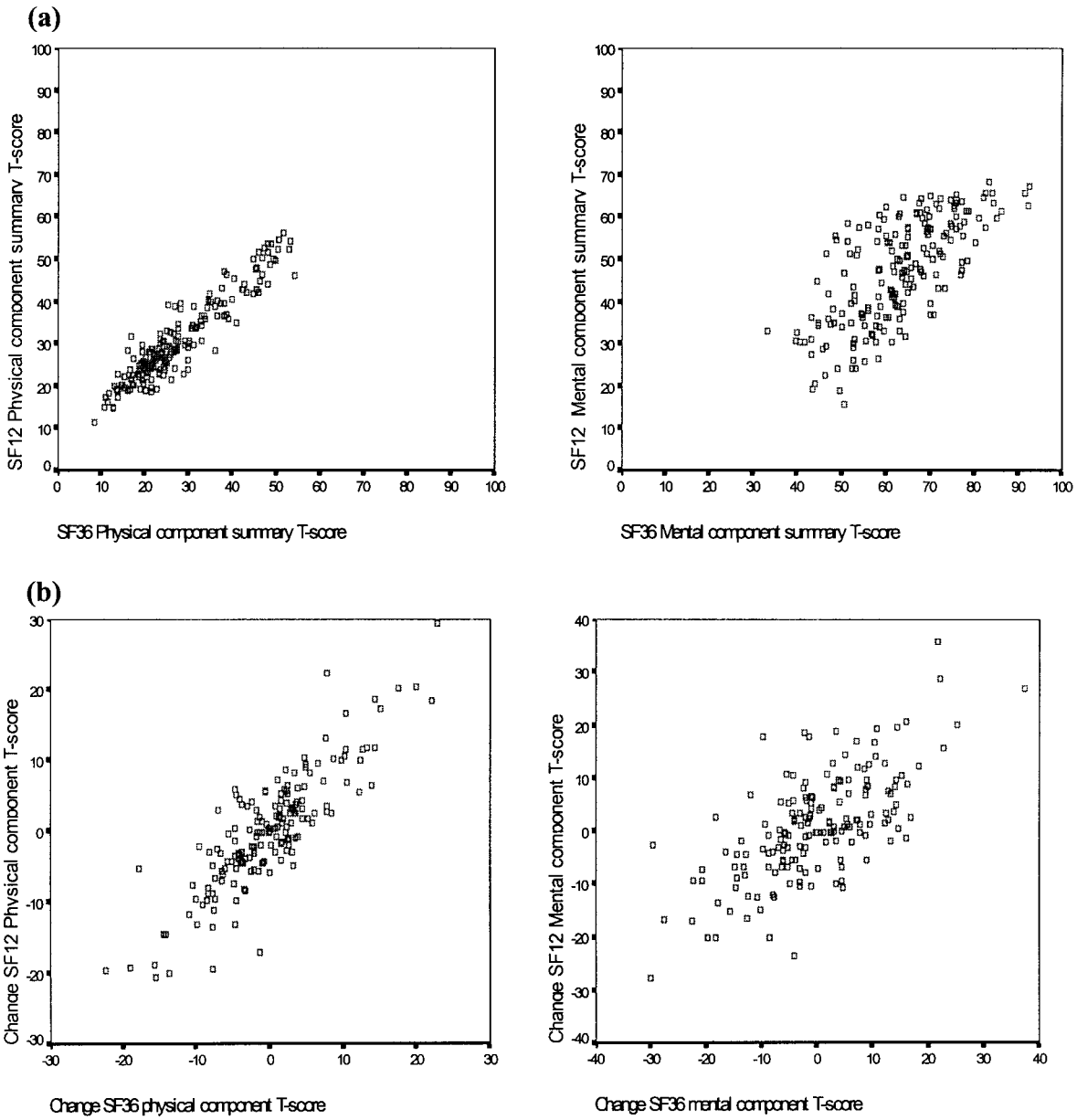

FIG. 2.—Correlation between SF12 and SF36 summary *T*-scores. (a) Correlation between *T*-scores for: (i) SF12-PCS *vs* SF36-PCS ($n = 191$) Pearson $R = 0.94$; (ii) SF12-MCS *vs* SF36-MCS ($n = 191$) Pearson $R = 0.71$. (b) Correlation between change in *T*-scores for: (i) SF12-PCS *vs* SF36-PCS ($n = 164$) Pearson $R = 0.87$; (ii) SF12-MCS *vs* SF36-MCS ($n = 164$) Pearson $R = 0.66$.

arthritis over 3 months, were very similar between SF12 and SF36 scales (Table Vb).

Sample size calculations

The sample sizes required to detect different size changes on each scale were obtained using the standard deviation of the change scores (Table Va) and a published nomogram [18]. Total samples required to detect different effect sizes at different power (90 or 95%), at a significance level of $P < 0.05$, are shown (Table VI). It can be seen that to detect comparable effect sizes, appreciably larger samples are needed with the SF12-PCS compared with the SF36-PCS.

Missing data

The data set of 36 items used in the SF36, which includes the 12-item subset used in the SF12, was examined for missing responses. Overall, there were missing data from 5 (8%), 4 (6%), 13 (22%) and 17 (34%) patients in functional classes 1–4, respectively. For each of these patients, at least one item was missing from the SF12 subset, thus preventing calculation of the SF12 summary scores.

Amongst patients in functional class 1 or 2, there were very few missing data. In class 1, one patient had clearly missed a page of 12 items by mistake, and another five patients had not completed one item each.

TABLE III
Correlation (Pearson coefficient) between (a) SF12-PCS or SF36-PCS scales and (b) SF12-MCS or SF36-MCS with MHAQ, HAD and SF36 subscales

(a) Physical component summary scales

	SF12-PCS	SF36-PCS
SF12-PCS	1.00**	0.94**
MHAQ scale	−0.71**	−0.77**
HAD scale	−0.37**	−0.34**
SF36 subscales		
Physical function	0.77**	0.84**
Role limitations—physical	0.77**	0.77**
Bodily pain	0.74**	0.78**
General health	0.54**	0.62 **
Energy/vitality	0.55**	0.55**
Social function	0.63**	0.62**
Role limitations—emotional	0.21**	0.20**
Mental health	0.22**	0.16*

(b) Mental component summary scales

	SF12-MCS	SF36-MCS
SF12-MCS	1.00	0.71**
MHAQ scale	−0.39**	−0.08 ns
HAD scale	−0.76**	−0.49**
SF36 subscales		
Physical function	0.36**	−0.12 ns
Role limitations—physical	0.39**	−0.12 ns
Bodily pain	0.49**	−0.27**
General health	0.61**	0.24**
Energy/vitality	0.71**	0.32**
Social function	0.66**	0.28**
Role limitations—emotional	0.84**	0.59**
Mental health	0.86**	0.67**

ns, not significant; *$P < 0.05$; **$P < 0.01$.

In class 2, one patient failed to complete one item and three patients failed to complete four items each out of the 36 questions.

In functional classes 3 and 4, however, there was a substantial number of missing responses relating to three questions (eight items) which relate to the impact of health problems on usual work or activities over the last 4 weeks. Five of these eight items are part of the SF12 subset (Table VII). Of the remaining 29 items, 26 had between one and three missing responses, and three had between one and four missing responses.

DISCUSSION

The SF12 health survey is designed to be quick to use while retaining the validity of the parent SF36 and the capacity to distinguish between the health of groups of subjects of different age, gender and with different conditions [6]. The loss of reliability associated with fewer defined health levels was regarded as an acceptable trade-off with practicality and length in the context of large group studies. The purpose of this present analysis was to examine the performance of the SF12 as compared to the SF36 in a sample of patients with RA stratified by functional class.

The validity of SF12-PCS was shown by its close correlation with SF36-PCS, the SF36 physical function subscale and the MHAQ, but weak correlation with the SF36 mental health subscale. A close correlation was also observed between the change scores for the SF12-PCS and SF36-PCS. Furthermore, the SF12-PCS discriminated almost as well as SF36-PCS between Steinbrocker functional classes: classes 1–4, respectively, had mean SF12-PCS scores 1σ, 2σ, 2.4σ and 2.7σ below the population norm. The decline in SF12-PCS and SF36-PCS score across functional classes was shown to be related to loss of physical function, rather than disease duration or other important covariates. Thus, 51% of the variance in the SF12-PCS score and 60% of variance in the SF36-PCS score was explained by MHAQ score with only a small contribution due to age. These results, taken together with our previous reports on the validity of SF36 in RA, suggest that SF12-PCS is a valid measure of physical health status in RA.

In tests of construct validity, the SF12-MCS was found to be much more closely related to measures of mental health than the SF36-MCS. MHAQ explained 10% of the variance in the SF36-MCS, showing that in part SF36-MCS is measuring physical disability. Thus, the SF12-MCS appears to be a better measure of mental health in patients with RA than the SF36-MCS. It is not clear why the SF12-MCS performs better than the SF36-MCS, but this may be due to less 'contamination' with items weighted more towards the physical health of RA patients. Whatever the reason, it is clearly advantageous to be able to identify mental health problems since these contribute significantly to poor perceived health and disability in various patient populations. In a recent study, for example, it was

TABLE IV
Relationship between SF12-PCS, SF12-MCS and Steinbrocker functional class

Functional class	SF12-PCS	SF36-PCS	SF12-MCS	SF36-MCS
1	41.4 (9.5)	40.2 (9.7)	53.1 (9.4)	60.3 (13.8)
2	30.0 (7.2)	26.7 (7.1)	50.3 (10.1)	66.0 (9.0)
3	26.3 (5.4)	21.7 (5.3)	39.9 (11.8)	61.2 (12.0)
4	23.4 (5.1)	20.3 (6.2)	43.6 (14.9)	63.2 (13.0)
F statistic†	35.8**	49.3**	8.8**	4.0*
RV‡	0.73	1.0	0.18	0.08

†*F* statistic from ANOVA (main effect = functional class, corrected for age and duration of RA).
‡RV = relative validity (ratio of *F* statistics compared with 'best' scale).

TABLE V
(a) Baseline score, mean change and standardized response means (SRM) for SF12 and SF36 summary scales in patients reporting improvement in arthritis over 3 months

Health scale	*n*	Baseline score (S.D.)	Mean change (S.D.)	SRM (95% CI)
SF12-PCS	42	31.4 (9.1)	4.1 (7.9)	0.52 (0.20, 0.84)
SF36-PCS	42	29.9 (10.4)	4.3 (7.1)	0.61 (0.30, 0.92)
SF12-MCS	42	49.9 (12.0)	2.6 (8.3)	0.31 (0.02, 0.66)
SF36-MCS	42	63.9 (11.0)	3.5 (9.9)	0.35 (0.04, 0.67)

(b) Baseline score, mean change, reliability coefficients (RC), and repeatability for SF12 and SF36 summary scales in patients reporting no change in arthritis over 3 months

Health scale	*n*	Baseline score (S.D.)	Mean change (S.D.)	RC (95% CI)	Repeatability*
SF12-PCS	75	32.0 (9.7)	0.20 (7.6)	0.75 (0.64, 0.87)	13.1
SF36-PCS	70	28.7 (10.7)	−0.10 (7.1)	0.81 (0.69, 0.93)	12.0
SF12-MCS	75	47.4 (12.3)	1.0 (9.4)	0.71 (0.60, 0.83)	16.2
SF36-MCS	70	62.6 (11.7)	0.27 (9.5)	0.74 (0.62, 0.86)	16.2

*Repeatability = size of score difference detectable with 95% confidence.

TABLE VI
Sample sizes required to detect different effect sizes on the SF12 and SF36 summary scales (paired sample)

Health scale	δ*	Standardized difference†	Total sample 95% power	Total sample 90% power
SF12-PCS	2.5	0.6	140	110
	5	1.3	34	28
SF36-PCS	2.5	0.7	105	82
	5	1.4	26	20
SF12-MCS	2.5	0.6	140	110
	5	1.2	36	28
SF36-MCS	2.5	0.5	210	170
	5	1.0	55	43

*δ = effect size (scale difference to be detected).
†Standardized difference = $2\delta/\sigma$, where σ = standard deviation of change score.

TABLE VII
Missing responses to SF36 transition questions relating to health and usual activities over the last 4 weeks

	Missing responses by functional class			
SF36/12 question	I	II	III	IV
During the past 4 weeks have you had any of the following problems with your work or other regular activities as a result of your physical health?				
(a) Cut down the amount of time you spent on work or other activities	0	0	6	9
(b) Accomplished less than you would like*	0	2	6	9
(c) Were limited in the kind of work or other activities*	2	1	7	7
(d) Had difficulty performing the work or other activities	0	2	5	6
During the past 4 weeks have you had any of the following problems with your work or other regular activities as a result of any emotional problems?				
(a) Cut down the amount of time you spent on work or other activities	0	3	8	13
(b) Accomplished less than you would like*	0	2	9	11
(c) Didn't do work or other activities as carefully as usual*	0	2	10	13
During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?*	0	0	4	5

*Items which are included in the SF12 subset.

found that mental and physical health problems contributed almost equally to variance in perceived health amongst rheumatology out-patients whether or not they have inflammatory joint disease (N.P. Hurst *et al.*, unpublished). For the investigator wishing to monitor or measure both physical and mental health, the SF12 has a clear advantage over the alternative and more cumbersome approach of using separate instruments such as the MHAQ and HAD scales or the separate subscales of the SF36. The MHAQ and HAD used together, excluding the subsections of the MHAQ, contain a minimum of 22 questions.

The reliability of SF12 measured as an intra-class correlation coefficient was comparable to the SF36 summary scores. Thus, reduction of scale length from 36 to 12 items does not seriously affect reliability. As far as its use in individuals is concerned, the SF12 is little different from the SF36 and only fairly large-scale changes of 13 (i.e. 1.3σ) may be detected with 95% confidence between two time points. Over several time points, though, trends in individual scores may be just as useful as trends in other measures such as joint scores or the erythrocyte sedimentation rate.

Responsiveness of the SF12-PCS and SF36-PCS measured by the SRM was very comparable and considerable better than the SRM for the modified Stanford Health Assessment Questionnaire [4, 5]. Estimates of the sample sizes required to detect change not surprisingly favoured the SF36 over the SF12, but were not substantially greater than for the SF36. The scale changes (effect sizes) of 2.5 and 5 used in these calculations are realistic, since mean change scores of about four were observed in this study amongst patients reporting improvement in their health. Comparing the SF12 and SF36 summary scales, there is clearly a trade-off between the disadvantage of the slightly reduced responsiveness and reliability of the SF12 and its advantage of being brief. In practical terms, SF12 requires a sample size $\sim$30% greater than the SF36, which is three times as long. From the patient's perspective, SF12 is a clear winner and the small increase in sample size should pose no great practical problem to the researcher or auditor seeking to measure modest changes in health.

The problems associated with non-completion of certain items relating to 'usual activities or work' has been commented on previously [5]. These questions clearly pose a problem for patients with severe disability, some of whom either disregard or cannot complete them. Since these items are included in the SF12, the applicability and reliability of the SF12 in very severely disabled subjects must be questionable, and it should therefore be used and interpreted with caution in such subjects.

An important issue is the role and usefulness of instruments such as the SF12 in the routine clinical setting and whether they have any advantage over more traditional instruments such as the HAQ or MHAQ. The aims of measuring health status include measurement of change in health, i.e. outcome, comparison of the health of different groups of patients with perhaps disparate diseases, or assisting in the identification of problems which may bear on clinical decision making. The MHAQ (or HAQ) is widely used to measure disability in RA and to a lesser extent in other rheumatic diseases. The advantage of the MHAQ is that it provides a quick guide to identifying which activities of daily living (ADL) are problematic for the patient and may help to guide the occupational therapist. Although it is quite reliable, it is not very sensitive to change [4, 5] and therefore of less use for measuring change over time. No normative values are available so adjustment for age, sex and other demographic variables is difficult. In contrast, the SF12-PCS is less useful than the MHAQ as an inventory of ADL problems, but is more sensitive to change in physical function over time. The SF12 would therefore be a better instrument for following change in physical function during anti-rheumatic therapy than the MHAQ. Since mental health is also a determinant of disability, the availability of a mental health score, as well as a physical health score, from the SF12 is an added advantage. Recognition and treatment of concomitant mood disorder in rheumatic disease, which may even sometimes be the dominant problem, is important and well known to most rheumatologists. Providing a quantitative measure of mental health may improve clinical recognition and help even the most experienced consultant to avoid overlooking a mood disorder. Finally, the fact that the SF12-PCS and SF12-MCS scores are expressed as *T*-scores makes them very easy to interpret. For example, an SF12-MCS score of 40, compared with the population mean of 50, represents a 1σ reduction and would suggest the possibility of a mental health problem; a score of 30 (i.e. 2σ reduction) is highly indicative. While the SF12 can clearly be used to monitor or audit the health change of small groups of patients, it is not clear whether it can be used in the individual patient. In practice, though, none of the measures used by rheumatologists are very reliable and clinicians tend to rely on a composite picture made up of several laboratory and clinical markers. Where there is uncertainty, clinicians often repeat these measures, thereby improving confidence limits and providing rough estimates of trends or change. Although in theory there is no reason why the SF12 or other short instrument should not be used on successive clinic visits, the fear is that in practice patients will become disenchanted and fed up with 'form filling'. To address this question, a randomized clinical trial to test the acceptability and usefulness of a health status instrument such as the SF12 in a routine setting would be of considerable interest.

In conclusion, the SF12 is a useful and valid measure of health in all but those with severely disabling RA, it is easy and quick to use, but slightly less reliable and less responsive than the SF36. The expression of SF12 in terms of *T*-scores allows immediate interpretation by comparison with population norms. Whether for the purpose of audit or clinical trials, the SF12 may be used as a simple generic measure of health alongside or instead of more traditional condition-specific meas-

ures of health outcome in patients with RA. The role of SF12, and other similar instruments, as an adjunct to clinical decision making and monitoring in individual patients needs further investigation.

Acknowledgements

This work was supported by a grant from the Chief Scientists Office of the Scottish Home & Health Department. The authors are very grateful to our two research assistants, Mrs Hunter and Mrs Stubbings, who collected and collated questionnaire data, and to all the patients who willingly gave of their time to complete the various assessments.

References

1. Jenkinson C, Coulter A, Wright L. Short form 36 (SF36) health survey questionnaire: normative data for adults of working age. Br Med J 1993;306:1437–40.
2. Greenfield S, Rogers W, Mangotich M *et al.* Outcomes of patients with hypertension and non-insulin dependent diabetes mellitus treated by different systems and specialties. J Am Med Assoc 1995;274:1436–44.
3. Hemingway H, Stafford M, Stansfield S, Shipley M, Marmot M. Is the SF-36 a valid measure of change in population health? Results from the Whitehall study. Br Med J 1997;315:1273–9.
4. Hurst NP, Kind P, Ruta D *et al.* Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of Euroqol (EQ-5D). Br J Rheumatol 1997;36:551–9.
5. Ruta D, Hurst NP, Kind P *et al.* Measuring health status in British patients with rheumatoid arthritis: reliability, validity, and responsiveness of the SF-36 health survey (SF-36). Br J Rheumatol 1998;37:425–36.
6. Ware JE, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey. Construction of scales and preliminary tests of reliability and validity. Med Care 1996;34: 220–33.
7. Ware JE, Kosinski M, Keller SD. SF-36 Physical and Mental Health Summary Scales: A user manual. Boston, MA: The Health Institute, New England Medical Center, 1994.
8. Ware JE, Kosinski M, Keller SD. SF12: How to score the SF12 Physical and Mental Health Summary Scales, 2nd edn. Boston, MA: The Health Institute, New England Medical Center, 1995.
9. Arnett FC, Edworthy SM, Bloch DA *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
10. Steinbrocker O, Traeger CH, Betterman RC. Therapeutic criteria in rheumatoid arthritis. J Am Med Assoc 1949;140:659–62.
11. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scand 1983; 67:361–70.
12. Pincus T, Summey JA, Sorraci SA *et al.* Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
13. Bland JM, Altman DG. Measurement error. Br Med J 1996;313:744.
14. Streiner GL, Norman DR. Health measurement scales: a practical guide to their development and use, 2nd edn. Oxford: Oxford University Press, 1996.
15. Helmstadter GC. Principles of psychological measurement. New York: Appleton-Century Crofts, 1964.
16. Kelly TL. Interpretation of educational measurements. Yonkers: World Books, 1927.
17. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171–8.
18. Altman DG. Practical statistics for medical research. London: Chapman & Hall, 1991.