
Article:
Brazier, J.E., Tsuchiya, A., Roberts, J. et al. (1 more author) (2004) A comparison of the EQ-5D and the SF-6D across seven patient groups. Health Economics, 13 (9). pp. 873-884. ISSN 1057-9230

https://doi.org/10.1002/hec.866

HEALTH ECONOMICS
Health Econ. **13**: 873–884 (2004)
Published online 9 January 2004 in Wiley InterScience (www.interscience.wiley.com). **DOI:**10.1002/hec.866

A comparison of the EQ-5D and SF-6D across seven patient groups

John Brazier[a,*], Jennifer Roberts[a], Aki Tsuchiya[a,c] and Jan Busschbach[b,c]
[a] *Sheffield Health Economics Group, University of Sheffield, Sheffield, UK*
[b] *Institute for Medical Technology Assessment (iMTA), Erasmus University, Rotterdam, The Netherlands*
[c] *The EuroQol Group*

Summary

As the number of preference-based instruments grows, it becomes increasingly important to compare different preference-based measures of health in order to inform an important debate on the choice of instrument. This paper presents a comparison of two of them, the EQ-5D and the SF-6D (recently developed from the SF-36) across seven patient/population groups (chronic obstructive airways disease, osteoarthritis, irritable bowel syndrome, lower back pain, leg ulcers, post menopausal women and elderly). The mean SF-6D index value was found to exceed the EQ-5D by 0.045 and the intraclass correlation coefficient between them was 0.51. Whilst this convergence lends some support for the validity of these measures, the modest difference at the aggregate level masks more significant differences in agreement across the patient groups and over severity of illness, with the SF-6D having a smaller range and lower variance in values. There is evidence for floor effects in the SF-6D and ceiling effects in the EQ-5D. These discrepancies arise from differences in their health state classifications and the methods used to value them. Further research is required to fully understand the respective roles of the descriptive systems and the valuation methods and to examine the implications for estimates of the impact of health care interventions. Copyright © 2004 John Wiley & Sons, Ltd.

Keywords preference-based measures of health

Introduction

Preference-based measures of health, sometimes called multi-attribute utility scales, are standardised multi-dimensional health state classifications which come with pre-existing preference or utility weights [1] and generate a single index score for each state of health where full health is one and zero is equivalent to death. Preference-based measures of health have become an important set of instruments for estimating the health state values used to calculate quality adjusted life years (QALYs) and are widely used in economic evaluations alongside clinical trials to value the benefits of health care.

There are currently an array of preference-based measures, including the EQ-5D, HUI, 15D, AQoL, QWB and most recently the SF-6D. These preference-based measures of health differ considerably in terms of their dimensions, items and preference weights and there is therefore no reason why they should generate the same values for a given patient. A recent review of these measures

*Correspondence to: Sheffield Health Economics Group, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK. E-mail: j.e.brazier@sheffield.ac.uk

Received 24 January 2002
Accepted 25 September 2003

against the criteria of practicality, reliability and validity of the measures concluded that what is required, among other things, is a series of head to head comparisons of the preference-based measures across a range of conditions and severity [2].

This paper presents a comparison of two of them, the EQ-5D and the SF-6D (recently developed from the SF-36) across seven patient/population groups (chronic obstructive airways disease, osteoarthritis, irritable bowel syndrome, lower back pain, leg ulcers, post menopausal women and elderly). These are two of the most widely used general measures of health and this paper presents the most extensive comparison of them. The paper begins by presenting a brief description of the instruments, before setting out the methods and results of the comparisons across the seven patient groups. The discussion seeks to explore the possible reasons for any divergence and to explore the implications for their use.

EQ-5D and the SF-6D

The EQ-5D instrument was developed by a multidisciplinary group of researchers from seven centres across five countries [3]. The five dimensions are mobility, self-care, usual activities, pain/discomfort and anxiety/depression (see Table 1) [4]. They each have three levels and together define 243 health states. Patients are classified onto the EQ-5D by self-completion or interviewer administration. The most influential valuation work to date with the EQ-5D has been a large-scale survey undertaken in the UK by the Measurement and Valuation of Health group at York [5,6]. Their work elicited trade-off (TTO) values for 42 health states defined by the EQ-5D using 2997 interviews of members of the general population. The 243 health states of the EQ-5D have been valued using regression methods on this sample. Separate algorithms are available for different socio-demographic groups. However, it is the so-called A1 tariff (based on 10-year TTO valuations on the whole survey sample) that is referred to in this paper[a] (See Table 2). Valuation studies using the MVH protocol have been repeated in Spain ($n = 979$), Germany ($n = 339$), Japan ($n = 621$) [7–9]. It has become one of the most widely used generic measures of health in Europe and has become commonly used in economic evaluation.

The SF-36 is another generic measure of health that generates scores across eight dimensions of health [10]. The eight dimensions are: physical functioning, role limitation due to physical problems, social functioning, bodily pain, role limitations due to emotional problems, mental health and vitality. It has become one of the most widely used generic measures of health through out the world, including the USA where it was originally developed, but was not originally developed for use in economic evaluation.

A research team at the University of Sheffield in collaboration with Dr Ware at Boston has estimated a preference-based single index measure of health from the SF-36 [11]. The index is estimated via a health state classification called the SF-6D derived from the SF-36 and is composed of six multi-level dimensions of health. It was constructed from a sample of 11 items selected from the SF-36 to minimise the loss of descriptive information and defines 18 000 health states. A selection of 249 states defined by the SF-6D have been valued by a representative sample of the UK general population ($n = 611$) using the standard gamble (SG) valuation technique. Like the EQ-5D, regression models were estimated to predict single index scores for all health states defined by the SF-6D. The resultant algorithm can be used to convert SF-36 data at the individual level to a preference-based index.[b]

These instruments have previously been reviewed against the criteria of practicality, reliability, and validity. Both instruments have been found to be practical to use in terms of rates of response to the questionnaire and levels of completion. The instruments have been found to be reliable between test and re-test [12], though there is evidence that rates of completion of the SF-36 and hence the SF-6D decline with age [13]. The assessment of *validity*, however, is rather more controversial in this area since there is no gold standard for assessing the social value of health care interventions [12].

This paper does not directly assess the validity of these instruments, but addresses the question of whether or not the two can be used interchangeably to estimate the health state value of patients with different medical conditions. It does this by comparing the two measures in t erms of the indices they generate across seven patient samples and examining the distribution of responses across their health state classifications.

Table 1. The SF-6D and EQ-5D

Level	SF-6D		EQ-5D
	Physical Functioning		*Mobility*
1	Your health does not limit you in *vigorous activities*	1	No problems walking about
2	Your health limits you a little in *vigorous activities*	2	Some problems walking about
3	Your health limits you a little in *moderate activities*	3	Confined to bed
4	Your health limits you a lot in *moderate activities*		*Self care*
5	Your health limits you *a little in bathing and dressing*	1	No problems with self-care
6	Your health limits you *a lot in bathing and dressing*	2	Some problems washing or dressing myself
		3	Unable to wash or dress self
	Role limitations		
1	You have *no* problems with your work or other regular daily activities as a result of your physical health or any emotional problems		
2	You are limited in the kind of work or other activities as a result of your physical health		
3	You accomplish less than you would like as a result of emotional problems		*Usual activities*
4	You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems	1	No problems with performing usual activities (e.g. work, study, housework, family or leisure activities)
		2	Some problems with performing usual activities
	Social functioning	3	Unable to perform usual activities
1	Your health limits your social activities *none of the time*		
2	Your health limits your social activities *a little of the time*		
3	Your health limits your social activities *some of the time*		
4	Your health limits your social activities *most of the time*		
5	Your health limits your social activities *all of the time*		
	Pain		*Pain/discomfort*
1	You have *no* pain	1	No pain or discomfort
2	You have pain but it does not interfere with your normal work (both outside the home and housework)	2	Moderate pain or discomfort
3	You have pain that interferes with your normal work (both outside the home and housework) *a little bit*	3	Extreme pain or discomfort
4	You have pain that interferes with your normal work (both outside the home and housework) *moderately*		
5	You have pain that interferes with your normal work (both outside the home and housework) *quite a bit*		
6	You have pain that interferes with your normal work (both outside the home and housework) *extremely*		
	Mental health		*Emotions*
1	You feel tense or downhearted and low *none of the time*	1	Not anxious or depressed
2	You feel tense or downhearted and low *a little of the time*	2	Moderately anxious or depressed
3	You feel tense or downhearted and low *some of the time*	3	Extremely anxious or depressed
4	You feel tense or downhearted and low *most of the time*		
5	You feel tense or downhearted and low *all of the time*		
	Vitality		None
1	You have a lot of energy *all of the time*		
2	You have a lot of energy *most of the time*		
3	You have a lot of energy *some of the time*		
4	You have a lot of energy *a little of the time*		
5	You have a lot of energy *none of the time*		

Methods

Data set

The data set consists of 2436 cases, and covers a wide rage of patients/populations: lower back pain chronic obtrusive pulmonary disease irritable bowel syndrome leg ulcer menopausal women osteo arthritis, and healthy older women aged 75+. These patients' samples are described in more detail below in terms of the study source, demographic variables and key severity indicators for each condition.

Lower back pain (LBP; number of observations = 265). These patients were recruited from General Practices in York (UK) into a randomised clinical trial of the cost effectiveness of alternative treatments for lower back pain [14]. Patients were included if they were aged 20 to 65 presenting with low back pain of at least 4 weeks duration. The mean age of the participants was 43 and 61% were women. On average they had been experiencing back pain for 17 weeks prior to recruitment, with most people experiencing pain every day. The data set consists of the baseline assessment and follow-up at 3 months following intervention using the SF-36 and EQ-5D.

Chronic obstructive pulmonary disease (COPD; number of observations = 284). Patients aged 35 years and over, and clinically judged to have COPD were recruited in a teaching hospital in Sheffield, UK [15]. Patients with a clinical diagnosis of asthma, lung fibrosis and pulmonary malignancy were excluded. Also excluded were those whose spirometeric tests gave $FEV_1 > 70\%$ FVC or $FEV_1 < 70\%$ FVC but with demonstrable reversibility. The data set consists of observations from 156 individuals, of which 76 were male and 80 were female, and mean ages (SD) were 67 (10.4) years and 62 (10.3) years respectively. Of these, 128 were participated in the follow-up observations 6 months later.

Irritable bowel syndrome (IBS; number of observations = 322). A sample of 161 IBS sufferers known to their GPs were recruited from six GP practices in Trent Region in the UK [16]. The practices themselves were chosen to be representative geographically and on the basis of deprivation level and social class. All suitable patients were sent a letter of explanation and an invitation to attend consultation for recruitment. This involved a 15 min screening consultation with a project researcher who took a medical history of their bowel symptoms. The diagnosis of IBS was determined by the use of the Rome Criteria. Their mean age was 47 and 86% were female. All patients were observed twice.

Leg ulcer (Lu; number of observations = 434). All respondents were recruited into a randomised control trial of the cost effectiveness of alternative treatments for venous leg ulcers from eight community-based clinics [17]. The inclusion criteria were: a venous ulcer below the knee and down to the foot that had lasted for three months, agreement to travel to the clinic for the trial, and written informed consent. The instruments were administered at 12 weeks and 12 months following recruitment. The overall response rate was 86%. Most of the sample was over retirement age and two thirds were women. More than half had a mobility problem and three quarters complained of leg ulcer pain. The median patients had a baseline ulcer area of 5.6 cm^2 and an ulcer that had lasted for 7 months.

Menopausal women (Mp; number of observations = 293). One thousand and eighty women aged 45–60 were randomly selected from six General practices in Sheffield, UK and sent a postal questionnaire containing the EQ-5D, SF-36 and questions a about menopausal symptoms [18]. Of these 758 (73%) were returned and had useable data. The sample used in this paper was the 293 who reported having menopausal symptoms as defined by the presence of hot flushes. The average age of the sample was 53.

Osteoporosis (OA; number of observations = 458). These patients were recruited from five UK clinics across two distinct settings: a knee replacement waiting list and a rheumatology clinic [19]. From these groups, patients were restricted to those patients with a diagnosis of OA of the knee made by a rheumatology or orthopaedic specialist. No further inclusion or exclusion criteria were applied, so that the recruited patients were likely to be representative of those seen in everyday hospital clinical practice in the UK. Patients were asked to complete the EQ-5D and SF-36 at baseline and at a 6 month follow-up. Both assessments have been used in this data set. The

Table 2. Comparing the SF-6D and EQ-5D predictive models

SF-6D Mean model		EQ-5D Random effects model	
c	1.000	*C*	0.919
PF23	−0.056	MOB2	−0.069
PF4	−0.072	MOB3	−0.313
PF5	−0.080		
PF6	−0.134	SC2	−0.104
		SC3	−0.213
RL234	−0.073		
		UA2	−0.036
SF2	−0.080	UA3	−0.094
SF3	−0.082		
SF4	−0.091		
SF5	−0.107		
PAIN23	−0.052		
PAIN4	−0.076		
PAIN5	−0.107	PA2	−0.123
PAIN6	−0.179	PA3	−0.385
MH23	−0.062		
MH4	−0.121	MOOD2	−0.071
MH5	−0.136	MOOD3	−0.237
VIT234	−0.017		
VIT5	−0.043		
Most	−0.032	N3	−0.269
N	249		35964
adj R^2	0.409		0.460
Predictive ability			
MAE	0.074		0.039
% > |0.05|	46		29
% > |0.10|	22		7
t(mean = 0)	−1.612		−0.571
JB	1.505		0.005
LB	123.00*		9.027

Note: Most data here is *not* directly comparable across models.
MAE = mean absolute error. JB = Jarque-Bera test [28].
LB = Ljung-Box, test result significant at $t_{0.001}$ [29].
Both models fail Reset and heteroscedasticity tests.
Models are estimated with White's heteroscedasticity consistent standard errors.
All coefficient estimates are significant at $t_{0.10}$.
The EQ-5D model presented here is formally equivalent to that presented in [5].
For mobility, for example, MOB2 = MO, and MOB3 − MOB2 = MO + M2.

mean age of respondents was 67 and around three quarters were female. Out of those attending the rheumatology clinic 41% were classified by the clinicians as having severe disease. Completion rates exceeded 90% for both the instruments.

Healthy older women (number of observations = 380). All respondents were women recruited into a pilot randomised trial of a treatment for osteoporosis from four General practices in Sheffield, UK [20]. The inclusion criteria for this study were all women aged 75 or over willing to participate in the trial. Women were excluded for a range of medically related conditions (e.g. receiving treatment for concurrent malignancy, women with bilateral hip arthroplasties or internal fixations of the neck and so forth), but the aim was to make the trial available to the vast majority of elderly women. The respondents were Caucasian females with a mean age of 80.1 and 49% living alone. Eighty six percent stated that they had a long limiting standing illness or disability compared to 73% in the General Household Survey, but overall they did not differ from the same age group in terms of use of health services.

These samples were chosen to represent a range of patient types in terms of dimensions of health effected by the medical condition, such as those that impact significantly on physical functioning (such as OA and COPD), pain (LBP and IBS) and mental health (Mp). They also represent a range in terms of severity, from those with severe COPD through to a general population sample (the older sample). The seven samples provide a good opportunity to understand the relationship between the two instruments over a wide range of health problems.

The overall data set has been formed by combining 'baseline' and 'follow-up' observations from five of the seven studies, but no distinction is made in the following analysis. Four of the studies did not have a specific intervention, and overall the changes in health that did take place are too small for the purposes of this paper.

Analysis

EQ-5D dimensions are what the respondents reported, and EQ-5D indices are derived using the MVH standard (A1) algorithm discussed above. SF-6D dimensions are obtained from the SF36 responses and the indices estimated using the methods described above. The analysis is carried out in two stages. The first is based on SF-6D and EQ-5D self-reported health classifications, with no

reference to single index scores. This includes an assessment of the degree of agreement between dimension of the two instruments using the Spearman rank correlation across the whole sample and by patient group. The distribution of responses across the dimensions in the two instruments is also examined for the whole sample and by patient group. Those patients who are classified into the lowest or highest health state for each instrument are looked at more closely in terms of their distribution of responses across the other instrument.

The second stage in the analysis will introduce the preference-based indices. Basic descriptive statistics including means, medians and ranges are compared by instrument and by patient group. The degree of agreement is also examined by calculating the single measure intraclass correlation coefficient (ICC). This is done by using the reliability analysis in SPSS ver. 10 (two-way random effects model based on absolute agreement). Further, to examine the nature of the relationship more closely, a series of OLS regressions have been run to explore the relationship between SF-6D indices and EQ-5D indices. The models used are:

$$\text{SF-6Dindex} = \alpha + \beta_1 \text{EQ-5Dindex} + u \quad (1)$$

$$\text{SF-6Dindex} = \alpha + \beta_1 \text{EQ-5Dindex} + \beta_2 \text{N3} + u \quad (2)$$

The regression analysis is not intended to suggest that EQ-5D index numbers can or should be used to predict SF-6D indices, or that SF-6D indices can or should be explained in terms of EQ-5D indices. It merely provides another way of understanding the relationship between them.

Results

Dimension-to-dimension comparison

Table 3 summarises the relationship between the dimensions of the EQ-5D and the SF-6D in terms of Spearman correlation coefficients. There is evidence for the convergent validity between similar dimensions: between physical functioning and mobility, between role limitation and social functioning with usual activities, between pain and pain/discomfort, and between mental health and anxiety/depression. These correlations exceeded the correlations between these and other dimensions. The lowest correlations are observed between mental health and mobility, between mental health and self care, and between physical functioning and anxiety/depression, of the SF-6D and the EQ-5D respectively.

Table 3. The correlation between SF-6D levels and EQ-5D levels

EQ-5D SF-6D	*M*	SC	UA	PD	AD
PF	**0.58**	0.51	**0.58**	0.39	<u>0.14</u>
RL	0.40	0.30	**0.50**	0.33	0.42
SF	0.47	0.43	**0.56**	0.38	0.34
P	0.46	0.39	0.59	**0.60**	0.30
MH	<u>0.12</u>	<u>0.17</u>	0.25	0.21	**0.55**
V	0.34	0.33	0.45	0.32	0.35

The correlations between like dimensions are indicated in bold. The three least correlated dimension pairs are underlined.
EQ-5D dimensions: *M*: mobility; SC: self care; UA: usual activities; PD: pain/discomfort; AD: anxiety/depression.
SF-6D dimensions: PF: physical functioning; RL: role limitation; SF: social functioning; *P*: pain; MH: mental health; V: vitality.

Table 4 presents the distribution of SF-6D and EQ-5D results. SF-6D indicates bimodal distributions in physical functioning, role limitation, social functioning, and to some extent pain. This reflects the heterogeneity in severity across the seven data sets. Breaking this information down by disease groups (not shown) indicates that for example, while the majority of IBS and of Mp patients report levels 1 or 2 physical functioning, an overwhelming 92% of OA patients report level 6 on this dimension. However, none of this is reflected in the EQ-5D results. A large proportion of responses lie within levels 1 and 2, and for example, only 0.2% of the whole sample (and none of them OA patients) report level 3 mobility. Overall, there is very little use of level 3 in four of the five dimensions of EQ-5D. This suggests that many patients feel that an 'extreme' problem in EQ-5D is much worse than any of the worst levels of the SF36 items used in SF-6D.

The EQ-5D has a larger proportion of respondents in the top category of each dimension than for the SF-6D (i.e. 3.6% to 34.6% compared to 17.1% to 72.2%). This suggests that the EQ-5D is not capable of distinguishing between health states close to full health. The extent of the distribution across the SF-6D is somewhat negated by the aggregation of some levels in the consistent version of the SF-6D model. The ceiling effect can be

Table 4. SF-6D and EQ-5D results

(a) Distribution of SF-6D (%) Level	Physical functioning	Role limitation	Social functioning	Pain	Mental health	Vitality
1	7.4	**28.0**	**34.6**	9.3	17.8	3.6
2	**21.7**	22.4	14.4	13.9	26.1	18.9
3	18.6	11.1	**24.5**	**23.9**	**37.2**	**32.6**
4	14.2	**38.4**	17.7	20.5	14.1	22.4
5	11.4	—	8.8	**23.8**	4.8	22.4
6	**26.6**	—	—	8.5	—	—
Total (*n*)	2339	2994	2320	2332	2333	3213

(b) Distribution of Eq-5D (%) Level	Mobility	Self care	Usual activities	Pain/discomfort	Anxiety/depression
1	39.9	**72.2**	35.8	17.1	**52.5**
2	**59.9**	26.8	**53.7**	**66.0**	43.6
3	0.2	1.0	10.5	16.9	3.9
Total (*n*)	2324	2295	2303	2313	2307

Modal level is in bold.
The distribution adds up to 100% by columns.

further explored by selecting those reporting full health in one instrument to see what they report in the other instrument. There are only 5 respondents who report full health in SF-6D, and all of them also report full health in EQ-5D, so there is not much to be said of this subgroup. However, there are 214 observations where the patient reports full health in EQ-5D, of which only 12 (6%) report full health in SF-6D. This means that, for the great majority of those reporting full health in EQ-5D, their health was not full in SF-6D. Table 5 summarises the SF-6D responses of these 214 by dimension. This indicates that those in full health in terms of EQ-5D may still have problems in physical functioning, mental health, and vitality. Vitality is not an EQ-5D dimension, so this to some extent is expected. Of interest is that of those who report 11111 in EQ-5D, 60% feels 'tense or downhearted and low' a little (level 2) or some (level 3) of the time.

There is only one observation (75+) where EQ-5D = 33333 and 12 observations where SF-6D = 645655 (2 with COPD, 2 with Lu, 7 with OA, and 1 from 75+). These respondents do not necessarily report the worst state in the alternative instrument, but the numbers are too small to merit generalisation. However, it is clear from the distribution across the individual dimensions that the SF-6D has larger proportion on its lowest level of physical functioning and role limitation than does the EQ-5D on mobility and usual activities (i.e. 24.6% and 38.4% versus 0.2% and 10.5%). For pain and mental health the proportions are more alike (at 8.2% and 4.8% versus 16.9% and 3.9%).

The preference-based index numbers

The SF-6D mean index exceeds the EQ-5D mean index across the whole sample by 0.045 (Table 6). The mean difference varies between –0.015 for patients suffering symptoms of the Menopause to 0.094 for patients with leg ulcers. The mean difference is within 0.05 for all patient groups except leg ulcers and osteoarthritis. ICC between the indices for the whole sample was 0.51, ranging from 0.28 for COPD and 0.55 for IBS.

For six out of the seven patient groups the mean SF-6D index exceeds the EQ-5D. By contrast, the median EQ-5D score exceeds the SF-6D in six patient groups and by –0.082 for the whole sample. This reversal in the sign of the difference reflects the different distributions of indices produced by the two measures. The ranges differ markedly, with the range for the EQ-5D covering –0.4 to 1.0 compared to 0.3 to 1.0 for the SF-6D. This negative skew in the EQ-5D data

Table 5. Ceiling effects of EQ-5D. Distribution of SF-6D of those with EQ-5D = 11111

Level	Physical functioning	Role limitation	Social functioning	Pain	Mental health	Vitality
1	29.0	**76.7**	**82.1**	**45.8**	35.2	7.5
2	**49.8**	10.0	10.8	32.1	**38.5**	**51.4**
3	11.1	8.6	5.2	18.9	22.1	29.2
4	2.4	4.8	1.4	2.4	2.3	9.4
5	3.4	—	0.5	0.9	1.9	2.4
6	4.3	—	—	0.0	—	—
Total (*n*)	207	210	212	212	213	212

Modal level is in bold. The distribution adds up to 100% by columns.

Table 6. SF-6D and EQ-5D indices by disease group

		n	Mean	SD	Median	Minimum	Maximum	ICC of indices
Whole	EQ5D index	2298	0.586	0.309	0.691	−0.594	1.000	0.51
	SF6D index	2192	0.631	0.149	0.609	0.300	1.000	
LBP	EQ5D index	265	0.636	0.266	0.691	−0.181	1.000	0.53
	SF6D index	263	0.658	0.144	0.634	0.370	1.000	
COPD	EQ5D index	255	0.540	0.309	0.620	−0.349	1.000	0.28
	SF6D index	230	0.572	0.112	0.577	0.296	0.944	
IBS	EQ5D index	314	0.662	0.260	0.725	−0.077	1.000	0.55
	SF6D index	296	0.666	0.146	0.628	0.373	1.000	
Lu	EQ5D index	431	0.552	0.307	0.620	−0.239	1.000	0.50
	SF6D index	430	0.647	0.145	0.626	0.296	1.000	
OA	EQ5D index	428	0.442	0.336	0.587	−0.239	1.000	0.38
	SF6D index	404	0.521	0.114	0.499	0.296	0.948	
75+	EQ5D index	320	0.614	0.299	0.691	−0.594	1.000	0.49
	SF6D index	291	0.662	0.141	0.651	0.296	1.000	
Mp	EQ5D index	285	0.729	0.262	0.796	−0.181	1.000	0.53
	SF6D index	278	0.716	0.143	0.718	0.370	1.000	

results in the mean being lower than the median in six out of seven cases. For the SF-6D mean exceeds the median in all patient groups, though the size of the difference is somewhat less. The EQ-5D indices are also associated to higher standard deviations.

A more detailed inspection of the plot of EQ-5D to SF-6D reveals other marked differences between the two measures (Figure 1). Whilst the overall correlation is 0.66, the pattern is not linear. There is a considerable degree of dispersion, with negative indices on the EQ-5D being associated with values on the SF-6D as high as 0.75. Furthermore, SF-6D indices being spread across a much narrower range compared to EQ-5D indices has a floor effect with a significant proportion of the SF-6D at or near the lowest possible value being associated with a large range of EQ-5D values. Conversely, the ceiling effect of EQ-5D relative to SF-6D can be observed in the wide range in SF-6D indices of those with EQ-5D index = 1.00. Some of these respondents can have quality adjustments by SF-6D as low as 0.56.

Another pattern is for EQ-5D indices to cluster. This is especially so at the upper extreme, where there is a large gap between EQ-5D index = 1.00 and those less than 1.00. Finally, there is another clear gap in the EQ-5D indices around 0.45. All observations to the right of this gap have N3 = 0 in the MVH A1 tariff, and all those to the left have N3 = 1, indicating that the N3 term used in the modelling is causing this bimodal distribution of EQ-5D indices.

Table 7 summarises the first and second regression models on the whole sample, and the second model by patient group. The patient specific

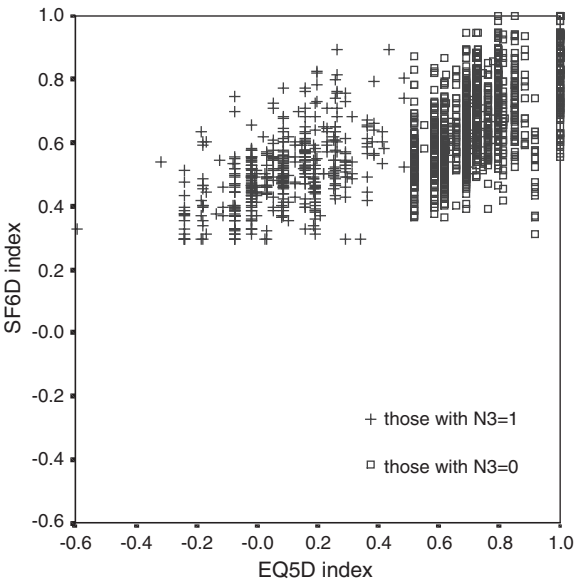

Figure 1. SF-6D indices vs EQ-5D indices

Table 7. Regressing SF-6D indices on EQ-5D indices

	Adjusted R^2	α	β_1	β_2
Whole (1)	0.44	0.44	0.33	—
Whole (2)	0.50	0.26	0.56	0.20
LBP	0.50	0.19	0.67	0.27
COPD	0.19	0.42	0.25	0.07
IBS	0.46	0.22	0.63	0.21
Lu	0.54	0.30	0.55	0.18
OA	0.48	0.29	0.43	0.14
75+	0.46	0.31	0.51	0.17
Mp	0.45	0.28	0.56	0.20

The regression 'whole (1)' is: SF-6Dindex $= \alpha + \beta_1$ EQindex $+ u$. All others are: SF-6Dindex $= \alpha + \beta_1$ EQindex $+ \beta_2$N3 $+ u$. All coefficients have $p < 0.001$ except β_2 for COPD where $p = 0.03$.

coefficients indicate the relationship is not uniform between the two index numbers. The patient specific β_1 coefficients were found to be similar for IBS, leg ulcer, the elderly and menopausal groups, but statistically different for Lower Back pain, COPD and Osteoarthritis. Another set of regressions were run where two sets of estimations were obtained for those with N3 = 0 and those with N3 = 1, but the differences in coefficients were small, and therefore not reported here. As can be expected from the scatter plot, the inclusion of the N3 term will improve the explanatory power somewhat, and makes the β_1 coefficient much larger.

Discussion

It is important to compare the different preference-based measures of health in order to inform an important debate on the choice of instrument. This paper has sought to address this question by comparing two of the more widely used generic measures in seven patient groups. The results presented in this paper broadly support the following conclusions. There is considerable overlap in the descriptive systems of the two instruments and significant agreement between the indices that they generate. A simple comparison of mean SF-6D and EQ-5D indices across the whole sample of patients found that on average SF-6D generates values that exceed the EQ-5D. The mean difference, though statistically significant, is only 0.045. This high degree of convergence provides some support for their construct validity [2]. However, this apparent similarity hides a significant degree of disagreement between these two measures. The intraclass correlation coefficient for the whole sample is 0.51. Furthermore, depending on the patient group, the difference between the SF-6D and EQ-5D ranges from −0.015 to 0.094 and the intraclass correlation coefficient ranges from 0.27 to 0.55. The median shows a different pattern, with the EQ-5D value exceeding the SF-6D value. This reflects the considerable negative skew in the EQ-5D values. The plot of the two indices for the whole group reveals marked differences over the range of ill health.

The similarity in mean health state values across the patient groups implies that for curative interventions aimed restoring patients to full health the size of the difference in estimated QALY gain may not be great for many of these conditions, with a difference of less than 0.05 for five out of the seven conditions. However, most interventions do not restore people to full health but achieve some partial relief, where the variation in the relationship over the range of ill health could have significant implications for the estimate of QALY gain. It is therefore important to understand the reason for these differences in order to inform the debate concerning which instrument to use and whether or not there is a case for revising one or both of these measures to solve any apparent limitations. This involved a discussion of their descriptive content, the valuation methods and the resultant scoring algorithm.

Descriptive content

The descriptive systems would seem to account for some part of the difference. The SF-6D is more concentrated at the milder end of health problems, whereas the EQ-5D covers a larger range and consequently suffers from being cruder. Around a third of respondents to the valuation survey thought that the worse state defined by the SF-6D was worse than death compared to a great majority for the EQ-5D. Correspondingly, at the upper end the EQ-5D has a far larger proportion on the top level than the SF-6D and that those on the ceiling of the EQ-5D can be differentiated by the SF-6D in terms of health problems. A similar result was found in a study of liver transplant patients (see Louise L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ*, in press). The two health classifications cover similar domains (except for energy), but at least for physical functioning, self care and usual activities describe different severity levels. For these dimensions the different levels are described as *limits* to activities. In the EQ-5D, the lowest levels of the comparable domains are described as *unable* and *confined to bed*. The SF-6D has a far larger number on the floor of these dimensions, but this does not extend to the dimensions for pain and mental health.

This difference between the dimensions is supported by the larger mean differences for conditions focused more on physical health problems such as leg ulcers and osteoarthritis compared to those more focused on pain and other forms of discomfort such as lower back pain and IBS. The implication would be that the researcher should choose between the instruments on the bases of the appropriateness of the descriptive system in terms of the severity of problems typically encountered in the patient group for each domain.

Valuation methods

Unfortunately the conclusion is not quite as simple, since we do not fully understand the consequences of the differences in valuation methods. The main difference in methods is in the valuation technique. It has been suggested that for a range of reasons SG would be expected to generate higher values than TTO across the entire severity range and this has been found in a number of studies [21]. A study undertaken in York comparing their own variants of TTO and SG, however, found evidence for a crossover. SG props values exceeded TTO props up to VAS values of 0.4, but then there was a cross over with TTO values exceeding SG values [22]. The size and pattern of the difference between TTO and SG depends, therefore, on the variants of the two techniques being used. We have undertaken such a comparison and provisional results have been reported elsewhere [23].

Another difference is that the SF-6D group uses a 'two stage', or 'chained' SG. That is, first the health states are valued using perfect health and the worst health state as anchor points. So in the first instance, death is no part of the SG. Then the worst state is valued using a standard gamble with the anchor points perfect health and death. The primary reason to use a chained SG is the notion that subjects might be reluctant to trade off health against immediate death. There is evidence that chained values may be higher than value obtained [24–26]. This would mean that the values of the SF-6D would show upward shift compared to the MVH valuation of the EQ-5D which did not use a chained procedure.

The scoring algorithms

The algorithms presented in Table 2 are the result of the descriptive system and the valuation method and therefore in one sense examining the algorithms does not add any additional explanations. However, there are two features of the algorithm that have an important effect on the indices they generate: the interpretation of the constant term and the interaction terms.

The expected value of the estimated constant term in the models was one, but in practice it was found to be significantly smaller than one. The SF-6D study made a theoretical case for restricting the intercept to unity [11]. The SG value for each state has been estimated by assuming SF-6D state 111111 health is to equal one. Whereas the EQ-5D study interprets the difference between the estimated constant and one as 'any move away from full health' (p. 1104); however, this has no theoretical justification. This difference in use of the treatment of the constant term would increase the value of all ill health states defined by the SF-6D compared to the EQ-5D. It results in the large gap between EQ-5D state

11111 with a value of 1.0 and the next state a value of 0.88.

Both studies accounted for interaction effects in a similar way. For the SF-6D the interaction term is a simple dummy, MOST, which takes the value 1 if any dimension in the health state is at the 'most severe' level, and 0 otherwise. 'Most severe' is defined as level 6 for physical functioning and pain, levels 4 and 5 for social functioning and mental health and level 5 for vitality. For the EQ-5D, N3 is a very similar dummy variable that takes the value 1 when any dimension is at level 3, and 0 if otherwise. The N3 term has a coefficient of −0.269 compared to that for MOST of −0.032. The N3 together with the value associated with a move from level 2 to 3 of 0.1–0.3 results in another large gap in the distribution of the EQ-5D index.

Conclusion

It is important not to lose sight of the key finding that at the mean level these instruments produced indices that were within 0.05 of each other. However, this does not imply that these two instruments can be used interchangeably since they generate different indices over the range of ill health. Such differences have been found between other generic preference-based measures [27] (See Louise L, Bryan S. An empirical comparision of EQ-5D and SF-6D in liver transplant patients. *Health Econ*, in press; O'Brien BJ, Spath M, Blackhouse G, Severens JL, Brazier JE. A view from the Bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index. *Health Econ*, accepted). These results raise the important research question of why different generic preference-based measures are giving different values. To address this question it would be necessary to extend this comparison to other data sets across the full range of ill health. It would also be important to compare the variants of the valuation methods used for each instrument [23].

The study has other important implications for further work. One is to extend this comparative work to other preference-based measures of health, and this paper demonstrates the different ways in which it is possible to compare measures. A second is to estimate the impact of the differences between these and other preference-based measures on the estimates of the gains from health care interventions. Another implication would be to consider revising one or both of these instruments to overcome their weaknesses, particularly in their descriptive systems. It could be possible, for example, to add more intermediate levels to the EQ-5D or, by the same token, add lower levels to the SF-6D dimensions, at least for the physical functioning and role limitation dimensions.

Acknowledgements

We are grateful to Stephen Walters and the two anonymous referees for their advice.

Notes

a. Those wishing to use the EQ-5D should contact the EuroQol Business Management, PO Box 4443, 3006 AK Rotterdam, The Netherlands; http://www.euroqol.org/

b. Those wishing to use the SF-6D should contact SF-6D, Sheffield Health Economics Group, School of Health and Related Research, The University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK.

References

1. Drummond MF, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programme*. Oxford Medical Publications: Oxford, 1987.
2. Brazier JE, Deverill M, Harper R, *et al*. A review of the use of Health Status measures in economic evaluation. *Int J Health Technol Assess* 1999; **3**: 1–164.
3. The Euroqol Group. Euroqol – a new facility for the measurement of health-related quality-of-life. *Health Policy* 1990; **16**: 199–208.
4. Brooks R, The EuroQol Group. EuroQol: the current state of play. *Health Policy* 1996; **37**: 53–72.
5. Dolan P. Modelling valuation for EuroQol health states. *Med Care* 1997; **35**: 351–363.
6. The MVH Group. *The Measurement and Valuation of Health: Final Report on the Modelling of Valuation Tariffs*. Centre for Health Economics, University of York, 1995.
7. Badia X, Roset M, Herdman M, *et al*. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Med Decis Making* 2001; **21**: 7–16.

8. Claes C, Greiner W, Uber A, *et al*. An interview-based comparison of the TTO and VAS values given to EuroQol states of health by the general German population. *Proceedings of the EuroQol Meeting 1999*, Hannover, 1999.
9. Tsuchiya A, Ikeda S, Ikegami N, *et al*. Estimating an EQ-5D population value set: The case of Japan. *Health Econ* 2002; **11**: 341–353.
10. Ware JE, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Med Care* 1992; **30**: 473–483.
11. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002; **21**: 271–292.
12. Brazier JE, Deverill M. A checklist for judging preference-based measures of health related quality of life: Learning from psychometrics. *Health Econ* 1999; **8**: 41–52.
13. Brazier J, Harper R, Jones NMB, *et al*. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *Br Med J* 1992; **305**: 160–164.
14. Thomas K, Thorpe L, Fitter M, *et al*. *Long-term clinical and economic benefits of offering acupuncture to patients with chronic low back pain assessed as suitable for primary care management—3 month clinical outcomes*. Medical Care Research Unit, University of Sheffield, 2000.
15. Harper R, Brazier JE, Waterhouse JC, *et al*. Comparison of outcomes measures of patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. *Thorax* 1997; **52**: 879–887.
16. Akehurst R, Brazier J, Mathers N. Health-related quality of life and cost impact of irritable bowel syndrome in a UK primary care setting. *Pharmacoecon* 2002; **20**: 455–462.
17. Walters SJ, Morrell CJ, Dixon S. Measuring health-related quality of life in patients with venous leg ulcers. *Qual Life Res* 1999; **8**: 327–336.
18. Zoellner Y, Oliver P, Platts M, Alt J, *et al*. Development and validation of a menopause-specific quality of life questionnaire. Poster presented at the *16th European Congress of Obstetrics and Gynaecology*, Malmö, 2001.
19. Brazier JE, Harper R, Munro J, *et al*. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. *Rheumatology* 1999; **38**: 870–877.
20. Brazier JE, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and Euroqol on an elderly population. *Qual Life Res* 1996; **5**: 195–204.
21. Green C, Brazier J, Deverill M. Review of health state valuation techniques. *Pharmacoecon* 2000; **17**: 151.
22. Dolan P, Sutton M. Mapping VAS scores onto TTO and SG utilities. *Soc Sci Med* 1997; **44**: 1289–1297.
23. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets in the UK. The Euroqol Meeting, York, September 2002.
24. Rutten-van Molken M, Bakker C, Doorslaer E, *et al*. Methodological issues of patient utility measurement. Experience from two clinical trials. *Med Care* 1995; **33**: 922–937.
25. Lalonde L, Clarke A, Joseph L, *et al*. Conventional and chained standard gambles in the assessment of coronary heart disease prevention and treatment. *Med Decis Making* 1999; **19**: 149–156.
26. Bleichrodt H. Probability weighting in choice under risk: An empirical test. *J Risk Uncertainty* 2001; **23**: 185–198.
27. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Ann Med* 2001; **33**: 358–370.
28. Bera A, Jarque C. Efficient tests for normality, heteroscedasticity and serial independence. *Econ Lett* 1980; **6**: 255–159.
29. Ljung G, Box G. On a measure of lack of fit in time series models. *Biometrika* 1979; **66**: 265–270.

