Title: **… another study showing that the EQ-5D and SF-6D are not interchangeable. But why should we expect them to be?**

Word count: 6619

Table count: 5

Figure count: 2

Authors: Mr David G T Whitehurst (1,2): d.whitehurst@cphc.keele.ac.uk
Prof. Stirling Bryan (3)

Affiliations:
(1) Arthritis Research Campaign National Primary Care Centre, Primary Care Sciences, Keele University, Staffordshire, UK
(2) Health Economics Unit, University of Birmingham, Birmingham, UK
(3) Centre for Clinical Epidemiology and Evaluation, 7th Floor, 828 West 10th Avenue, Research Pavilion, Vancouver, Canada

Abstract

Aims
Previous work has demonstrated poor agreement between EQ-5D and SF-6D index scores for individual responses and group mean estimates. Our objective was to further compare these two measures, using contemporaneous index scores collected from patients with nonspecific neck pain, to provide insight into the nature of between-measure discrepancies.

Methods
Initially, the expected 'poor agreement' was explored using intraclass correlation coefficients and Bland & Altman plots. Subsequently, techniques were used to investigate the reasons for poor agreement, focusing on practical considerations and the respective descriptive and valuation components of the two measures; response rates, item-completion rates, question formats, dimension-to-dimension correlations, floor and ceiling effects, and construct validity.

Results
A poor level of agreement was confirmed. The two measures do not permit respondents to value their health state in the same manner, due, primarily, to differences in the contextual framing of items and the number of available response options. The EQ-5D and SF-6D were indistinguishable with regard to the statistical significance of linear trends across theoretical constructs. Response rates were consistently better for the EQ-5D; poorer completion on the SF-6D was related to a single item.

Conclusions
The non-interchangeable nature of EQ-5D and SF-6D index scores is not surprising given the differences in the descriptive content of the measures. Selecting the 'better' measure is problematic, although, in their current format, the wider scoring range and better completion rates associated with the EQ-5D are sufficient for it to remain the standard measure for use in economic evaluation.

1. Introduction

Within a cost-utility framework, outcome measures must provide a single index value that reflects respondents' preferences for different health states. Preferences can be measured directly using preference-elicitation techniques or indirectly captured through the use of health state classification questionnaires (also known as multi-attribute utility scales (MAUS), or preference-based health-related quality-of-life instruments). In the UK, two indirect measures are often used; the EQ-5D and the SF-6D. The use of indirect approaches is increasingly popular due to the availability of UK population-based health state valuations for every response permutation and the ease of administration (with associated time and cost savings). However, the availability of several instruments raises concerns about the cross-study comparability of cost-utility results when studies have used different outcome measures.

Across a range of clinical conditions and community-based samples, previous research has shown that the EQ-5D and SF-6D do not provide individual-level index scores or group mean index scores that are interchangeable (Brazier et al, 2004; Whitehurst et al, 2009). Specific to the field of musculoskeletal disorders, studies have shown considerable variation in the difference between EQ-5D and SF-6D index scores, ranging from 0.022 for chronic low back pain (Brazier et al, 2004) to 0.180 for patients with a confirmed diagnosis of a lumbar spine disorder (McDonough et al, 2005). Irrespective of the study population and magnitude of the difference in mean scores, studies have consistently identified important differences between the two measures. Knowing that two outcome measures, which purport to measure the same underlying construct (i.e. preference-based health-related quality of life), provide different estimates is an important finding. However, within clearly defined patient populations, there is a need to inform an important debate about the appropriate choice of instrument and provide guidance for the design of future studies, as well as understanding the reasons for between-measure discrepancies.

There are a number of reasons why two utility measures may provide different valuations for a given health state. Preference-based measures are made up of two constituent parts; a descriptive system that defines respondents' health-related quality of life as one of a finite number of health states and a valuation system that scores each health state as a single index score, usually based on community-derived preferences. Either element, or a combination of the two, could offer an explanation as to why measures differ in their contemporaneous valuation of a health state (Bryan and Longworth, 2005). In addition to the descriptive and valuation components of the EQ-5D and SF-6D, other considerations play a key role in the decision to select between alternative health measurement scales, such as issues of practicality, patient burden and administrative burden (Lohr et al, 1996; Brazier and Deverill, 1999).

This paper provides a direct head-to-head comparison of the EQ-5D and SF-6D for a sample of patients with nonspecific neck pain, with the primary objective being to provide insight into the nature of between-measure

discrepancies. Throughout the remainder of this paper, 'SF-6D' will refer to the preference-based measure derived from the SF-12. There are two stages to the analysis. Initially, it was necessary to assess the level of agreement between the EQ-5D and SF-6D. It was hypothesised that such analysis would demonstrate poor agreement, in line with the evidence from other clinical areas. Following this, techniques that cover three broad themes were used to explore the reasons for the lack of congruence between the two measures. These themes were the assessment of practicality in terms of response rates and item-completion rates, examination of question formats and response data for individual dimensions to compare the respective *descriptive* components, and the theoretical validation of index scores to compare the respective *valuation* components.

2. Methods

2.1 Data source: overview of the PANTHER study

The PANTHER (**P**hysiotherapy **A**rc **N**eck **T**rial, **H**ands on, **E**lectrotherapy **R**esearch) study was a randomised controlled trial that sought to assess three alternative treatment packages for patients aged 18+ with a clinical diagnosis of nonspecific neck pain (Dziedzic et al, 2005). The primary outcome measure was neck pain disability measured using the Northwick Park Neck Pain Questionnaire (NPQ); a 0-100 scale, where 0 = no neck pain or disability, 100 = maximum neck pain and disability (Leak et al, 1994). Of 735 potential participants screened between July 2000 and June 2002, 346 were eligible, consented to be randomised and were followed-up. Outcome data, including the EQ-5D and SF-12, were collected at baseline (in an assessment clinic), 6 weeks and 6 months (by self-report postal questionnaires). The EQ-5D and the 7 SF-12 items that comprise the SF-6D were replicated in line with the official UK validated versions (Brooks, 1996; Brazier and Roberts, 2004); the EQ-5D appeared immediately before the SF-12 in all questionnaires. Both measures were located in a 'general health' section of the questionnaire in order to make a clear distinction from prior sections that focused on neck pain-specific outcome measures.

2.2 Preference-based measures: the EQ-5D and SF-6D

Index scores from utility instruments are interpreted on a 0-1 scale, where 0 indicates a health state valuation equivalent to death and 1 is full health. The EQ-5D and SF-6D have been described in detail elsewhere (Whitehurst *et al*, 2009). Briefly, The EQ-5D is a 5-dimension self-classification system covering mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension containing 3 levels, which provides 243 (3^5) distinct health states that can be defined by a unique five-digit number. Health states for 'unconscious' and 'dead' have been added, resulting in a total of 245 health states. The UK scoring algorithm was derived from a valuation study performed by the Measurement and Valuation of Health (MVH) group at the University of York (Dolan et al, 1995). This was the first large-scale EQ-5D valuation study and is still the most widely-used scoring algorithm (Räsänen, 2006). Preferences were elicited from a representative sample

of 3395 members of the UK adult population using time trade-off methodology. The resultant algorithm provides utility scores within a range of -0.594 (state 33333, the lowest level on each dimension) to 1.000 (state 11111, the highest level on each dimension). Negative values reflect that some health states are considered to be worse than death.

The SF-6D can be derived from two widely-used generic health profile measures, the Short Form 36 (SF-36) (Ware et al, 1992) and the Short Form 12 (SF-12) (Ware et al, 1996). The SF-6D comprises 6 dimensions (physical functioning, role limitations (physical & emotional), bodily pain, vitality, social functioning, and mental health), which is constructed from 7 items of the 12-item instrument, each with between 2 and 5 levels of severity. This descriptive systems defines 7,500 health states. Preferences for the scoring function were measured using standard gamble methodology, based on a representative sample of members of the non-institutionalised UK adult population (n=611). The range of index scores covered by the algorithm is 0.345 to 1.000. Although the minimum value does not include zero, index scores are still interpreted on a 0-1 scale.

2.3 Methods of Analysis

Assessment of the level of agreement between the EQ-5D and SF-6D index scores was based on graphical and statistical approaches, using Bland and Altman plots (Bland and Altman, 1986), with the associated limits of agreement, and the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979). For the quantification of absolute agreement, a single measure ICC based on a two-way mixed analysis of variance model was calculated, where the two outcome measures are treated as a source of variability. For this study, the following benchmarks were used to interpret correlation coefficients; 0.00 to 0.10 representing virtually no correlation/agreement, 0.11 to 0.40 slight, 0.41 to 0.60 fair, 0.61 to 0.80 moderate and 0.81 to 1.00 substantial correlation/agreement (Shrout, 1998).

For the assessment of practicality, a response was defined as positive if there were sufficient data to allow for the derivation of a utility score, whereas 'item-completion' focused on the individual dimensions of the two measures. The SF-12 has clearly defined procedures for handling missing responses in order to allow calculation of the mental and physical component scores. However, these procedures do not impute missing values for individual questions and, therefore, do not provide any assistance in dealing with missing responses when using SF-12 data for the SF-6D. Similarly, the EQ-5D requires complete responses to all 5 dimensions to generate an index score. Differences between responders and non-responders were explored to identify the underlying reasons for missing utility scores.

Three approaches were used to explore potential reasons for disagreement; a descriptive reflection on the contextual setting of questions within the two measures (terminology, time frame etc.) and the assessment of dimension-to-dimension correlations and floor and ceiling effects.

The EQ-5D and SF-6D contain dimensions that would, on first thought, be expected to have a strong relationship with each other, such as EQ-5D 'pain/discomfort' and SF-6D 'pain', and EQ-5D 'anxiety/depression' and SF-6D 'mental health.' However, failure to consider the context within which respondents are asked about health dimensions may lead to important oversights when addressing the comparability of the descriptive components of two measures. The questionnaire formats of the two instruments were examined in order to identify any contextual differences.

To go beyond this discursive examination, dimension-to-dimension correlations were explored in order to quantify the extent of the relationship between dimensions across the two instruments. The purpose of this analysis was to identify whether the correlation structure between the domains of the two instruments identifies the highest correlations between those dimensions that purport to capture the same specific aspects of health-related quality of life. Several *a priori* assertions were made regarding paired dimensions that were expected to be the highest correlations. For each of the seven items on the SF-6D (the role limitations dimension requires the combination of two items; physical aspects and emotional aspects of role limitations), an EQ-5D dimension was chosen that was thought to best reflect the same health domain. An additional assertion was made in relation to the physical functioning dimension of the SF-6D as the correlation was expected to be high with two different EQ-5D dimensions (the 8 assertions are identified in Table 2).

For the assessment of floor and ceiling effects, responses that were classified as 'full health' or 'worst possible health' on one measure were examined more closely in terms of their distribution on the other measure. In addition to looking at index scores, responses to both measures were broken down into the individual dimensions to further explore sources of disagreement.

The comparison of EQ-5D and SF-6D index scores focused on descriptive statistics (mean, standard deviation, median, inter-quartile range, minimum, maximum, 95% confidence interval and 99% confidence interval) and empirical validity testing. Paired observations at each time point were used to compute the same descriptive statistics for the difference between index scores. Within-subject differences in mean utility scores were tested using paired *t*-tests; the Wilcoxon Signed Rank test was used to test for the equality of median values. Box-plots were used to graphically describe the distribution of data points. Within this study, the 'whiskers' of the box-plots extend the smallest and largest values deemed to be a reasonable distance from the box that represents the inter-quartile range. This 'reasonable' distance is defined in two-stages, which are reflective of a Normal distribution assumption. Firstly, values which are between 1.5 and 3 box lengths from either end of

the box are denoted by a circle. Beyond this range, where values are more than 3 box lengths from either end of the box, data points are denoted with an asterisk.

Validity testing has been described as the 'acid test' for preference-based measures as it provides a rationale for assessing whether such measures *do* reflect preferences as opposed to whether they *could* reflect preferences (Brazier and Deverill, 1999). A number of checklists for evaluating health status questionnaires have been published, addressing many different aspects of validity (Lohr et al, 1996; Terwee et al, 2007), including a checklist specifically for preference-based measures (Brazier and Deverill, 1999). However, the application of some of the validity techniques listed in such checklists was not considered relevant for a *comparative* analysis of EQ-5D and SF-6D responses. Face validity and content validity were not judged to be important when comparing two previously validated measures. Similarly, convergent validity and criterion validity were deemed to be inappropriate for this analysis. Convergent validity is concerned with the extent to which a measure correlates with another measure of the same concept. While correlation was considered to be an important analytical technique when exploring the relationship between similar dimensions across the EQ-5D and SF-6D, it is not appropriate for the comparison of index scores. Criterion validity relates to the extent to which the scores on an instrument correlate to a gold standard or *criterion* measure. The gold standard for preference-based measures of health-related quality of life is revealed preference data (Brazier and Deverill, 1999). However, there are well-documented problems regarding the application of revealed preference methods within a health care context (Donaldson et al, 2004).

The theoretical validity concept explored within this chapter focused on construct validity, which is a technique that assesses the ability of an instrument(s) to discriminate between study populations thought to differ on the basis of a theoretically constructed hypothesis. Hypothetically constructed preference rules were devised for a purposive selection of generic and neck pain-specific outcome measures in order to explore issues deemed to be important to utility measures and reflective of the opportunities specific to the PANTHER data set. Analysis of these preference rules consisted of computation of descriptive statistics and appropriate statistical tests to explore the significance of differences in utility scores across patient subgroups.

Although utility measures are generic health instruments, some theoretical constructs were based on neck pain-specific outcome measures as the focus of this chapter was to assess the *comparative* performance of the EQ-5D and SF-6D, rather than the evaluation of preference-based measures *per se*. As an extension to this focus on pain, additional constructs explored responses to the EQ-5D and SF-6D pain-related dimensions. Four pain-related constructs were explored; neck pain severity (3-level ordinal variable: mild, moderate, severe), neck pain-specific assessment of change compared with baseline (5-level ordinal variable: much better, better, same, worse, much worse) and three groups were defined using tertile values for responses to the Northwick Park neck pain questionnaire and average pain severity over the last three days (0-10 numerical rating scale),

The constructs for generic outcome measures consisted of self-reported health status categories (5-level ordinal variable: excellent, very good, good, fair, poor), tertile-defined groups for responses to the EuroQol visual analogue scale (EQ-VAS) (0-100 scale, where 100 = 'best imaginable health state' and 0 = 'worst imaginable health state') and time-off work for respondents currently in paid employment (inclusive of part-time and full-time employment).

The appropriate statistics to empirically test the theoretical constructs depended on the type of variables involved and the number of levels defined within the respective categorical variable. When the construct consisted of examining an interval variable against a 2-level nominal variable, *t*-tests were used. Analysis of variance (ANOVA) was used for those preference rules consisting of interval variables and ordinal variables (each ordinal variable had at least 3 levels), testing for linear trends in mean scores across groups. For example, mean utility values were expected to increase monotonically with improving self-reported health status. The exploration of some preference rules relating to the pain-related dimensions required testing two ordinal variables. In these instances, the Mantal-Haenszel chi-square test (sometimes called the 'chi-square test-for-trend' or the 'linear-by-linear association chi-square') was used to test whether one variable significantly increase/decreases with an increase in the other variable. As a general rule, approximately 80% of cells in a cross-tabulation of two categorical variables should have expected values of at least 5 (Bland, 1987). In the event of violating this requirement, the merging of categories with low marginal counts was considered as a remedial measure.

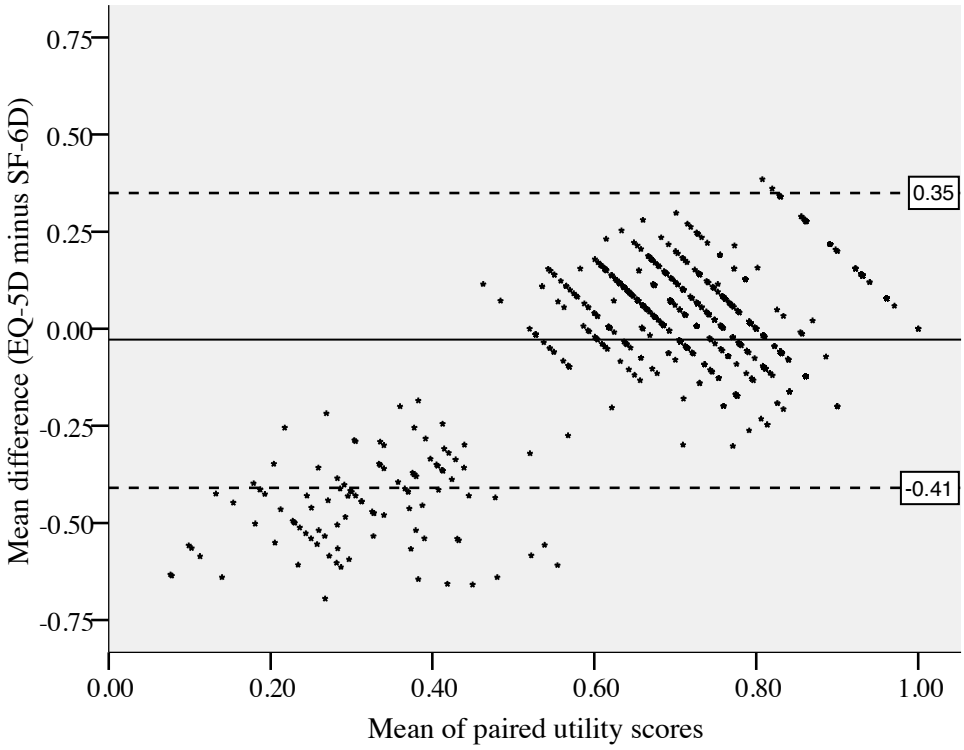
Statistical considerations

Repeated EQ-5D and SF-6D observations were available within the PANTTHER data set. In some scenarios, including previous empirical comparisons of EQ-5D and SF-6D scores (Brazier et al, 2004), such observations have been combined to form a single study population. Typically, this is an inappropriate action within health care research because within-subject observations are not independent (Bland and Altman, 1994). With regard to statistical rigour, the importance of ignoring the non-independence of observations within a combined sample will depend on the analysis being performed. In this study, the starting point was to account for the non-independence of observations by performing analyses separately for different time points. To prevent unnecessary repetition of work, analyses of a combined sample population is reported when findings and interpretation were analogous to the respective time point-specific analysis.

3. Results

The intraclass correlation coefficient for the analysis of absolute agreement ranged from 0.518 (at 6 week follow-up) to 0.550 (at 6 month follow-up), which is interpreted as a 'fair' level of agreement according to the chosen benchmark. The Bland and Altman plot for the combined data set is presented in Figure 1. The width of the limits of agreement, or the 'range of expected variation', ranged from 0.713 (at 6 weeks) to 0.835 (at 6 months). These values provide an indication of the expected level of variation between any pair of future EQ-5D and SF-6D observations, specific to this clinical area, which demonstrates a substantial lack of agreement between the two measures. Consistent results were observed within all time point-specific analyses (not shown). Within a conventional Bland and Altman plot, the purpose of the dashed lines representing the limits of agreement is to reflect a consistent assessment of agreement across the whole range of the measurement scale, i.e. the lines are horizontal, based on the assumption that there is no relationship between the magnitude of the scores (represented by the average of the EQ-5D and SF-6D score) and the difference between scores. However, Figure 1 shows that the degree of disagreement is not constant. At the lower end of the scale, there were 68 (8%) paired observations below the lower limit of 95% agreement. Conversely, at the upper end of the scale, there were only two (<1%) observations above the upper limit of 95% agreement. This is a reflection that agreement is poor across the entire range but is *poorer* at the lower end the scale.

Figure 1: Bland and Altman plot for paired EQ-5D and SF-6D scores for the combined sample (n=853)



A breakdown of EQ-5D and SF-6D response rates at each data collection stage is provided in Table 1. At each time point, the response rate was greater for the EQ-5D, ranging from a 2% difference at baseline, to a 7% difference at 6-month follow-up. The largest between-measure difference was in relation to the consecutive reporting of utility measures, across the three data collection phases, where a significantly greater proportion of the sample provided complete EQ-5D responses compared to complete SF-6D responses (difference = 10%; $\chi^2(1, N = 346) = 177.96, p < 0.001$).

Table 1: Response rates (percentages) by data collection point

Data collection point	EQ-5D	SF-6D	Both
Baseline	345 (100)	338 (98)	337 (97)
6 weeks	291 (84)	274 (79)	270 (78)
6 months	271 (78)	247 (71)	246 (71)
All time points	255 (74)	222 (64)	216 (62)
Total responses for the combined sample	907 (87)	859 (83)	853 (82)

With regard to individual dimensions, for the EQ-5D, the consistency of item-completion rates indicated that the inability to generate an index value at either follow-up stage was not related to the completion of a single dimension (not shown). For the SF-6D, the 'role limitations' dimension had a poorer item completion rate than the other SF-6D dimensions at both 6-week and 6-month follow-up. Disregarding the role limitations dimension, similar item-completion rates were identified for all remaining dimensions, across both instruments (item completion rates were ≈100% at baseline, ≈85% at 6 weeks and ≈79% at 6 months).

There were 54 instances where the EQ-5D was the only index score that could be generated from a respondents questionnaire; of these 54 observations, 44 (81%) were due to non-response to the 'physical' item of the role limitations dimension. Further analysis was conducted to compare respondents with complete EQ-5D and SF-6D data to the sample of respondents with missing data on the 'physical' role limitations item (but complete data for every other EQ-5D and SF-6D dimension). It was found that respondents with missing data on the 'physical' role limitations were more likely to be older and not in paid employment at the time of recruitment ($p < 0.05$). There were no statistically significant differences between groups with regard to gender, proportion of employed participants reporting a period of work absence or clinical outcome measures.

Examination of the wording used in the EQ-5D and SF-6D identified three key differences between the measures that questions whether respondents are being asked to describe their health state in the same way.

Firstly, completion of the EQ-5D requires responses to statements that focus exclusively on the level of severity within the respective health dimension, e.g. the response options for the pain dimension relate to the *level* of pain felt by the respondent (no pain or discomfort, moderate pain or discomfort, extreme pain or discomfort). The description of some health dimensions within the SF-6D involves a different approach, where respondents are asked to consider the extent to which specific health dimensions impact on their lives. For example, the pain dimension asks, '*how much did pain interfere with your normal work (including work both outside the home and housework?*' These two approaches differ greatly with regard to their measurement objective and, therefore, the EQ-5D and SF-6D provide respondents with two contextually different sets of dimensions from which to describe their health state. For example, of the 624 level-2 responses on the EQ-5D pain/discomfort dimension from the 853 combined sample observations providing matched EQ-5D and SF-6D scores, 70 (11%) observations had a matched response at level 1 on the SF-6D pain dimension, whereas 118 observations had a matched response at level 4 or level 5.

The second contextual difference relates to the emphasis on 'time' within the SF-6D, which is not seen within the EQ-5D. Duration of impairment within the last week (the approach of the SF-6D) and severity of impairment today (the approach of the EQ-5D) provide different descriptive backdrops for respondents to describe their health state. For example, it is feasible for individuals to have severe mental health impairment intermittently or, alternatively, mild mental health impairment constantly. Therefore, poor agreement or correlation between paired responses to the SF-6D mental health dimension and the EQ-5D anxiety/depression dimension does not infer inconsistent or contradictory responses.

The third difference is in regard to the length of time respondents are asked to consider when completing the questionnaires. The EQ-5D asks respondents to indicate the response option that best describes their health state on the day of completion (i.e. 'today'), whereas the acute version of the SF-12 asks respondents to consider their experiences '*during the last week*.' One previous study has suggested that an instrument with a 7-day recall period will provide lower utility values for recently resolved adverse events when compared to an instrument with a 1-day recall period (Bansback et al, 2008), although limitations within this study mean that further research is necessary to validate the results. Contrary to these findings, scoring algorithms for the SF-6D make no allowance for the different recall periods of the standard (4-week) and acute (1-week) versions of the SF-12 and SF-36, which imparts an implicit assumption that different lengths of recall do not result in systematically different descriptions and/or valuations of health status.

Table 2 presents the matrix of Spearman correlation coefficients for the combined sample, which summarises the relationship between dimensions across the two instruments (correlations for the 8 *a priori* assertions are underlined, whereas the 5 highest correlations are in bold). Of the 8 paired dimensions that were expected to be among the highest correlations, 4 had a correlation coefficient greater than 0.500 and were among the highest 5 coefficient values. The results suggest there is evidence for the convergent validity between similar

dimensions: between physical functioning and usual activities, between role limitations and anxiety/depression, between pain and pain/discomfort, and between mental health and anxiety/depression. The unexpected correlation within the highest 5 values was between pain (SF-6D) and usual activities (EQ-5D), which had the second highest correlation coefficient of 0.612. The four lowest correlations all involved the mental health dimension of the SF-6D and the remaining EQ-5D dimensions. Of the *a priori* assertions, only the correlation between vitality (SF-6D) and usual activities (EQ-5D) had a value below 0.400.

Table 2: Correlation between EQ-5D and SF-6D dimensions for the combined sample[a]

SF-6D \ EQ-5D	Mobility	Self-care	Usual activities	Pain / discomfort	Anxiety / depression
Physical functioning	0.488	0.391	**0.619**	0.483	0.233
Role limitations	0.322	0.256	0.442	0.353	**0.511**
Pain	0.467	0.362	**0.612**	**0.590**	0.318
Vitality	0.340	0.267	0.323	0.279	0.300
Mental Health	0.194	0.171	0.221	0.231	**0.596**
Social functioning	0.412	0.330	0.443	0.416	0.432

[a] The five most correlated dimensions are in bold. The underlined correlations were identified *a priori* as purporting to capture similar aspects of quality of life.

Within the combined sample, there were 107 (13%) paired observations with an EQ-5D classification at full health. For this sample of responses, the mean (sd) SF-6D score was 0.852 (0.09). In 3 (3%) instances, the corresponding SF-6D score indicated full health; the lowest SF-6D score with a paired full health valuation on the EQ-5D was 0.615. Full health was reported on the SF-6D on 7 (<1%) occasions; matched EQ-5D responses were full health (3 times) or a value of 0.800 (4 times). At the bottom end of the respective scales, no respondents report the lowest possible health state on the EQ-5D, whereas 2 SF-6D responses had the lowest possible health state valuation of 0.345. The corresponding EQ-5D index scores for these two 'worst possible' SF-6D states were both below the SF-6D valuation (-0.080 and 0.090).

Table 3 reports the dimension-level responses on the SF-6D for observations where the corresponding EQ-5D valuation is full health. All dimensions on the SF-6D have at least 10% of responses indicating some level of impairment. More significantly, the role limitations, pain, vitality and mental health dimensions of the SF-6D identified impairment in at least 25% of responses. This indicates that the EQ-5D has problems in reflecting small to moderate levels of impairment for these health dimensions in respondents that are in relatively good/mild health states. For example, only 4% of respondents reporting full health on the EQ-5D indicate having 'a lot of energy' (the vitality dimension) 'all of the time' (response category 1) on the SF-6D.

Table 3: Distribution of data points for each SF-6D dimension for the sample of values where the paired EQ-5D valuation is full health (n=107)

SF-6D dimension	Number (%) of responses for each level of severity				
	level 1	level 2	level 3	level 4	level 5
Physical functioning	92 (86)	15 (14)	0 (0)	-	-
Role limitations	71 (66)	13 (12)	14 (13)	9 (8)	-
Pain	81 (76)	22 (21)	3 (3)	1 (1)	0 (0)
Vitality	4 (4)	53 (50)	38 (36)	12 (11)	0 (0)
Mental Health	75 (70)	24 (22)	6 (6)	2 (2)	0 (0)
Social functioning	96 (90)	6 (6)	5 (5)	0 (0)	0 (0)

The distribution of all EQ-5D and SF-6D responses for each response category was also explored (not shown). EQ-5D item-responses are in the top category of the respective dimensions in 17% (pain/discomfort) to 89% (self-care) of the time, compared to a range of 3% (vitality) to 53% (social functioning) for the SF-6D. The opposite effect is shown for the lowest response categories of each dimension. For the EQ-5D, the most severe response category for the mobility dimension (level 3, '*I am confined to bed*') was reported once, while there were no level 3 responses for the self-care dimension ('*I am unable to wash or dress myself*'). This contrasts to the physical functioning and role limitations dimensions of the SF-6D, where the lowest categories were reported in 17% and 41% of responses, respectively. The range of responses for the vitality dimension on the SF-6D, both in the combined sample (not shown) and for those respondents reporting full health on the EQ-5D (Table 3), indicates that it is a health domain that is unique to the SF-6D. Further support for this supposition is seen in the low correlation coefficients between vitality and all 5 EQ-5D dimensions; the highest correlation was with the mobility dimension, with a coefficient of 0.340 (see Table 2).

A box-plot of EQ-5D and SF-6D index scores for the combined sample is shown in Figure 2; descriptive statistics are presented in Table 4. The box-plot identifies EQ-5D outliers that correspond to health state values beyond the lowest possible value for the SF-6D. This means that the magnitude of the differences between paired scores will inevitably be greater at the lower end of the scale. It also highlights the skewed nature of EQ-5D index scores, with SF-6D index scores follow resemble a Normal distribution. The mean index score for the EQ-5D was lower than for the SF-6D at each data collection stage (only the combined sample data is presented in Table 4). The differences between paired index scores ranged from 0.019 (6-week follow-up) to 0.035 (baseline) and were statistically significant at baseline and for the combined sample ($p < 0.001$). The reverse was true for median values, where the EQ-5D provided higher median scores at each time point compared to the SF-6D, although no significant differences were found. This is reflective of the negatively skewed distribution of EQ-5D scores.

Figure 2: Boxplots for observed utility scores from the EQ-5D and SF-6D by data collection point

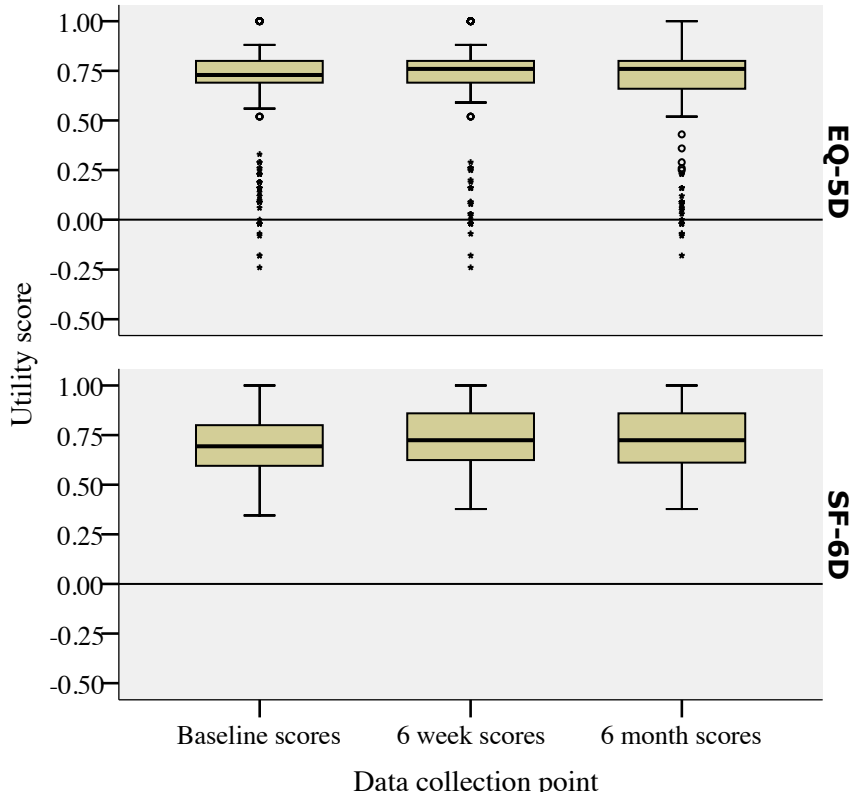


Table 4: Descriptive statistics of observed EQ-5D scores, SF-6D scores and the difference between paired observations for the combined sample[a]

	EQ-5D	SF-6D	Difference[b]
Mean	0.686	0.716	0.028
Standard deviation	0.251	0.141	0.195
95% CI	0.670 to 0.703	0.706 to 0.725	0.015 to 0.041
99% CI	0.664 to 0.708	0.703 to 0.728	0.011 to 0.046
Median	0.760	0.720	-0.007
Inter-quartile range	0.690 to 0.800	0.611 to 0.859	-0.100 to 0.103
Minimum	-0.240	0.345	0.000
Maximum	1.000	1.000	0.695

[a] Difference equals SF-6D minus EQ-5D for the respective sample of paired observations. See table 7.2 for the corresponding number (%) of responses.

[b] The minimum and maximum values are the smallest and largest absolute values.

Table 5 presents mean scores for the theoretical constructs concerning EQ-5D and SF-6D index scores for the combined sample (descriptive statistics for the constructs regarding change scores and pain-dimension responses are not reported but are available from the authors on request). For the analysis of construct validity, statistically significant linear trends were identified for all analyses that focused on index scores ($p<0.001$); such trends were consistent across the individual time points and the combined sample. For the neck pain-related index score constructs, the EQ-5D and SF-6D generated scores that decreased linearly with (i) increased neck pain disability, measured by the Northwick Park neck pain questionnaire, (ii) increased reporting of pain on a 0-10 numerical rating scale, and (iii) average pain category (mild, moderate and severe). For the generic index score constructs, both measures generated scores that decreased linearly with deteriorating self-reported health status on a 5-point ordinal measure and decreasing self-rated health status on the EQ-VAS. The t-test performed on respondents in paid employment showed that EQ-5D and SF-6D scores are significantly lower for individuals that reported a period of neck pain-related work absence during follow-up compared to those individuals who did not report a period of neck pain-related work absence.

There were no significant linear trends in EQ-5D or SF-6D change scores across the 5-level ordinal measure of self-reported neck pain change at 6-week follow-up. However, at 6-months, significant linear trends were identified for EQ-5D and SF-6D change scores. With regard to the constructs based on EQ-5D and SF-6D pain dimension response options, statistically significant linear trends were identified for both measures ($p<0.001$). This means that increments in the response options on the EQ-5D pain/discomfort dimension and the SF-6D pain dimension are associated with significant increases in mean scores on the Northwick Park neck pain questionnaire and significant decreases in mean scores on the 0-10 numerical rating scale for self-rated neck pain.

4. Discussion

Indirect preference-based measures of health-related quality of life are increasingly being used in all areas of health services research, primarily because they provide outcome measurements suitable for use in economic evaluation. The analysis reported in this paper investigated the comparative performance of the two most widely-used UK-specific preference-based measures, the EQ-5D and the SF-6D, within a sample of clinical trial participants with nonspecific neck pain. Many previous studies have demonstrated poor agreement between these two measures and, consequently, the primary research question was to provide insights as to why the EQ-5D and SF-6D do not provide comparable utility estimates for given health states. The expected *dis*agreement between EQ-5D and SF-6D scores was observed across the full utility scale range, although the magnitude of disagreement was greater towards the lower end.

Table 5: Mean EQ-5D and SF-6D scores for the hypothesised preference rules for the construct validity analysis (combined sample)[a]

	EQ-5D		SF-6D	
	n (%)	mean (sd)	n (%)	mean (sd)
Northwick Park NPQ[b,c]				
Group 1 (lowest scores)	297 (33)	*0.826 (0.16)*	283 (33)	0.801 (0.11)
Group 2	298 (33)	*0.727 (0.13)*	283 (33)	0.737 (0.12)
Group 3 (highest scores)	298 (33)	*0.509 (0.30)*	284 (33)	0.610 (0.12)
Pain severity NRS[d]				
0	75 (8)	*0.876 (0.22)*	71 (8)	0.806(0.12)
1	81 (9)	*0.858 (0.13)*	74 (9)	0.806 (0.11)
2	103 (11)	*0.784 (0.12)*	99 (12)	0.789 (0.12)
3	126 (14)	*0.742 (0.17)*	121 (14)	0.770 (0.12)
4	124 (14)	*0.696 (0.19)*	121 (14)	0.722 (0.12)
5	138 (15)	*0.656 (0.21)*	129 (15)	0.683 (0.13)
6	97 (11)	*0.649 (0.20)*	93 (11)	0.661 (0.11)
7	75 (8)	*0.547 (0.31)*	68 (8)	0.633 (0.11)
8	55 (6)	*0.435 (0.32)*	52 (6)	0.580 (0.12)
9	18 (2)	*0.203 (0.24)*	17 (2)	0.526 (0.08)
10	13 (1)	*0.257 (0.29)*	13 (2)	0.522 (0.13)
Pain severity category				
Mild	509 (56)	*0.778 (0.18)*	486 (57)	0.773 (0.12)
Moderate	310 (34)	*0.627 (0.24)*	290 (34)	0.664 (0.12)
Severe	86 (10)	*0.359 (0.32)*	82 (10)	0.560 (0.11)
Self-reported health status				
Excellent	19 (2)	*0.808 (0.16)*	18 (2)	*0.856 (0.11)*
Very good	193 (21)	*0.812 (0.15)*	188 (22)	*0.808 (0.11)*
Good	436 (48)	*0.728 (0.20)*	415 (48)	*0.733 (0.12)*
Fair	213 (24)	*0.579 (0.26)*	198 (23)	*0.622 (0.10)*
Poor	43 (5)	*0.190 (0.29)*	40 (5)	*0.496 (0.09)*
EuroQol visual analogue scale[c,e]				
Group 1 (lowest scores)	301 (33)	*0.516 (0.30)*	284 (33)	0.616 (0.12)
Group 2	301 (33)	*0.740 (0.17)*	285 (33)	0.727 (0.12)
Group 3 (highest scores)	302 (33)	*0.804 (0.16)*	285 (33)	0.805 (0.11)
Employed respondents[f]				
Reported time off work	92 (18)	0.670 (0.23)	89 (18)	0.682 (0.14)
Did not report time off work	421 (82)	0.770 (0.17)	404 (82)	0.766 (0.12)

[a] All tests for linear trend were statistically significant ($p<0.001$), in line with the theoretical constructs. Figures in italics denote analyses where equality of variances across groups (Levene's test) could not be assumed.

[b] Higher NPQ scores reflect a greater degree of disability.

[c] Groups were generated using tertile values.

[d] NRS = numerical rating scale

[e] Higher EQ-VAS scores reflect a greater self-reported general health status.

[f] Sample consists of respondents in full-time or part-time employment.

The two striking differences between the EQ-5D and the SF-6D related to response/completion rates and ceiling effects. The EQ-5D provided a higher proportion of valid utility estimates at each data collection stage. Multiple data collection stages provided an opportunity to compare response rates for successive completion of the instruments. This is an important practical consideration for researchers because utility measures are primarily used to calculate QALYs. On our data set, there was a statistically significant 10% difference between the measures, in favour of the EQ-5D, in terms of the proportion of responders providing complete data throughout the study (74% compared to 64%).

The discrepancy between response rates was due to a single item on the SF-6D, the physical component of the role limitations dimension. Previous research has suggested that the SF-36 and SF-12 have a number of problematic questions that may adversely affect response rates, which include the two items that comprise the role limitations dimension on the SF-6D. Firstly, it has been suggested that the rubric that precedes the two items and the subsequent response options do not complement each other; the two questions are phrased as problems ('how much of the time have you has any of the following problems'), although the response options are not phrased as responses to problems (Jenkinson, 1995; Mallinson, 1998).

A second consideration focused on reference to 'work' within the two role limitation items of the SF-6D (Hayes et al, 1995), suggesting that such a term may cause confusion for respondents that are not engaged in employment. Within our data, respondents that had failed to complete the physical role limitation question were more likely to be older and not currently in paid employment. This result concurs with the findings of Barton et al (2008), where the two items in the role limitation dimension were less likely to be completed by individuals who were not currently in work. However, neither of the issues raised previously in the literature offers an explanation why the item-completion rate for the physical role limitation item was the only anomalous result. Other items, such as the emotional role limitation item, are phrased as problems and refer to 'work' within the rubric, yet provided comparable item-completion rates with other SF-6D and EQ-5D items.

Ceiling effects were observed within EQ-5D index scores and individual dimensions. For the EQ-5D, 13% of paired observations within the combined sample indicated full health, compared to less than 1% on the SF-6D. Within individual EQ-5D dimensions, the top-level response option was reported in over 47% of responses with the exception of the pain/discomfort dimension; for the SF-6D, only the social functioning dimension had a top-level response rate over 40%. The presence of ceiling effects is suggestive that extreme items are missing towards the upper end of the scale. The consequence is that patients with the highest possible score (an index score of 1.00 or a top-level dimension-specific response) cannot be distinguished from each other. As has been pointed out by previous commentators, such deficiencies raise concerns about content validity and indicate poor reliability (Terwee et al, 2007). More specifically, these findings suggest that the 3-level response option within the EQ-5D is insufficient as it does not allow respondents to indicate levels of impairment when in relatively mild health states.

A third difference between the EQ-5D and SF-6D emerged from examining the dimension-to-dimension correlations and distribution of responses within individual dimensions. The vitality dimension on the SF-6D had relatively small correlation coefficients with each dimension of the EQ-5D and the distribution of responses was markedly different than any other dimension across both measures. These findings suggest that vitality is an aspect of health that is not picked up by any of the EQ-5D dimensions. Previous research has indicated that vitality (Grieve et al, 2009), social functioning (Grieve et al, 2009) and role limitations (Konerding, 2009) are SF-6D dimensions that are not adequately captured by the EQ-5D, although the techniques and judgments used to reach such conclusions are varied.

One of the 'indistinguishable' elements of the comparative analysis was the examination of theoretically derived constructs. It was not possible to choose the 'better' measure based on the linear trends explored for each theoretical construct. While this is true, an assessment of whether mean scores are able to distinguish between explicit categories on a defined variable (e.g. self-reported general health status) takes no account of the absolute value of the mean scores observed in the categories. Consideration of the location of mean scores on the respective scoring ranges is necessary to make inferences about the *comparative* performance of the two measures. SF-6D index scores associated with the lowest response categories on the pain severity numerical rating scale, pain severity 3-level ordinal variable and self-reported general health status provided mean SF-6D index scores that would not be considered low despite their association with poor/severe health states (see Table 5). Each of the mean EQ-5D scores for the same response categories were below the minimum possible score on the SF-6D but someway short of the lowest possible EQ-5D score, which indicates that the EQ-5D is better suited to capture the magnitude of severity for poor health states.

Given that preference-based utility measures are intended to be used across all clinical conditions, the narrow scoring range of the SF-6D and the associated inability to reflect the magnitude of disutility associated with poor health states creates a problem because of the normative application of preference-based utility measures. The SF-6D will overestimate QALY estimates for participants in severe health states. The consequence is that utility increments (or, more specifically, QALY increments) for health care interventions aimed at treating severely ill patients may be smaller when using the SF-6D compared to the EQ-5D.

A second component of the analysis for which the two measures were deemed indistinguishable related to the contextual framing of the questions. An examination of question formats led to the supposition that so-called 'similar dimensions' across the two measures should not be expected to provide similar response patterns due to the different framing of the questions, i.e. the focus on severity within the EQ-5D compared to the SF-6D requirement for respondents to consider factors such as duration and interference with normal daily activities. This difference, coupled with the ceiling effects observed for the EQ-5D and the apparent unique contribution of the SF-6D vitality dimension, shows that the descriptive classification systems differ to such an extent that

contemporaneous EQ-5D and SF-6D valuations attached to health states should not be expected to provide similar estimates, irrespective of elicitation techniques used in the respective valuation studies.

The analysis applied to this sample of EQ-5D and SF-6D scores provides a number of generalisable insights into between-measure discrepancies. However, two primary caveats exist. Firstly, the application of parametric statistical methods (*t*-tests and ANOVA) to explore linear trends in the construct analyses requires the data to be approximately normally distributed. Typically, the distribution of SF-6D scores within a given data set will be similar to a Normal distribution, while EQ-5D scores are negatively skewed. However, it has been demonstrated that parametric techniques are robust to violation of the normality assumption that is common to many quality-of-life outcomes that have discrete, bounded and skewed distributions (Walters and Campbell, 2004).

The second limitation relates to the 'completeness' of psychometric properties permitted by the type of data available. A number of commentators have highlighted the importance of reproducibility (or repeatability) and responsiveness in method comparison studies (Bland and Altman, 1999; Terwee at al, 2007). The general premise behind such propositions is that lack of agreement in studies with only one observation per individual may be because of poor repeatability or responsiveness within the standard method; i.e. if the standard comparator is poor, a new method would not agree with it even if it were a perfect measurement tool. The classical definition of responsiveness asserts that health status questionnaires should have the ability to detect clinically important differences over time, irrespective of the magnitude of the difference (Guyatt et al, 1989). There has been no consideration of clinically important differences throughout this analysis, due to the authors' view that it is inappropriate quantify a clinically important difference for outcomes that reflect societal preferences. A less stringent definition of responsiveness relates to 'longitudinal validity,' i.e. the assessment of pre-defined hypotheses regarding change scores (Terwee et al, 2007). Adopting this approach, the analysis of theoretical constructs using EQ-5D and SF-6D change scores infers that the two measures were indistinguishable in their responsiveness to changes in self-reported neck pain status.

The analysis provided in this chapter adds to the growing body of literature confirming that the EQ-5D and SF-6D are empirically valid preference-based measures of health-related quality-of-life but that they do not provide comparable utility estimates for given health states. Irrespective of the oft-cited dissimilarity of scoring ranges across the two measures, differences across the respective descriptive classification systems in terms of the context of question formats, number of available response options and the unique contribution of vitality to the SF-6D instrument demonstrate why utility estimates should not be expected to be comparable. Selecting the 'better' measure is problematic. In their current format, the wider scoring range and better completion rates of the EQ-5D are sufficient for it to remain the industry standard.

References

Bansback N, Sun H, Guh DP, Li X, Nosyk B, Griffin S, Barnett PG, Anis AH; OPTIMA TEAM. Impact of the recall period on measuring health utilities for acute events. Health Econ. 2008; 17: 1413-9

Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, Muir KR. A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years. Health Econ. 2008; 17(7): 815-32

Bland M. An Introduction to Medical Statistics. Oxford: Oxford University Press; 1987

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476): 307-10

Bland JM, Altman DG. Correlation, regression, and repeated data. BMJ. 1994; 308: 896

Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999; 8(2): 135-60

Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. Health Econ. 1999; 8(1): 41-51

Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004; 42(9): 851-9

Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004; 13(9): 873-84

Bryan S, Longworth L. Measuring health-related utility: why the disparity between EQ-5D and SF-6D? Eur J Health Econ. 2005; 6(3): 253-60

Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37(1): 53-72

Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a UK general population survey. York: Centre for Health Economics (Discussion Paper No. 138); 1995

Donaldson C, Gerard K, Mitton C, Jan S, Wiseman V. Economics of health care financing: the visible hand. 2nd ed. Basingstoke, UK: Palgrave Macmillan; 2004

Dziedzic K, Hill J, Lewis M, Sim J, Daniels J, Hay EM. Effectiveness of manual therapy or pulsed shortwave diathermy in addition to advice and exercise for neck disorders: a pragmatic randomized controlled trial in physical therapy clinics. Arthritis Rheum. 2005; 53(2): 214-22

Grieve R, Grishchenko M, Cairns J. SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility. Eur J Health Econ. 2009; 10(1): 15-23

Guyatt GH, Deyo RA, Chalson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol. 1989; 42: 403-8

Hayes V, Morris J, Wolfe C, Morgan M. The SF-36 health survey questionnaire: is it suitable for use with older adults? Age Ageing. 1995; 24(2): 120-5

Jenkinson C. Evaluating the efficacy of medical treatment: possibilities and limitations. Soc Sci Med. 1995; 41(10): 1395-401

Konerding U, Moock J, Kohlmann T. The classification systems of the EQ-5D, the HUI II and the SF-6D: what do they have in common? Qual Life Res. 2009; 18: 1249–1261

Leak AM, Cooper J, Dyer S, Williams KA, Turner-Stokes L, Frank AO. The Northwick Park Neck Pain Questionnaire, devised to measure neck pain and disability. Br J Rheumatol 1994; 33: 469–74

Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther. 1996; 18(5): 979-92

Mallinson S. The Short-Form 36 and older people: some problems encountered when using postal administration. J Epidemiol Community Health. 1998; 52(5): 324-8

McDonough CM, Grove MR, Tosteson TD, Lurie JD, Hilibrand AS, Tosteson AN. Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among spine patient outcomes research trial (SPORT) participants. Qual Life Res. 2005; 14(5): 1321-32

Räsänen P, Roine E, Sintonen H, Semberg-Konttinen V, Ryynänen OP, Roine R. Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. Int J Technol Assess Health Care. 2006; 22(2): 235-41

Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res. 1998; 7(3): 301-17

Shrout, PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 1979; 86: 420-428

Søgaard R, Christensen FB, Videbæk TS, Bünger C, Christiansen T. Interchangeability of the EQ-5D and the SF-6D in Long-Lasting Low Back Pain. Value in Health. 2009; 12(4): 606-612

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC.Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007; 60(1): 34-42

Walters SJ, Campbell MJ. The use of bootstrap methods for analysing Health-Related Quality of Life outcomes (particularly the SF-36). Health Qual Life Outcomes. 2004; 2: 70

Ware J Jr, 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(3): 220-33

Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992; 30(6): 473-83

Whitehurst DGT, Bryan S, Lewis M. The Interchangeability of Utility Measures: a Systematic Review of Contemporaneous EQ-5D and SF-6D Group Mean Scores. Summer 2008 HESG meeting, Aberdeen.