Article:
Brazier, J.E., Roberts, J. and Deverill, M. (2002) The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics, 21 (2). pp. 271-292. ISSN 0167-6296

https://doi.org/10.1016/S0167-6296(01)00130-8

Not for citation or quotation

Accepted for publication in the Journal of Health Economics (September 2001)

The Estimation of a Preference-Based Measure of Health from the SF-36

John Brazier, Jennifer Roberts and Mark Deverill

Sheffield Health Economics Group, University of Sheffield, UK

Abstract

This paper reports on the findings of a study to derive a preference-based measure of health from the SF-36 for use in economic evaluation. The SF-36 was revised into a six dimensional health state classification called the SF-6D. A sample of 249 states defined by the SF-6D have been valued by a representative sample of 611 members of the UK general population, using standard gamble. Models are estimated for predicting health state valuations for all 18,000 states defined by the SF-6D. The econometric modelling had to cope with the hierarchical nature of the data and its skewed distribution. The recommended models have produced significant coefficients for levels of the SF-6D, which are robust across model specification. However, there are concerns with some inconsistent estimates and over prediction of the value of the poorest health states. These problems must be weighed against the rich descriptive ability of the SF-6D, and the potential application of these models to existing and future SF-36 data set.

JEL classification: I10

Key words: Preference-based health measure, SF-36, modelling stated preference data

Address for correspondence:
John Brazier
Sheffield Health Economics Group
School of Health and Related Research
University of Sheffield
Regent Court, 30 Regent Street
Sheffield, S1 4DA, UK.
Telephone: (UK) 0114 2220715.
Fax: (UK) 0114 2724095
Email: J.E.Brazier@sheffield.ac.uk

1. Introduction

Measures of health related quality of life (HRQoL) have become widely used by clinical researchers and can provide useful descriptive information on the effectiveness of health care interventions covering such disparate range of outcomes for HRQoL. However, these measures have not been designed for use in economic evaluation. The main shortcoming of using such instruments in economic evaluation is that they do not explicitly incorporate preferences into their scoring algorithms.

This paper reports on a study to derive a preference-based measure of health from the SF-36, which is one of the most widely used generic measures of HRQoL in clinical trials. It has the potential to considerably extend the scope for undertaking economic evaluation in health care using existing and future SF-36 data sets. The paper also seeks to address the methodological issues this research task raises.

The next section of this paper briefly describes the SF-36 and some of the problems of using it in its current form in economic evaluation. This is followed by a section describing the methods of the study, including: the rationale for the choice of approach, the changes made to the SF-36, the valuation survey using a version of standard gamble and the issues around modelling the data. The valuation survey is reported in sections four and five and the modelling reported in section six. These types of stated preference data are complex to model due to their hierarchical nature and skewed distribution, and section six outlines alternative specifications for dealing with these problems. The final section considers how the results from this work can be used.

2. The Short Form-36 (SF-36) Health Survey

The SF-36 health survey is a standardised questionnaire used to assess patient health across eight dimensions (Ware et al, 1993). It consists of items or questions which present respondents with choices about their perception of their health. The physical functioning dimension, for example, has 10 items to which the patient can make one of three responses: 'limited a lot', 'limited a little' or 'not limited at all'. These responses are coded 1, 2 and 3 respectively and the ten coded responses summed to produce a score from 10 to 30. These raw dimension scores are transformed onto a 0 to 100 scale, which are not comparable across dimensions.

There is extensive evidence of the ability of these scores to describe the health differences between different patient groups and more importantly for evaluation, their ability to detect health changes in populations following intervention (Garratt et al, 1993). However, the method of scoring the SF-36 is not based on preferences. The simple scoring algorithm it uses assumes equal intervals between the response choices (e.g. the change from 'no limitation' to 'limited a little' is regarded as the same the change from 'limited a little' to 'limited a lot'). Furthermore, it assumes the items are of equal importance; for example, being limited in walking has the same importance as being limited in climbing flights of stairs. The evidence has confirmed these concerns with the scoring. Studies have found only low to moderate correlations between HRQoL measures and preference-based measures (Revicki and Kaplan, 1993). The absence of preference data makes it impossible to undertake any trade-offs between SF-36 dimensions, or between its dimensions and survival and/or cost.

The remainder of this paper describes a study that introduces preferences into the scoring of the descriptive data in order to generate the health state utility values needed to construct QALYs and hence conduct cost-utility analyses.

3. Methods

There are three components to this study. Firstly, the SF-36 has been reduced in size and complexity in order that respondents can process the information and hence give reliable valuations of health states. Secondly, a preference based valuation survey has been undertaken. Thirdly, the results of the survey were used in a model to predict values for all states of health described by the reduced form version of the SF-36, via alternative econometric techniques.

Econometric methods to estimate a model to predict health state values was chosen over techniques based on multi-attribute utility (MAU) theory, such as used to value the HUI (Torrance, 1996), due to the structure of the SF-6D system. The dimensions of the SF-6D are not strictly independent, so for example a health state with one dimensions at its worse level and all the other being at the best level is extremely unlikely to occur in practice and would not be credible with respondents. This presents problems in using MAU theory since it becomes necessary to back-off from these 'corner state'. Econometric methods do not rely on such corner states.

The feasibility of this approach was demonstrated in a pilot survey. A specially derived reduced version of the SF-36 (the 'SF-6D') was valued by a convenience sample (n=165) using standard gamble (SG), and the results were modelled to estimate a scoring algorithm for deriving a preference based single index from the SF-36 (Brazier et al, 1998). This pilot study was constrained by the unrepresentativeness of the sample of respondents and limitations in the modelling owing to the small size of the dataset. Therefore the study reported in this paper was designed using a much larger representative sample of the UK population.

3.1 Deriving the SF-6D Health State Classification

The SF-36 has 35 multi-level items used in its current scoring system, many of which have no obvious ordinal relationship; hence many millions of health states can be defined from this classification. The valuation of such a large multi-attribute function would present enormous estimation problems. Furthermore, experience from other research in transport and health, suggests that individuals can only process between five and nine pieces of information at a time (Miller, 1956; Pearmain et al, 1991; Dolan et al, 1995). The aim of this stage of the project was to produce a health state classification which was amenable to valuation by respondents subject to the constraint that responses to the SF-36 could be unambiguously mapped onto it.

The main task was to substantially reduce the number of items for the health state classification. The principle criterion was to minimise the loss of descriptive information. This item selection process was able to benefit from the research undertaken by Dr John Ware and his colleagues on the descriptive importance of the items of the SF-36 in terms of their overall contribution to longer versions of the dimension scores (Ware et al, 1995). They undertook extensive factor analyses to determine the relative contribution of items to their overall dimension scores. This work has already contributed to a further shortened version of the instrument, the SF-12, which has become widely used in the USA.

3.2 The SF-6D health state classification

The number of dimensions was reduced from eight to six. This was achieved firstly, by excluding all general health items since the purpose is to generate a single index for health

and it would be illogical to include a general health dimension as a constituent element. Secondly, the dimension 'role limitation due to physical problems' was combined with 'role limitations due to emotional problems' to form a single dimension for simplicity. However, this important distinction between different sources of role limitation is not lost in the derived health state classification system.

The final selection of items uses 8 from the SF-12 and three other items from the SF-36 physical functioning dimension to extend the scale to cover the full range of functioning problems. The result is a six dimensional health state classification shown in Table 1, which has been called the SF-6D. The SF-36 items used in the SF-6D are listed at the bottom of the table. This version of the SF-6D differs in a number of important respects from the pilot version published in Brazier et al, (1998).

The SF-6D has six dimensions ($\delta = 1,2,\ldots,6$), each with between two and six levels (λ). An SF-6D health state is defined by selecting one statement from each dimension, starting with *physical functioning* and ending with *vitality* (see Table 2 for examples). A total of 18000 health states can be defined in this way. All responders to the original SF-36 questionnaire can be assigned to the SF-6D provided the 10 items used in the six dimensions of the SF-6D have been completed.

4. The valuation survey

The basic design of the survey was that a sample of 249 health states defined by the SF-6D was valued by a representative sample of the general public (n = 836). Each respondent was asked to rank, and then value, six of these states using a variant of the SG technique.

4.1 Selection of respondents

A representative sample of the general population should be used in this survey since the purpose is to inform the allocation of public resources (Gold et al, 1996). The aim of the sampling has been to ensure the sample should reflect the variability of the population in terms of characteristics such as age, socio-economic status and level of education. The sample was drawn using a two-stage cluster random selection design. The primary units were postcode sectors stratified by percentage of households with a non-manual occupation. Fifty one postcode sectors were selected, and addresses randomly selected from each of these, resulting in 1445 potential interviews. Where more than one adult (i.e. 16 or over) was found in household, one was selected at random by the interviewer using a standard Kish selection grid.

4.2 Selection of health states

For such a large descriptive system, where it is not possible to value all possible combinations of each dimension or attribute, there is little guidance in the statistics literature on selecting samples for valuation. The minimum sample of health states was identified using an orthogonal design (by applying the Orthoplan procedure of SPSS), which generated 49 health states (out of 18,000) required in order to estimate an additive model. It was anticipated that more complex specifications, allowing for some form of interaction between dimensions, would be estimated and therefore it was desirable to value more states. Another reason for valuing more states was to provide scope for examining the predictive ability of the models subsequently estimated. However, resources constrained the survey to around 800 interviews. The choice was therefore to maximise the number of valuations per state (hence choose the

minimum number of states, 49), or maximise the number of health states valued or some combination of the two. The latter course was chosen. States were classified as mild, moderate or severe and a stratified sampling method was used to supplement the 49 states selected by Orthoplan with a further 200 states, to provide 249 health states for valuation.

Each respondent was asked to value six health states. Care was taken to ensure each person was asked to value a range of health states across the space defined by the SF-6D rather than a predominantly 'good' or 'bad' selection of states (Brazier et al, 1999b). The allocation procedure was also designed to maximise the chance that each of the 249 cards would be valued by an equal number of respondents.

4.3 The interviews

A trained and experienced interviewer conducted the interviews in the respondent's own home. The interviewers were employed by the Social and Community Planning Research (SCPR), who are a private survey organisation that has undertaken numerous surveys for Government agencies and academics, including the MVH EQ-5D valuation survey (Dolan et al, 1995). The interview began with the respondent being asked to complete a short self-completion questionnaire about his or her own state of health, that included completing the SF-6D in the format that appears in Table1. This familiarised the respondent with the idea of describing health in terms of the SF-6D. It also provides a self-assessment of health which could be subsequently used in the modelling to estimate the impact of respondent's own health on their valuation of other health states.

At the next stage of the interview, the respondent was asked to rank a set of eight cards: one for each of the health states they would have to value, along with the best state defined by the SF-6D, the worst state and immediate death. This task provided an opportunity for the

respondent to familiarise themselves with the cards and the notion of having preferences for one health state over another.

The main part of the interview was the SG valuation of six health states. This study employed a variant of the SG using props developed by a team at McMaster (Furlong et al, 1990). In the interview, the respondent is asked to choose between the certain prospect (A) of living in an intermediate state defined by the SF-6D and the uncertain prospect (B) of two possible outcomes, the best state defined by the SF-6D or the worst ('pits'). The chances of the best outcome occurring is varied until the respondent is indifferent between the certain and uncertain prospects. At all times the probabilities are displayed on a chance board, both numerically and in the form of a pie chart. This 'ping pong' with props version of SG was chosen for its ease of use by interviewers. The chance board is designed to make the interview as straightforward as possible, by leading the interviewer through a set of questions depending on the respondents answer to the previous question, and minimise the risk of interviewer variation. The developers have tried and tested the procedure and its associated prop over many years and it has become widely used in health economics. The McMaster team were able to provide training to the study investigators and produce the chance boards for the survey interviewers. All interviewers were trained in the use of this SG valuation technique by the investigators.

In the SG valuation task respondents were asked to value each of the five certain SF-6D health states against the best and 'pits' health state. For calculating QALYs it is necessary to transform the results onto a scale where 1 is full health and 0 equivalent to death. The best health state defined by the SF-6D is full health. The worst state defined by the SF-6D must be valued on the full health to death scale and the five health state values adjusted accordingly. All respondents were therefore asked a sixth SG question. Depending on

whether they thought the 'pits' state was better or worse than death they would be asked to consider a choice between either: i) the certain prospect of being in the 'pits' state and the uncertain prospect of full health (state 111111) or immediate death, or ii) the certain prospect of death and the uncertain prospect of full health or the 'pits' state.

The use of the 'pits' state as the reference state is an important feature of the SG valuation task used in the survey. It is more conventional to use death as the worst outcome (and more convenient for the purposes of deriving QALYs where it is necessary to place the health state values directly onto a scale where 1 is full health and 0 is regarded as equivalent to death). The 'pits' state was chosen for two reasons. The first arose from a concern that the 'ping pong' version of SG used in the survey was insensitive at the upper end of the scale. Respondents are asked to consider probabilities up to 0.95 and yet the pilot study using an earlier version of the SF-6D found many respondents would only consider having the operation at higher odds (Brazier et al, 1998). The two stage valuation process allows respondents who believe the 'pits' state is better than death (and most did) to value the intermediate state above 0.975. The second benefit is that it enables people who regard the 'pits' state and other health states defined by the SF-6D to be worse than death to do so in one go at the end of the interview rather have to incorporate the states worse than death gamble into every question.

It was necessary to 'chain' the health state values in order to place them on the zero to one scale. For health states better than death, where the best outcome is set at 1 and death is 0, then expected utility theory would indicate that the health state value of the intermediate state is the probability of the better outcome at the respondent's point of indifference. For states worse than death, the equivalent value would be -P/(1-P); where P is the valuation of the 'pits' state. However, this results in a scale ranging from -∞ to +1, which gives greater weight

to negative values in the calculation of mean scores and presents problems for the statistical analysis. It has therefore been recommended in the literature that states valued worse than death should be simply the negative of the indifference probability of the best outcome (Patrick et al, 1994). This has the effect of bounding negative values at minus one. It is acknowledged that this has no theoretical support and is only one of a number of possible ways of dealing with the problem, but it is one that has become widely used elsewhere in the literature and a similar transformation has been used on TTO values in the MVH study (Dolan, 1997). Furthermore, it is less of an issue in the valuation of the SF-6D since there are proportionately fewer SG observations below zero and very few below minus one using the formula -P/(1-P) than has been found for the HUI and EQ-5D.

Having valued the 'pits' state (P), the final step is to adjust the five intermediate SF-6D health state valuations (SG) onto the scale where the best SF-6D state is 1 and death 0. The health state value used in the modelling is therefore: SGADJ = SG + (1-SG) * P.

5. The Data

Out of the 1445 addresses contacted for interview, 167 proved to be ineligible[1]. Of the usable addresses there were 836 successfully conducted interviews (a 65% response rate). Respondents were found to be representative of the national population in terms of the distribution by age group, education and social class (Sturgis and Thomas, 1998).

One hundred and thirty respondents had to be excluded from the analysis for failing to value the 'pits' state; therefore it was not possible to generate an adjusted SG value (see below). A further 9 were excluded for not valuing two or more health states. Finally, there were 86

respondents whose health state values did not change between the five states. This last group have been excluded because the lack of variation is likely to indicate that the respondent did not understand the task. Other grounds for excluding individuals were considered, such as logical inconsistencies between their responses and the ordinal properties of the SF-6D, but these were discounted in favour of leaving individuals in the data set where possible.

A comparison of included and excluded respondents is presented in Table 3. The 225 excluded cases tended to be older, were marginally more likely to be male and unmarried, and less likely to have children under 16. They were more likely to rent rather than own their home and were less likely to be in full-time employment. They tend to have less educational qualifications and were slightly more likely to have problems understanding the standard gamble valuation task. Out of the 611 individuals included in the data set there were 148 missing values from 117 individuals. This results in 3518 observed SG valuations across 249 health states and these form the data set reported and analysed below.

5.1 Health state values

Descriptive statistics for 50 of the 249 health states are shown in Table 4. Each health state has been valued an average of 15 times. Mean health state values range from 0.10 to 0.99, and generally have large standard deviations. Median health state values usually exceed mean values, reflecting the positive skewness of the data. The relative health state valuations broadly conform with the logical ordering of the SF-6D.

[1] These were addresses which contained no resident household for various reasons including: insufficient address, not traced, not yet built, derelict/demolished, business only, empty, institution only, weekend/holiday home.

At the level of individual observations the degree of skewness is even more evident. A histogram and descriptive statistics for the 3518 individual adjusted health state valuations is shown in Figure 1. Negative observations did occur but were comparatively rare (245/3518) and over 23% of observations lie between 0.9 and 1.0. Interestingly, even for the 'pits' health state most respondents valued it as better than death (445/611). However, very few health states were valued at 1.0 (20/3518), indicating the willingness of respondents to risk a worse health state in order to have the chance of a better state of health.

6. Modelling

The overall aim is to construct a model for predicting health state valuations based on the SF-6D. The appropriate modelling strategy is not clear a priori, and the econometric analysis is necessarily of an exploratory nature (Busschbach et al, 1999). The data generated by the valuation survey described above, has a complex structure which creates a number of problems for econometric estimation. Firstly, the data are skewed and bimodal (see Figure 1). Conventional power transformations are therefore not appropriate. The skewness in the data also raises questions about the appropriate measure of central tendency. There are statistical and political (e.g. median voter) arguments for using the median, but for economic evaluation the mean is usually recommended. The choice of dependent variable in this respect is also influenced by the second consideration - the form of heterogeneity that characterises this data.

Variation is both between respondent and within respondent (across health states). Furthermore, health state valuations are likely to be clustered by respondent. Level 1 denotes the individual health state valuations, which are clustered according to level 2 – the respondents. Respondents did not value the same set of states, although allocation of states to

respondents was essentially random, differences between health state values may be partly due to differences in the preferences of the respondents, rather than the attributes of those states. Disentangling the respondent effect is a complex task and can only be tackled at the individual level, where each valuation is regarded as a separate observation, rather than using the mean value for each health state. The former has the advantage of greatly increasing the number of degrees of freedom available for the analysis (from 249 to over 3500) and enabling the analysis of respondent background characteristics on health state valuations. Despite these apparent advantages, it is not clear whether one is necessarily superior for the purposes of predicting mean health state values (Gravelle, 1995) and hence models have been estimated at both the individual and aggregate levels.

6.1 Models

A number of alternative models can be formulated for predicting the SG gamble scores generated in the valuation survey. The general model is defined as:

$$y_{ij} = g(\beta' \mathbf{x}_{ij} + \theta' \mathbf{r}_{ij} + \delta' \mathbf{z}_j) + \varepsilon_{ij} \qquad (1)$$

where i = 1, 2, …, n represents individual health state values and j = 1,2, …, m represents respondents. The dependent variable**,** y_{ij}, is the adjusted SG score for health state i valued by respondent j (SGADJ). **x** is a vector of dummy explanatory variables ($x_{\delta\lambda}$) for each level λ of dimension δ of the SF-6D. For example, x_{31} denotes dimension δ = 3 (social functioning), level λ = 1 (health limits social activities none of the time). For any given health state, $x_{\delta\lambda}$ will be defined as

$x_{\delta\lambda}$ = 1 if, for this state, dimension δ is at level λ

$x_{\delta\lambda} = 0$ if, for this state, dimension δ is not at level λ

In all there are 25 of these terms, with level $\lambda = 1$ acting as a baseline for each dimension. Hence for a simple linear model, the intercept represents state 111111, and summing the coefficients of the 'on' dummies derives the value of all other states.

The **r** term is a vector of terms to account for interactions between the levels of different attributes. The estimation of all possible interaction terms would have required a substantially larger proportion of the 18,000 health states of the SF-6D to be valued. There are, for example, 465 first order interactions alone. Given, the large number of possible interactions, and little evidence on which are likely to be important, there is a risk of finding significant interactions due to the play of chance. Further discussion of interaction effects is given below. **z** is a vector of personal characteristics that may also affect the value an individual gives to a health state, for example, age, sex and education. The role of personal characteristics is not discussed in this paper. g is a function specifying the appropriate functional form. ε_{ij} is an error term whose autocorrelation structure and distributional properties depend on the assumptions underlying the particular model used.

This is an additive model, which, apart from additivity, imposes no restrictions on the relationship between dimension levels of the SF-6D. For example, it does not enforce an interval scale between the levels of each dimension. Earlier empirical work on valuing the Euroqol assumed equal intervals, but this has since been found to be invalid for certain dimensions (van Hout and McDonnell, 1991, and Dolan, 1997). This additive model does not impose ordinality on the levels.

6.2 Alternative Model Specifications

The starting point is OLS estimation of model (1), with g as a linear function. This simple specification assumes a standard zero mean, constant variance error structure, with independent error terms, that is $\text{cov}(\varepsilon_{ij}, \varepsilon_{i'j}) = 0$, $i \neq i'$. This specification ignores the potential multilevel variation in the data and assumes that each individual health state value is an independent observation, regardless of whether or not it was valued by the same respondent.

An improved specification, which takes account of variation both within and between respondents, is the one-way error components random effects model. This model explicitly recognises that n observations on m individuals is not the same as n×m observations on different individuals. For the random effects model the errors from model (1) are subdivided such that,

$$\varepsilon_{ij} = u_j + e_{ij} \qquad (2)$$

u_j is respondent specific variation, which is assumed to be random across individual respondents. e_{ij} is an error term for the i^{th} health state valuation of the j^{th} individual, and this is assumed to be random across observations, with $e_{ij} \sim [0, \sigma_e^2]$. In addition $\text{cov}(u_j, e_{ij}) = 0$ which signifies that allocation of health states to respondents is random. Estimation is via generalised least squares (GLS) or maximum likelihood (MLE).

A one-way error components fixed effects model can also be specified. This differs from the random effects specification in that the respondent specific effects u_j are not assumed to be random, but are a set of fixed effects to be estimated, together with the vector of coefficients on the explanatory variables; hence $\text{cov}(u_j \mathbf{x}_{ij}) \neq 0$.

The choice between random and fixed effects specification depends largely on the sample design and the purpose of the study. In this case, respondents constitute a random sample and

the assumption of the random effects specification are met. Ultimately the choice is an empirical matter and will be determined by the Hausman test.

The multi-level nature of variation in these data suggests a further class of models for consideration, those developed specifically to deal with hierarchical data structures. The two-level multi-level model is similar to the one-way error components random effects model, and algebraically can be denoted by the specification given in (1) and (2) above. Estimation is by iterative GLS (IGLS), and this allows for more complex modelling of the variance components observed at both levels of the hierarchy (Goldstein, 1995).

Finally we consider alternative functional forms - g in (1) - to account for the skewed distribution of health state valuations. Four functional forms are used. Firstly, a Logit transformation and two complementary log-log transformations suggested by Abdalla and Russell (1995). These are chosen to map the data from the range (-1,1) to the range $(-\infty,\infty)$ via the unit range (0,1).

Before applying these transformations it was necessary to transform the SGADJ data to get rid of negative values using:

$$SGADJU=(SGADJ+1)/2 \qquad (3)$$

Secondly, a Tobit transformation which, although designed to deal with truncated data, can approximate for the left skew in this data, where 25% of the values lie between 0.9 and 1. This is done by specifying a Tobit model with upper censoring at 1.

All modelling was done using EVIEWS 3.1, STATA 6.0 and MLwinN 1.02.

6.3 Interaction Effects

Analysis of first order interaction effects was problematic, since the large number of possible effects means there is a risk of finding some are significant purely by chance. Also when first order interaction terms were found to be significant, they generally displaced the main effects due to collinearity between main effects and first order interaction terms.

It was therefore necessary to investigate alternative ways of accounting for interactions. Extreme level dummies were created to represent the number of times a health state contains dimensions at the extreme ends of the scale (Dolan, 1997). Least severe is defined as level 1 or 2 on each dimension. Most severe is defined as levels 4 to 6 for physical functioning (PF), levels 3 and 4 for role limitation (RL), 4 and 5 for social functioning (SF), mental health (MH) and vitality (VIT), and 5 and 6 for pain. A number of alternative definitions of most and least severe were investigated but these made little difference to the results. The most and least extreme dummies are denoted by EM_δ and EL_δ respectively, where $\delta = 1,2,\ldots,6$ and describes the number of times the least or most severe levels appear in a state.

Two further methods for accounting for any additional effect from dimensions at the most severe levels were tried. Firstly, count variables represent the number of dimensions at the least (most) severe level. Secondly, dummy variables LEAST (MOST) take a value of 1 if any dimension in the health state is at the least (most) severe level, and 0 otherwise. Further dummies MOSTn (n = 2, 3, …,5) takes the value 1 if at least n dimensions are at the most severe level.

Functional Form

Four transformations have been attempted to get remove the skew in the health state valuation data. Three ad hoc adjustments suggested by Abdalla and Russell (1995)[2] and a TOBIT transformation. All of these transformations were modelled with random effects since Breusch Pagan and Hausman tests suggest these are appropriate.

(i) Logit transformation, SGAG1 = ln (SGADJU /(1 – SGADJU)

(ii) complementary log-log transformation SGAG2 = ln (-ln (1 – SGADJU)

(iii) complementary log-log transformation SGAG3 = ln (- ln (SGADJU)

(iv) Tobit transformation (see Breen, 1996).

6.4 Results

Basic Models – Main Effects

The results are shown in Table 5 for OLS and random effects models at the individual level, and OLS models using mean and median health state values. These models include only the main effects dummies.

The main effects dummies represent progressively worse problems on each dimension compared to a base line of no problem for that particular dimension. As such the coefficient estimates are expected to be negative and increasing in absolute size. An inconsistent result occurs where a coefficient on the main effects dummies decreases in absolute size with a worse level.

[2] Abdalla and Russell (1995) did not attempt to model transformed data while also allowing for individual effects. Here we attempt to simultaneously cope with these two characteristics of the data.

For the OLS model (1) the vast majority of coefficients have the expected (negative) sign. In all 13 of the 26 coefficients are significant and there are 2 inconsistencies, where the estimated effect decreases from RL3 to RL4, and VIT2 to VIT3. The explanatory power of the model is 0.204. Diagnostic tests reveal problems with non-normal and heteroscedastic residuals[3]. Further these data represent repeated observations on 611 individuals and a Breusch-Pagan LM test for individual effects reveals that these are important ($\chi^2 = 1717.02$, $p = 0.000$). In addition Hausman's test suggests that random, rather than fixed, effects, is the appropriate specification (($\chi^2 = 27.11$, $p = 0.35$)[4].

For the random effects specification all coefficients have the expected negative sign. There are 17 significant coefficient estimates and 2 inconsistencies, with a decrease in the size of the coefficient from PF4 to PF5 and SF2 to SF3. Explanatory power is 0.200 and the variance decomposition suggests slightly more variation between respondents than within respondents. The Breusch–Pagan test for heteroscedasticity suggests that a problem still exists. The Ramsey RESET test shows no evidence of specification problems which is surprising given the skewness of the residuals.

The mean (3) and median (4) models presented in Table 5 have much greater explanatory power than the individual level models, explaining almost 60% of the variation in health state values. The mean model has serial correlation and heteroscedasticity problems, while the median model appears to have non-normal residuals.

Coefficients can be compared directly across the first 4 models presented in Table 5. There are similarities in that the important effects are found among the most severe levels of each dimension. Most of these effects are robust across model specification.

[3] The model was estimated using White's heteroscedasticity consistent standard errors.

Predictive Ability

Given our overall aim of predicting health state valuations the best way to compare these models is via their predictive ability. Summary statistics for inside sample predictions are presented in the lower half of Table 5. The median model appears to be the worst but there is little to chose between the other three, which have similar mean absolute errors (MAE) and result in similar numbers of errors greater than 0.05 and 0.10 in absolute value. The proportion correctly predicted to within |0.1| was nearly 80%, and 54% to within |0.05|. In all cases the predictions are unbiased (t-test), and prediction errors are normally distributed (JB test).

The most serious problem at this stage is that Ljung-Box (LB) statistics reveal significant autocorrelation in the prediction errors of all models, when the errors are ordered by actual mean health state valuation. Figure 2 shows actual and predicted health state valuations for the random effects model (2). This reveals a tendency to over predict at low health state values (i.e. poor health states) and under predict at high health state values. A similar result is found for all models (1) to (4).

Restricting the intercept to unity

There are strong theoretical arguments for restricting the intercept to unity. The adjusted SG value for each state has been estimated according to the axioms of EUT by assuming SF-6D state 111111 health is to equal one and death is equal to zero. For state 111111 to hold any other value would change the scale. Furthermore, for use in CUA it is necessary to assume

[4] Hierarchical models were estimated using MLwiN 1.02. The results are identical to those for the random effects models to 4 decimal places.

that health state 111111 is equivalent to full health and hence has a value of one. The best way to ensure health state 111111 has a value of one is to restrict the intercept to unity.

For models (5) and (6) in Table 5 the intercept has been restricted to unity. Coefficient estimates can be directly compared to those for models (1) to (4). For both of these models there is a substantial increase in the number of significant coefficient estimates and a slight increase in the number of inconsistencies. While there is a slight increase in error size compared with models (2) and (3) there is less autocorrelation in the errors; although the LB statistics are still significant. Figure 3 shows actual and predicted health state valuations for the random effects model (5). This shows that while the tendency to under predict at good health states has been removed there is still a problem of over prediction at poor health states. A similar result is found for the mean model.

Interaction Effects

Models which include some of the interaction effects discussed above are presented in Table 6. A number of ways of dealing with interaction effects were investigated and these three models are the most successful. The random effects and mean models (7 and 8) include the dummy variables MOST and LEAST, which take a value of 1 if any dimension in the health state is at the most or least severe level. The coefficient estimates suggest a further negative effect if any dimension is at the most severe level which is slightly reduced by a positive effect of dimensions at the least severe level. The coefficients on the main effects dummies are slightly reduced as expected but are robust to the inclusion of the interaction effects. These models show little improvement in predictive ability above models (2) and (3).

Models (9) and (10) are the equivalent with the intercept forced to unity; here only the MOST dummy is significant at $t_{0.05}$. Again these results are very similar to those of models (5) and (6) with little or no improvement in predictive ability.

The alternative functional forms do not perform well in terms of predictive ability. They all give biased predictions (t-test) and in general they result in larger errors than the untransformed models[5].

7. Discussion and conclusion

The results of this study offer a method for analysing existing SF-36 data from trials and other sources of evidence where there is no other means of estimating the preference-based health values for generating QALYs. It also provides an alternative to existing preference-based measures of health for use in cost utility analysis. Two of the leading preference-based measures are the EQ-5D (Brooks, 1996) and the Health Utility Index (Torrance et al, 1995). Whether or not the SF-6D offers an improvement on these existing measures depends on one's view of the appropriate definition of health, the valuation techniques and the best method for modelling health state values (Brazier et al, 1999). There is insufficient space in this paper to go into these issues. However, one of the advantages of the SF-6D over the EQ-5D could come from the much larger size of its descriptive system and hence a possibly greater degree of sensitivity. This must be weighed against the inconsistencies between the coefficients at the upper levels of some SF-6D dimensions. The sensitivity of the new index needs to be compared to other preference-basecd measures before drawing any conclusion on this point. Any greater sensitivity would be most likely in groups experiencing mild to

[5] Predicted values have been retransformed using the smearing estimator (Rutten-van Mölken et al, 1994)

moderate health problems and in those expected to experience comparatively small changes or where small differences are expected between interventions.

An important question is whether the derivation of the SF-6D health state classification has compromised the descriptive richness and sensitivity of the original SF-36. The selection of items was intended to minimise the potential loss of information but the loss may offset the advantages of the SF-36. This is an empirical question to be addressed in future research.

The models have produced significant coefficients for levels of the SF-6D with the expected negative sign. These main effects are robust across model specification and in most cases they are consistent with ordinal levels of the SF-6D. However, there are concerns with the individual level models low explanatory power. At the individual level explanatory power reached 0.2 compared with 0.45 for the York MVH models for the EQ-5D (Dolan, 1997). The size of the mean absolute error was correspondingly smaller. Comparisons between these two pieces of work is difficult since the valuation of the SF-6D is much larger undertaking describing nearly 75 times more states. More relevant for CUA is the ability of the model to predict mean health state values and the best mean model achieved an adjusted R-squared of 0.58.

Another concern is the existence of inconsistencies between coefficients on the SF-6D levels. In many cases the estimated coefficients on lowest levels of each dimension are not statistically significant (e.g. the coefficients on PF2 and PF3 in the recommended model 10), hence the fact that PF3 attracts a point estimate lower than PF2 is not an inconsistency, since they are both interpreted as zero. Therefore we interpret an inconsistency as only occurring between significant coefficients and the number of these is quite low compared to the number of consistent coefficients. Those inconsistencies that occur in more than one of the four models reported in Table 6 are as follows: PF4 vs PF5, RL3 vs RL4, VIT2 vs VIT3 and

PAIN2 vs PAIN 3. There is no clear ordinal relationship between PF4 and PF5 and hence this may not be an inconsistency at all. RL3 vs RL4 have similar coefficients across all models and this indicates that most respondents did not distinguish between them. For VIT2 vs VIT3 one possible explanation is that this dimension is worded in the positive rather than the negative and this may have caused some confusion for respondents. Finally, PAIN2 and PAIN3 are not significant in models (7) and (8) and similar in models (9). In model (10) PAIN3 attracts an insignificant coefficient estimate, whereas PAIN2 is significant suggesting an inconsistency. But like the remaining 3 inconsistencies it occurs only once across the four specifications. We do not believe these inconsistencies have any serious implication for the performance of the model as whole except for a reduction in sensitivity at the upper end for some dimensions. Of course, a larger sample size and the valuation of additional health state may have overcome some of these problems.

Of more concern is the existence of systematic prediction errors resulting from all the models. Introducing interaction terms leads to little improvement in predictive ability and we still have a problem of under predicting the value of good health states and over predicting the value of poor states in the models with an estimated intercept terms. Restricting the intercept to unity eliminated the former problem, whilst the latter remains. We have attempted numerous other alternative specifications for interactions not reported in this paper, but these did not produce significant results.

A number of models have been presented for predicting preference-based health state values from SF-36 data. Whilst we have shown the RE model to be better than OLS at the individual level, we do not believe the RE offers any clear advantages over the mean level models. Indeed, the mean model is marginally better across the different tests of fit. Given the task is to predict mean health state values there is no reason to favour the individual level models

and therefore we recommend using one of the mean level models. The interaction terms lead to very modest improvements in the model and should therefore be used. As argued earlier in this paper, we also favour restricting the intercept to unity for the purposes of generating models for use in CUA. The preferred model for use in CUA is therefore (10) in Table 6[6].

This paper has presented a study to estimate a preference-based single index from one of the larger generic profile measures of health related quality of life. It is only the second time this has been done, the first being essentially a pilot to this study (Brazier et al, 1998). This research demonstrates that it is possible to estimate preference weights for measures of health related quality. The paper presents the key methodological issues involved in undertaking such a task, including the derivation of a health state classification, the valuation survey and modelling. The results can be applied to any SF-36 data set and hence considerably expand the available evidence base for conducting economic evaluation of health care interventions.

[6] A computer algorithm for deriving a preference-based index from SF-36 data via the SF-6D is available from the corresponding author. The algorithm is copyrighted, though it is free of charge for non-commercial uses.

Acknowledgments

We would like to thank GlaxoWelcome for supporting this study and Roger Thomas and Patrick Sturgis at SCPR for conducting the valuation survey. The usual disclaimer applies.

References

Abdalla, M., Russell, I. (1995). Tariffs for the Euroqol health states based on modelling individual VAS and TTO data of the York Survey In : MVH Group *Final Report on the Modelling of Valuation Tariffs* Centre for Health Economics, University of York, UK.

Brazier, JE, Deverill M., Harper R., Booth A. (1999a) A review of the use of Health Status measures in economic evaluation. *Health Technol Assess* 1999;3(9)

Brazier, J. E., Roberts, J., Deverill ,M. (1999b) *The estimation of a utility based algortihm from the SF-36 Health Survey*. Report prepared for GlaxoWelcome, Mimeo.

Brazier JE, Harper R, Thomas K, Jones N, Underwood T (1998) Deriving a preference based single index measure from the SF-36 *J.Clinical Epidemiology* 51 (11):1115-1129

Breen, B (1996) *Regression models: Censored, Sample Selected or Truncated Data*. Sage: London.

Briggs, A., Sculpher, M., Buxton, M. (1994) Uncertainty in the Economic Evaluation of Health Care Technologies: The Role of Sensitivity Analysis. *Health Economics* 3(2): 95-104

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

Busschbach, J.V., McDonnell, J., Essink-Bot, M-L., van Hout, B.A (1999) Estimating parametric relationships between health state description and health valuation with an application to the EuroQol EQ-5D. *J.Health Econ*. 18:551-571.

Dolan, P., Gudex, C., Kind, P. and Williams, A. (1995) A social tariff for Euroqol: Results from a UK general population survey, Centre for Health Economics Discussion Paper 138, University of York.

Dolan, P. (1997). Modelling valuation for Euroqol health states. *Medical Care* 35:351-363

Drummond, M.F., Stoddart, G.L., Torrance, G.W. *Methods for the economic evaluation of health care programmes.* Oxford: Oxford Medical Publications, 1987.

Furlong, W., Feeny D., Torrance, G.W., Barr, R., Horsman J. (1990) Guide to design and development of health state utlity instrumentation. Centre for Health Economics and Policy Analysis Paper 90-9, McMaster University, Hamilton, Ontario.

Garratt, A.M., Ruta, D.A., Abdalla, M.I. et al (1993). The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS. *British Journal of Medicine*; 306: 1440-4.

Gold, M.R., Siegel J.E., Russell L.B., Weistein M.C. (1996) *Cost-Effectiveness in Health and Medicine.* Oxford University Press, Oxford.

Goldstein (1995) *Multilevel Statistical Methods* London: Edward Arnold, New York: Halstead Press.

Gravelle, H. (1995) Valuations of Euroqol health states: comments and suggestions. Paper presented at the ESRC/SHHD Workshop on Quality of Life, Edinburgh, unpublished.

Kaplan, R.M., & Anderson, J.P. (1988). A general health policy model: update and application. *Health Services Research*, *23,* 203-235.

Miller, G.A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psych Rev* 1956;63:81-97.

MVH group. (1994). The measurement and valuation of health: first report on the main survey. Centre for Health Economics, University of York.

Patrick, D.L., Starks, H.E., Cain, K.C., Uhlmann, R.F., & Pearlman, R.A. (1994). Measuring preferences for health states worse than death. *Medical Decision Making*, *14,* 9-18.

Pearmain, D., Swanson, J., Kroes, E., Bradley, M. (1991). Stated preference techniques: a guide to practice. Steer Davis Gleave and Hague Consulting Group, Hauge.

Revicki, D.A., & Kaplan, R.M. (1993). Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Quality in Life Research*, *2,* 477-487.

Rutten-van Mölken, M.P.M.H., van Doorslaer, E.K.A. and van Vliet, R.C.J.A. (1994) Statistical analysis of cost outcomes in a randomized controlled clinical trial. *Health Economics*, 3(5), 333-46

Streiner, D.L., Norman, G.R. (1989). *Health Measurement Scales: a practical guide to their development and use*. Oxford: Oxford University Press.

Sturgis, P; Thomas, R. (1998) *Deriving a preference based utlity score for the SF-6D: Technical report*. Survey methods Centre at SCPR.

Torrance, G.W., Furlong, W., Feeny, D., & Boyle, M. (1995). Multi-attribute preference functions. Health Utilities Index. *Pharmaco Economics*, *7,* 503-520.

Torrance, G.W. (1986). Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics*, *5,* 1-30.

Torrance, G.W., Boyle, M.H., & Horwood, S.P. (1982). Applications of Multi-Attribute Utility Theory to measure social preferences for health states. *Operations Research*, *30,* 1043-1069.

van Hout, B.A. and McDonnell, J. (1992). *Estimating a parametric relation between health description and health valuation using Euroqol Instrument.* In: Euroqol Conference Proceedings, IHE Working Paper 1992:2, Lund, Sweden.

Ware, J. E., Snow, K. K., Kolinski, M., Gandeck, B. (1993) *SF-36 Health Survey manual and interpretation guide*. Boston: The Health Institute, New England Medical Centre, Boston, MA.

Ware, J. E., Kolinski, M., Keller SD (1995) *How to score the SF-12 physical and mental health summaries: a user's Manual*. Boston: The Health Institute, New England Medical Centre, Boston, MA.

Williams, A. (1992). Measuring functioning and well-being, by Stewart and Ware. Review article, *Health Economics*; 1 (4): 255-258.

Figure 1: Histogram and Descriptive Statistics for Adjusted Health State Valuations (SGADJ)

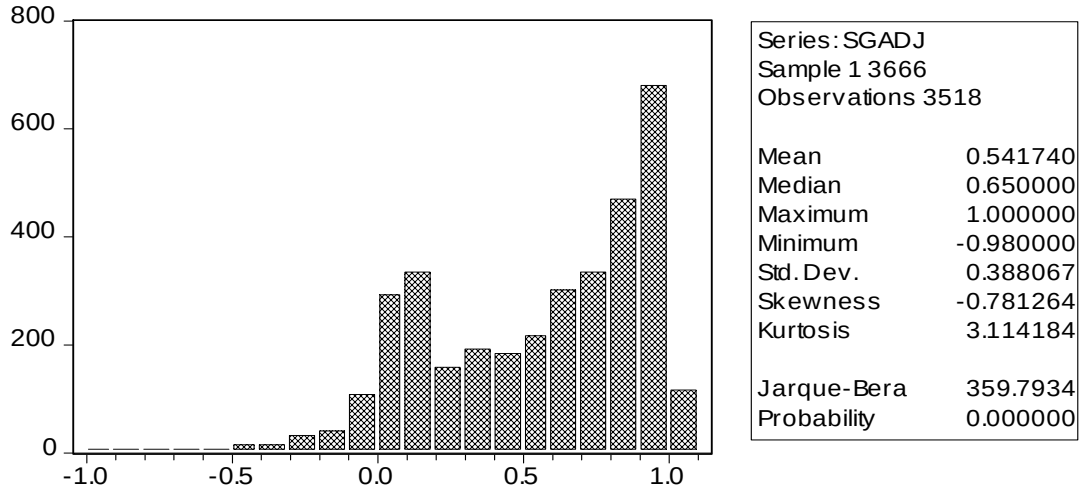

Figure 2: Actual and Predicted Health State Valuations for the Random Effects Model (2)

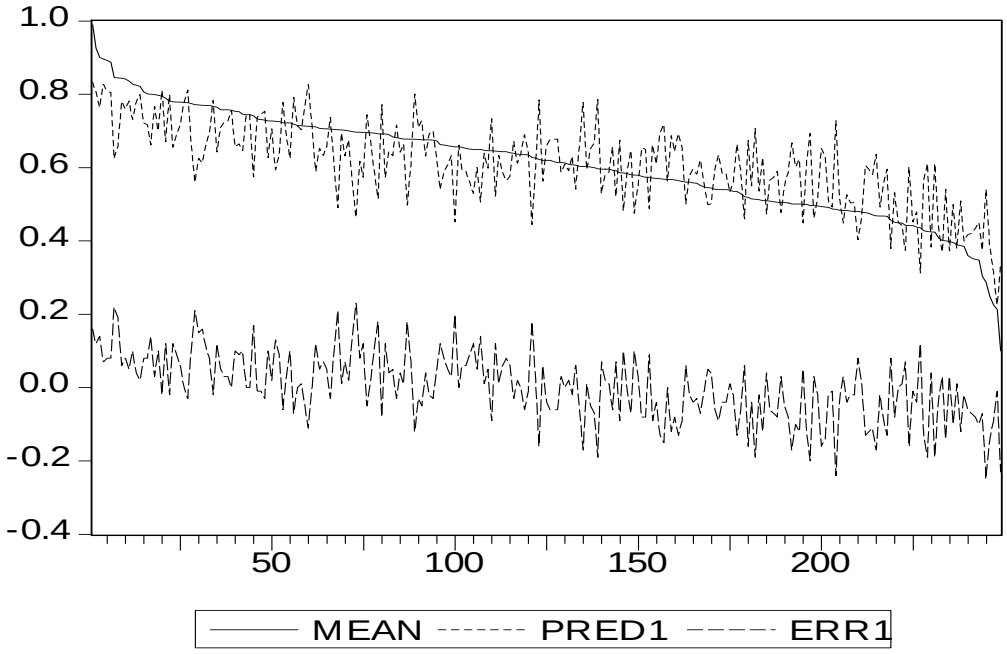

Figure 3: Actual and Predicted Health State Valuations for the Random Effects Model (5)

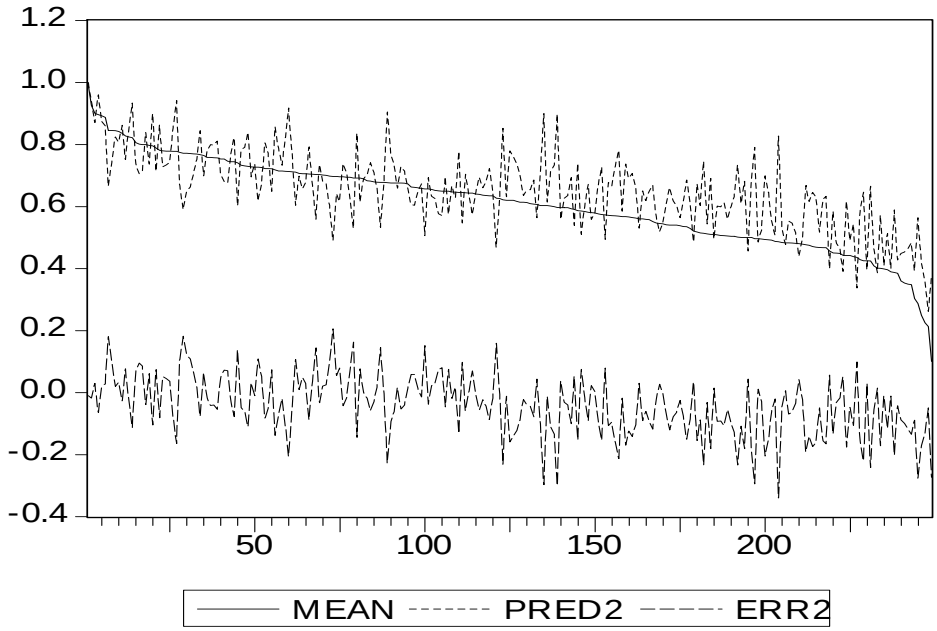

Table 1: The Short Form 6D

Level	Physical Functioning
1	Your health does not limit you in vigorous activities
2	Your health limits you a little in vigorous activities
3	Your health limits you a little in moderate activities
4	Your health limits you a lot in moderate activities
5	Your health limits you a little in bathing and dressing
6	Your health limits you a lot in bathing and dressing

Level	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
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

Level	Social functioning
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

Level	Pain
1	You have no pain
2	You have pain but it does not interfere with your normal work (both outside the home and housework)
3	You have pain that interferes with your normal work (both outside the home and housework) a little bit
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

Level	Mental health
1	You feel tense or downhearted and low none of the time
2	You feel tense or downhearted and low a little of the time
3	You feel tense or downhearted and low some of the time
4	You feel tense or downhearted and low most of the time
5	You feel tense or downhearted and low all of the time

Level	Vitality
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

Footnote: The SF-36 items used to construct the SF-6D are as follows: physical functioning items1, 2 and 10; role limitation due to physical problems item 3; role limitation due to emotional problems item 2; social functioning item 2; both bodily pain items; mental health items 1 (alternate version) and 4; and vitality item 2.

Table 2: A sample of health states defined by the SF-6D

Health state 111111

Your health does not limit you in **vigorous activities** (e.g. running, lifting heavy objects, participating in strenuous sports).

You have no problems with your work or other regular daily activities as a result of your **physical health or any emotional problems.**

Your health limits your **social activities** (like visiting friends or close relatives) a little or none of the time

You have no **pain**

You feel **tense or downhearted and low** a little or none of the time.

You have a lot of **energy** all of the time

Health state 223222

Your health limits you a little in **vigorous** activities (such as running, lifting heavy objects, participating in strenuous sport)

You are limited in the kind of work or other activities as a result of your **physical health**

Your health limits you in your **social activities** some of the times

You have **pain** but it does not interfere with your normal work (both work outside the home and housework)

You feel **tense or downhearted and low** a little of the time.

You have a lot of **energy** most of the time.

Health state 424334

Your health limits you a lot in **moderate** activities (such as moving a table, pushing a vaccum cleaner, bowling or playing golf)

You are limited in the kind of work or other activities as a result of your **physical health**

Your health limits you in your **social activities** most of the time

You have **pain** that interferes with your normal work (both outside the home and housework) a little bit.

You feel **tense or downhearted and low** some of the time.

You have a lot of **energy** a little of the time.

Health state 645655 ('pits')

Your health limits you a lot in **bathing and dressing yourself.**

You are limited in the kind of work or other activities as a result of your **physical health** and you accomplished less than you would like as a result of **emotional problems**

Your health limits your **social activities** all of the time

You have **pain** that interferes with your normal work (both outside the home and housework) extremely.

You feel **tense or downhearted and low** all of the time.

You have a lot of **energy** none of the time.

Table 3: Characteristics of included and excluded respondents

	Included n = 611	**Excluded n = 225**
Age: mean (s.d)	46 (18.1)	51 (19.6)
%		
female	61	56
married	53	48
with children < 16	28	21
renting property	28	34
in FT employment	37	32
Highest qualification		
degree	16	13
A levels	21	18
No qualifications	30	37
Found valuation task difficult[1]	4	5
Poor understanding of valuation task[2]	4	9

[1] judged by respondent
[2] judged by interviewer

Table 4: Descriptive Statistics for 50 SF-6D health state valuations

state	n	Min	Max	Mean	Median	s.d.
111111	13	0.92	1.00	0.99	1.00	0.02
111215	13	0.53	1.00	0.90	0.97	0.14
321221	11	0.57	0.98	0.84	0.89	0.13
122233	15	0.14	1.00	0.83	0.91	0.23
112221	11	0.51	0.98	0.82	0.89	0.17
221432	10	0.53	0.98	0.81	0.84	0.15
224223	14	0.53	1.00	0.80	0.85	0.17
532124	12	0.29	1.00	0.79	0.84	0.21
211111	12	0.19	1.00	0.78	0.90	0.27
221211	10	0.42	0.98	0.77	0.85	0.19
341123	11	0.10	0.99	0.76	0.92	0.31
241531	17	0.28	0.99	0.75	0.88	0.24
213323	13	0.12	0.98	0.74	0.79	0.25
222113	17	0.10	0.99	0.73	0.80	0.17
221212	15	0.05	0.98	0.72	0.85	0.33
112521	10	0.19	0.94	0.71	0.73	0.21
124314	10	0.06	0.99	0.70	0.94	0.35
541432	13	0.10	1.00	0.69	0.75	0.29
323333	9	0.05	0.98	0.68	0.76	0.32
443215	12	-0.06	1.00	0.67	0.81	0.35
342353	10	0.29	0.98	0.66	0.79	0.23
222121	11	0.05	1.00	0.65	0.75	0.34
345122	15	0.29	1.00	0.64	0.67	0.25
214535	12	0.00	0.99	0.63	0.78	0.37
413511	15	-0.24	0.99	0.62	0.75	0.39
523634	12	0.05	0.99	0.61	0.57	0.33
321455	10	0.10	0.99	0.60	0.65	0.33
424421	16	0.05	0.98	0.59	0.64	0.30
334254	13	-0.66	0.98	0.58	0.80	0.46
423433	10	-0.15	1.00	0.58	0.60	0.36
134322	17	0.10	1.00	0.57	0.59	0.27
315515	16	0.19	0.97	0.56	0.55	0.25
545122	13	0.10	0.98	0.55	0.61	0.31
432623	14	0.07	1.00	0.55	0.56	0.30
241635	17	-0.09	0.99	0.54	0.57	0.37
312552	14	0.10	0.95	0.53	0.64	0.35
344145	11	-0.57	0.98	0.51	0.63	0.48
412152	11	0.10	0.93	0.50	0.59	0.29
323443	12	-0.66	1.00	0.49	0.61	0.45
432255	12	0.00	1.00	0.48	0.48	0.42
325455	14	-0.19	0.91	0.47	0.54	0.36
431623	15	-0.88	0.99	0.45	0.67	0.47
423343	15	0.00	1.00	0.44	0.38	0.31
544352	11	-0.57	0.98	0.43	0.47	0.48
131542	19	-0.66	0.96	0.42	0.45	0.41
323644	10	0.10	0.99	0.40	0.29	0.31
141653	12	0.00	0.91	0.39	0.36	0.34
434654	16	-0.85	1.00	0.38	0.55	0.61
534644	11	-0.28	0.98	0.35	0.32	0.32
535645	8	-0.56	0.76	0.10	0.10	0.39

TABLE 5: Models with main effects

					Constant forced to unity	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	OLS	**RE**	**Mean**	**Median**	**RE**	**Mean**
c	**0.826**	**0.833**	**0.827**	**0.945**	**1.000**	**1.000**
PF2	-0.009	-0.021	-0.014	-0.011	**-0.058**	**-0.060**
PF3	0.008	-0.026	0.008	0.026	**-0.051**	-0.020
PF4	-0.036	**-0.065**	-0.027	0.001	**-0.088**	**-0.060**
PF5	-0.032	**-0.044**	**-0.043**	**-0.064**	**-0.061**	**-0.063**
PF6	**-0.115**	**-0.135**	**-0.096**	**-0.097**	**-0.160**	**-0.131**
RL2	-0.023	**-0.027**	-0.019	-0.026	**-0.056**	**-0.057**
RL3	**-0.035**	**-0.055**	**-0.043**	-0.035	**-0.076**	**-0.068**
RL4	**-0.034**	**-0.055**	**-0.036**	-0.026	**-0.078**	**-0.066**
SF2	-0.015	**-0.034**	-0.027	-0.029	**-0.066**	**-0.071**
SF3	**-0.041**	-0.022	**-0.049**	**-0.079**	**-0.048**	**-0.084**
SF4	**-0.047**	**-0.041**	**-0.057**	**-0.053**	**-0.066**	**-0.093**
SF5	**-0.085**	**-0.089**	**-0.073**	**-0.113**	**-0.109**	**-0.105**
PAIN2	0.011	-0.001	0.008	0.003	**-0.042**	**-0.048**
PAIN3	0.006	-0.018	-0.001	0.002	**-0.046**	-0.034
PAIN4	-0.034	**-0.026**	-0.032	-0.018	**-0.055**	**-0.070**
PAIN5	**-0.065**	**-0.068**	**-0.062**	**-0.102**	**-0.103**	**-0.107**
PAIN6	**-0.159**	**-0.155**	**-0.149**	**-0.191**	**-0.178**	**-0.181**
MH2	-0.033	-0.019	-0.026	**-0.058**	**-0.043**	**-0.057**
MH3	-0.025	**-0.032**	-0.022	**-0.043**	**-0.055**	**-0.051**
MH4	**-0.098**	**-0.093**	**-0.095**	**-0.133**	**-0.115**	**-0.121**
MH5	**-0.131**	**-0.106**	**-0.114**	**-0.165**	**-0.125**	**-0.140**
VIT2	**-0.043**	-0.006	-0.044	**-0.051**	**-0.040**	**-0.094**
VIT3	-0.036	-0.008	-0.037	-0.034	**-0.030**	**-0.069**
VIT4	-0.033	-0.011	-0.029	-0.048	**-0.040**	**-0.069**
VIT5	**-0.077**	**-0.068**	**-0.076**	**-0.090**	**-0.087**	**-0.106**
N	3518	3518	249	249	3518	249
adj R^2	0.204	0.200	0.583	0.577	#	0.508
inconsistencies	2	2	2	3	4	5
MAE	0.072	0.073	0.071	0.097	0.078	0.074
No > |0.05|	120	122	117	136	122	118
No > |0.10|	49	53	52	78	59	52
t(mean=0)	0.544	0.250	†	†	**-6.717**	†
JBPRED	0.376	1.178	0.737	1.725	2.461	0.681
LB	**333.01**	**386.63**	**520.71**	**560.88**	**185.3**	**169.57**

All models are estimated with White's heteroscedasticity consistent standard errors.
Estimates shown in **bold** are significant at $t_{0.10}$
no R^2 statistics (GEE estimation)
† Mean error is zero by definition.

TABLE 6: Models with interaction effects

			Constant forced to unity	
	(7)	**(8)**	**(9)**	**(10)**
	RE	**Mean**	**RE**	**mean**
c	**0.799**	**0.788**	**1.000**	**1.000**
PF2	-0.023	-0.015	**-0.050**	-0.053
PF3	-0.021	0.011	**-0.038**	-0.011
PF4	**-0.054**	-0.018	**-0.069**	**-0.040**
PF5	**-0.035**	-0.034	**-0.046**	**-0.054**
PF6	**-0.119**	**-0.084**	**-0.145**	**-0.111**
RL2	**-0.030**	-0.021	**-0.051**	**-0.053**
RL3	**-0.042**	-0.030	**-0.058**	**-0.055**
RL4	**-0.041**	-0.024	**-0.063**	**-0.050**
SF2	-0.030	-0.023	**-0.054**	**-0.055**
SF3	-0.012	**-0.040**	**-0.032**	**-0.067**
SF4	**-0.025**	**-0.042**	**-0.044**	**-0.070**
SF5	**-0.071**	**-0.058**	**-0.096**	**-0.087**
PAIN2	-0.005	0.005	**-0.037**	**-0.047**
PAIN3	-0.013	0.004	**-0.034**	-0.025
PAIN4	-0.020	-0.025	**-0.040**	**-0.056**
PAIN5	**-0.055**	**-0.049**	**-0.081**	**-0.091**
PAIN6	**-0.141**	**-0.136**	**-0.167**	**-0.167**
MH2	-0.022	-0.030	**-0.036**	**-0.049**
MH3	**-0.028**	-0.019	**-0.045**	**-0.042**
MH4	**-0.085**	**-0.089**	**-0.099**	**-0.109**
MH5	**-0.098**	**-0.109**	**-0.115**	**-0.128**
VIT2	-0.006	**-0.044**	**-0.032**	**-0.086**
VIT3	-0.002	-0.031	-0.019	**-0.061**
VIT4	-0.001	-0.019	-0.022	**-0.054**
VIT5	**-0.054**	**-0.064**	**-0.073**	**-0.091**
Most	**-0.052**	**-0.041**	**-0.084**	**-0.070**
Least	**0.049**	**0.048**		
n	3518	249	3518	249
adj R^2	0.201	0.591	#	0.526
inconsistencies	2	1	6	5
MAE	0.073	0.070	0.076	0.073
No > |0.05|	121	115	119	120
No > |0.10|	57	52	59	51
t(mean=0)	0.293	†	**-5.110**	-1.146
JBPRED	1.336	1.017	1.038	0.173
LB	**388.30**	**524.64**	**164.18**	**189.87**

All models are estimated with White's heteroscedasticity consistent standard errors.
Estimates shown in **bold** are significant at $t_{0.10}$
no R^2 statistics (GEE estimation)
† Mean error is zero by definition.