**Monograph:**
McCabe, C., Stevens, K., Roberts, J. et al. (2005) Health State Values for the HUI 2 descriptive system: results from a UK survey. Discussion Paper. Health Economics


# HEDS Discussion Paper 03/03



**Published paper**
McCabe C, Stevens K, Roberts J, Brazier JE. Health state values for the HUI2 descriptive system: results from a UK survey. *Health Economics* 2005;14(3):231-44.


ScHARR
School of Health and
Related Research

**The University of Sheffield**
**ScHARR**
**School of Health and Related Research**

# Sheffield Health Economics Group

**Discussion Paper Series**

**November 2003**

**Ref: 03/3**

# Health State Values for the HUI 2 descriptive system: results from a UK survey

Christopher McCabe, Katherine Stevens, Jennifer Roberts, John Brazier

Sheffield Health Economics Group

Corresponding Author:
Chris McCabe
Sheffield Health Economics Group
School of Health and Related Research
University of Sheffield
Regent Court, Sheffield, UK
S1 4DA
Email: C.McCabe@sheffield.ac.uk

## Abstract

This paper reports the results of a study to estimate a statistical health state valuation model for a revised version of the Health Utilities Index Mark 2, using Standard Gamble health state preference data. A sample of 51 health states were valued by a sample of the 198 members of the UK general population. Models are estimated for predicting health state valuations for all 8,000 states defined by the revised HUI2. The recommended model produces logical and significant coefficients for all levels of all dimensions in the HUI2. These coefficients appear to be robust across model specifications. This model performs well in predicting the observed health state values within the valuation sample and for a separate validation sample of health states. However, there are concerns over large prediction errors for two health states in the valuation sample. These problems must be balanced against concerns over the validity of using the VAS based health state valuation data of the original HUI2 valuation model.

## Introduction

The United Kingdom Paediatric Intensive Care Outcomes Study (PICOS) is using the Health Utilities Index Mark 2 to examine long term outcomes after paediatric intensive care in 23 centres in England, Scotland, Wales and Northern Ireland. Part of the study involves the construction of a UK valuation algorithm for the Health Utilities Index Mark 2 (HUI 2) health state classification system. This paper reports initial results of this work.

The HUI2 is the only preference based multi-attribute health related quality of life instrument specifically developed for use with children.[1] It consists of seven dimensions (sensation, mobility, emotion, cognition, self care, pain and fertility), each of which has between three and five levels. The levels describe a range, from 'normal functioning for age' to 'extreme disability'. (Appendix One gives the dimensions and level descriptions.)

The first applications of the system were in paediatric oncology. The fertility dimension was added to the original six dimensions proposed by Cadman and colleagues,[2] in order to capture side effects of chemotherapy. The developers state that by assuming fertility to be normal, the HUI 2 can be used as a generic health status instrument.[3] [4]

Preference based quality of life weights can be calculated for all 24,000 health states in the descriptive system using a multiplicative multi-attribute utility function

(MAUF) developed by Torrance and colleagues. This MAUF is based on health state valuation interviews with 194 parents of school age children in Hamilton, Ontario, Canada. The valuation interview followed the standard McMaster Valuation Framework, whereby each health state is valued using a Visual Analogue Scale (VAS) called the Feeling Thermometer. The value data obtained using the VAS are converted to utilities by means of a power curve transformation. The power curve transformation is estimated using the person mean values and utilities of four health states, which are valued using VAS and Standard Gamble methods. [4]

Three assumptions underlie the McMaster Valuation Framework: (i) VAS data capture individual's strength of preference, in terms of value, for different health states; (ii) individuals have a measurable value function which is related to their utility function by their relative risk attitude; (iii) the relative risk attitude is well described by a power curve relationship (i.e. a regression of the natural log of value on the natural log of utility).

There is now a significant literature that questions whether these assumptions hold in practice. Research by Nord,[5] Robinson et al., [6] and Morris and Durrand,[7] suggest that VAS data may provide ranking information but do not reflect strength of preference across health states. Schwarz [8] and Robinson et al. [9] report that, after adjustment is made for Range-Frequency context effects, VAS valuations appear to be consistent with the existence of a measurable value function. In addition a number of studies have reported that the power curve relationship does not hold at the individual level. [9] [10] [11] [12]

It is these concerns about the validity of VAS for collecting health state preference data, that motivated our aim to develop a statistical inference health state valuation algorithm using Standard Gamble data. This paper describes the valuation survey and the modelling results.

## Methods

Although the HUI2 consists of 7 dimensions, only six of these are required for the measurement and valuation of generic health status. As we are interested in using the instrument for generic health state valuation purposes the fertility dimension is excluded and health states are described using the six dimensions: sensation, mobility, emotion, cognition, self care and pain. The revised health state classification describes a total of 8,000 health states.

The study consists of two components; the valuation survey and the construction of a statistical model to predict health state values for all 8,000 health states defined by the classification system.

## Valuation Survey

## Selection of respondents

Following the recommendations of the Washington panel,[13] we aimed to achieve a representative sample of the UK general population. The sampling method aimed to achieve a sample that reflected the geographical distribution and socio-economic characteristics of the general population, in terms of age, sex, employment status and education. However, with the relatively small sample size, we were aware that our sample data could not be considered genuinely representative of UK population.

The target sample size for the valuation survey was 200, approximately equal to the 194 used in the original HUI 2 valuation exercise.[4] The sample was stratified by

mainland UK socio-economic region, on the basis of the proportion of population in each region according to the 1991 census. Age group, ethnicity, gender and socio-economic status were monitored through out the fieldwork in an attempt to balance the sample to reflect the UK population characteristics according to the 1991 census. The 1991 census was used because the 2001 census data were not available at the time of the valuation interviews.

## Selection of Health States

It is not feasible to value all the 8000 states in the descriptive system and there are no clear principles for identifying which health states to value for the construction of statistical inference models for health state preferences. Equally, we could find no guidance on how many observations are required for each health state nor how many health states any one individual should be asked to value. With a constrained sample size, a judgement has to be made, balancing the number of states valued and the number of observations per state; for example the SF-6D was estimated with an average of 15 observations per state, on a sample of 249 states.[16] We were also concerned to minimise the risk of respondent fatigue. The Measurement and Valuation of Health study asked respondents to undertaken 10 valuations tasks per interview. Using these valuation studies as guiding precedents we decided to ask respondents to complete a conservative eight standard gamble questions and aim for 25 respondents per health state.

The health states to be valued were identified from an orthogonal array constructed using the 'Orthoplan' module of SPSS 10 software. The minimum number of states

necessary to construct an orthogonal array for the six dimensions of the revised HUI2 was 25. The largest orthogonal array that met the constraints described above consisted of 49 cards. The full health state was selected in this orthogonal array, but as the value of full health is fixed in all the standard gamble valuation exercises and in the analyses, this state was not valued. For purposes other than the work reported here, two substitute states were identified and valued instead of perfect health. In addition all respondents valued the worst health state described by the HUI 2 (the PITS state). The 51 health states valued are listed in Table 1.

The health states in the orthogonal array were reviewed for plausibility in the light of the observation of the original HUI 2 valuation study that implausible states existed within the descriptive framework. For example, 'unable to control or use arms and legs' cannot plausibly be combined with 'eats, bathes, dresses and uses the toilet normally for age'). Where necessary, 'back off' states were substituted for the originals. We wished to keep the 'back off' states as close to the original states as possible, therefore, we constructed the 'back off' state by making the fewest dimension level changes required to create a plausible health state. Appendix 2 gives example health state descriptions from the valuation survey.

The backing off process does infringe the orthogonal design. However attempting to value implausible health states would have been unlikely to produce meaningful data for those specific states, and may have impacted upon the values obtained for other states in the same interview. It is not possible to state *a priori* to what degree the backing off process reduced the appropriateness of the sample of health states selected for valuation. If the estimated valuation models perform well across the range of

health states, this will indicate that the backing off process did not have a negative impact upon the sample of health states selected for valuation. The orthogonal array design means that the study was set up to measure main effects only. A separate valuation survey was undertaken to estimate a Multiplicative Multi-attribute utility function (M-MAUF) for the same health state classification. The M-MAUF explicitly models interactions. One of the objectives of the UK PICOS study was to compare the performance of the simple linear additive valuation model with that of a model that allowed for interactions. This paper reports the development of the simple linear additive model. The issue of interactions between dimensions will be discussed in more detail below.

## Interviews

Trained and experienced interviewers carried out all interviews. The interviews took place in the respondents' own home. Statistical Services Research Centre (SSRC), a professional survey and research group within the Sheffield Hallam University, employed the interviewers.

The interview consisted of four phases. In the first phase the purpose of the research was described to the respondent and consent to the interview was obtained. In the second phase the respondent was asked to rank 9 health states from the HUI2 classification system, plus immediate death. This allowed the individual to familiarise themselves with the descriptive system and with the task of comparing health states. The ranking data was also used to identify which version of the Standard Gamble question should be used in Phase 3; i.e for health states better or worse than death.

The third phase of the interview consisted of 8 standard gamble exercises. All respondents valued the PITS state And the seven remaining states were a sample from the states shown in Table 1. In order to ensure that all health states were valued a similar number of times, the health states were split in to 7 groups, each covering the range of functioning seen within the orthogonal array. Each interviewer was issued with seven envelopes, and instructed to work sequentially through the envelopes from 1 to 7 and then start again at envelope 1 until the sample was reached. The survey company monitored the returned scripts, to ensure that any differences in recruitment

rates between interviewers did not lead to significant imbalance in the number of valuations per state.

The study employed the version of the SG developed by the HUI2 development team at McMaster.[14] This version uses a prop called the Chance Board.

In the interview, the respondent is asked to choose between the certain prospect (A) of living in an intermediate health state defined by the HUI2 and the uncertain prospect (B) of two possible outcomes, the best (full health) state defined by the HUI2 and immediate death. The chances of the best outcome occurring is varied until the respondent is indifferent between the certain and the uncertain prospect. The respondent was asked to imagine that they were a child aged 10 years, and that they would expect to live for another 60 years.

For health states that were ranked as worse than death, the health state worse than death form of the SG question was used. In this the respondent is asked to choose between a certain prospect of immediate death and an uncertain prospect of perfect health (with probability p) and the health state (with probability 1-p). The health state value was calculated as –p.

The Chance Board prop adopts the 'ping-pong' approach to varying the probability; i.e. the respondent is asked to choose between the options with a very high probability of the best outcome occurring, before being asked to choose between the options when there is a very low probability of the best outcome occurring and so on, until the

respondent is indifferent between the uncertain and certain choice. The Chance Board uses 0.05 interactions in the probability of the best outcome. At all times the probabilities are displayed on a chance board, both numerically and in the form of a pie chart. The chance board is designed to lead the interviewer through a set of questions depending upon the interviewee's response to the previous question and thereby minimise the risk of interviewer variation.

The lead investigator (CM) received training in the use of this method from the McMaster team, who also produced the chance boards used in the survey. The lead investigator trained all interviewers.

For health states that were ranked better than immediate death in the ranking exercise the reference states were full health and immediate death. For health states that were ranked worse than death in the ranking exercise, respondents were asked to confirm, in the valuation exercise, that they still believed the health state to be worse than immediate death. Those that confirmed this view undertook a Standard Gamble question where Full Health and the Impaired health state were the reference states, and immediate death was the current state. The respondent was asked to identify what risk of the impaired health state they would be willing to accept in order to receive a therapy that avoided the state of immediate death and returned them to full health. This allowed us to calculate how much worse than death the impaired health state was felt to be.

The direct valuation of health states worse than death using SG represents a departure from the methods of the original HUI2 valuation survey. [4] In the original survey,

respondents who indicated that a health state was worse than immediate death were not asked to value the state and any respondents who valued 2 or more states as worse than death were excluded from the estimation sample.

The fourth phase of the interview consisted of a series of questions about the respondents' socio-economic circumstances. Finally the respondent was asked to rate how easy or difficult they had found each set of questions, on a five point scale. After the interview had been completed, the interviewer completed a brief assessment of the respondents understanding and effort.

## The Data

One hundred and ninety eight interviews were completed. Twenty three were excluded because they gave the same value for all health states or they valued the PITS state more highly than at least one other health state which had a higher level of functioning on all six dimensions. Whilst these exclusion criteria are consistent with other valuations studies, they represent an imposition of our expectations on the data; i.e we do not believe it is reasonable to value all health states the same and we do not believe it is reasonable to value the lowest health state in the classification system more highly than health states which have higher functioning on all dimensions. These criteria were applied to the full dataset before any models were estimated. The included and excluded respondents were different. Excluded people tended to be older, with fewer qualifications and were less likely to be in full-time employment. Compared to the UK population women and older people were over-represented in the sample of included respondents. Table 2 gives the socio-economic characteristics for the included and excluded respondents.

Descriptive statistics for the 51 health states valued are given in Table 1. Each state has been valued 24 times on average (minimum 9, maximum of 29). The PITS state was valued by 168 of the 175 respondents included in the analysis. The mean health state value ranged from –0.064 to +0.79. The median health state value normally exceeded the mean health state value. The mean and median health state values are consistent with logical orderings within the HUI2.

The individual data is bimodal, with concentrations of valuations around zero and 0.7. A histogram of the 1370 individual health state values included in the valuation dataset is shown in Figure 1. There were relatively few negative valuations (7%), and also very few health state valuations at 0.95 or greater (4%). This suggests that individuals were willing to accept a significant risk of death to achieve full health.

## Modelling

The aim of the project is to construct a model for predicting health state values for all 8,000 health states described by the revised HUI2 health state classification system.

The basic model structure for a statistical inference health state valuation model is:

$$Uij = g(\beta xij + \theta rij\ ) + \varepsilon ij$$

Where:

i = 1,2,….,n represents individual health states in the descriptive framework;

j= 1,2,…..,m represents individual respondents in the health state valuation survey;

Uij = the standard gamble value for health state i valued by respondent j;

g = appropriate functional form;

x = a vector of dummy variables for each level on each dimension of the health state classification.

r = is a vector of interaction terms between the levels of different attributes;

εij = the error term.

The vector x contains 21 terms. Level 1 on each dimension acts as the baseline for that dimension. The dummy variables take on a value of 1 for a health state that includes the dimension level and 0 otherwise. For a simple linear model, the intercept represents the estimated value of the full health state 1,1,1,1,1,1 and the value for all other health states are derived by summing the coefficients of the appropriate 'on' dummies.

In the preferred models the value of the full health state is restricted to equal unity. This is consistent with the theoretical construct within which the models are estimated.[16]

The orthogonal design of the survey allows the study of main effects only. However, it is standard practice in health state valuation modelling to look for interactions between the dimensions.[15][16] For completeness we looked for evidence of interactions in this dataset. The number of possible interactions within the HUI2 classification system is very large. Modelling all of them would require valuation data on a much larger sample of health states than the one available. There would be some risk of finding statistically significant interactions due to the play of chance. Therefore, the choice of interactions considered in the modelling is based on the type of interactions which other researchers have found to be significant [15] [16] and interactions within the HUI2 classification that have been acknowledged in the literature.[3] Specifically we tested for first order interactions between the mobility and self-care dimensions of the descriptive system; the presence of the lowest level of

functioning on at least one dimension; and the presence of the highest level of functioning on at least one dimension.

## Model specifications

The choice of the appropriate model specification depends upon the characteristics of the data. Standard OLS regression assumes the standard zero mean, constant variance error structure, with independent error terms, i.e. cov(εij,εi'j)=0, i≠i`. This assumption means that the 1370 observations from 175 respondents are treated as though 1370 respondents provided them.

The Random Effects (RE) model acknowledges that the error term may not be independent of the respondent, and therefore separates out within and between respondent error terms.

$$\varepsilon ij = uj + eij$$

Where;

uj = the respondent specific variations. This is assumed to be random across individuals; and

eij = the error term for the ith health state valuation of the jth individual, which is assumed to be random across observations.

This model also assumes that the allocation of health states to respondents is random i.e. cov(uj, eij)=0.

A third specification is a fixed effects model, which also recognises that the importance of individual effects but instead of assuming that these are random, the respondent specific variation is estimated along with coefficients on the explanatory variables.

Hausman and Breusch-Pagan tests are employed to test the importance of individual effects and the appropriate specification. We used the Ramsey RESET test to for model misspecification.

A number of transformations of the dependent variable, including a tobit model and a log-log transformation, were attempted in order to alleviate the problem of left skew in the data. However, none of these outperformed the models reported below so they are not discussed here.

For completeness we also estimated aggregate level mean and median health state value models. While these models do not make the most efficient use of the data, it may be argued that they utilise the information that is of most interest to policy makers, i.e. a central estimate of the value of each health state. The literature does not provide any clear principles for choosing between mean, median and individual models. The choice between mean and median health state valuations is a question of how preferences should be aggregated for public decision making. In the democratic paradigm all preferences should count equally, thus the median model would seem to be more appropriate. By contrast, in the welfare economics paradigm, the strength of preference is important and therefore the mean model is more appealing.[17]

## Results

Table 3 reports the OLS, Random Effects and mean models where the intercept term is restricted to unity. Unrestricted models were estimated but their performance was inferior and they are theoretically difficult to justify so they are not reported, here. The Breusch-Pagan test suggests that individual effects are present in the data ($\chi$=1261.62, p=0.000) and the Hausman test suggests that random (as opposed to fixed) effects are preferred ($\chi$=6.05, p=0.99).

All three models perform reasonably well. All the coefficients in the OLS model are significant, with only one inconsistency; (mobility 4 has a 0.0205 higher decrement than mobility 5.) and the explanatory power is 0.77. Similarly, all the coefficients in the random effects model are significant. There are two inconsistencies, neither of which is large; one between mobility level 4 and 5 and one between pain levels 2 and 3 .

The mean model has two inconsistencies (mobility levels 4 and 5; emotion levels 4 and 5) and explanatory power of 0.97. The median was inferior to the mean model and it not reported here.

## Predictive performance

Within sample predictive performance is examined by identifying the proportion of states that were predicted within 0.1 (absolute value) of the observed mean value. (See the bottom of Table 3). We also examined the Root Mean Square Error, the

Mean Absolute Error and looked at autocorrelation in the prediction errors using a Ljung-Box test.[18]

The OLS and mean models perform the best in this regard, predicting 94% and 92% to within +/-0.1 respectively. The RE model is not as good at 84%. These figures compare well with the performance of other instruments.[16 15] None of the models display autocorrelation in the within sample prediction errors. However the RE model appears to give biased predictions as shown by the t-test of the null hypothesis that the mean prediction error is zero. Whilst the R-squared statistics are high compared to both the EQ-5D and SF-6D models, it is important remember that these models are estimated without constants and therefore such comparisons have limited meaning.

Fifteen health states from the revised HUI2 descriptive system, which were not in the orthogonal array used in the main valuation survey, were valued using identical methods, in a separate valuation survey (n=51, mean of 25 observations per health state). To identify these states we constructed a new orthogonal array and selected the first 14 states not in the valuation sample. As with the valuation survey, all respondents valued the PITS states. These data were used to assess out of sample predictive performance.

The random effects model predicts 100% of the states to within +/- 0.1; the OLS model 93% and the mean model 87%.

## Interactions

We estimated models with interactions for the presence of the lowest or the highest level of functioning on any of the six dimensions. These variables were not significant in any of the models.

We also estimated models with first order interaction terms for mobility (levels 2,3,4 and 5) and self care (levels 2,3 and 4); i.e.12 interaction terms in total. The majority of the interaction terms for the mobility and self-care did not reach statistical significance and a number of them did not have the expected sign. In addition, the number of inconsistencies increased markedly in these models, whilst the explanatory power was not improved.

## Discussion

The Health Utilities Index questionnaire is one of the most widely used health state classification systems. It has been applied in a wide range of contexts, including population health assessment, economic evaluations alongside clinical trials and decision analytic modelling.[19]

The methods used for estimating the quality of life weights of different health states within the HUI descriptive systems have been the subject of much debate, raising substantial concerns that the weights may not reflect population preferences for health. In particular, there is considerable doubt that VAS data capture health state preferences, and that they are related to utilities by a simple power curve transformation.[5] [6] [9] [10] [11]

In this paper we report the results of a valuation survey and modelling project that has used direct utility values obtained through Standard Gamble interviews. We have examined a number of alternative models for predicting health state values for the HUI2 classification.

Using the assessment criteria described above, the models estimated with the constant forced to unity perform better than those estimated with an unrestricted constant. No other transformations of the data lead to improved models. Of the restricted models, those estimated on the individual data are superior to those estimated using either the mean or median data. This is perhaps unsurprising given the small number of observations in these datasets. Whilst the mean model performs well on many of the

assessment criteria, the non-significant co-efficient on mobility level 5, and value of the same co-efficient (-0.098) lacks face validity when mobility level 4 is given a co-efficient of –0.140. This is not reflected in the assessment of predictive performance, as there is only one health state in either the valuation or validation sample that includes mobility level 5. This highlights the importance of the health states selected to be in the validation dataset. With hindsight, it might have been desirable to include much milder states and much more severe states than those in the valuation dataset.

On balance the OLS model is slightly superior to the RE model, with fewer inconsistencies and superior predictive performance. The OLS model is preferred despite the importance of individual effects suggested by the Breusch-Pagan test. Statistical theory dictates that both OLS and RE estimators are unbiased and consistent, but that the RE estimator is more efficient in these circumstances. Given the sample size these efficiency gains are not expected to be great and are outweighed by the biased predictions generated by the RE model. Our choice of the OLS model is a judgement on our part. We are giving greater weight to the significance of the coefficients and the predictive performance, than to the test for individual effects.

There remain some areas for concern, notably the inconsistency in the coefficients between mobility level 4 and mobility level 5. The apparent inconsistency may not be serious, as the difference between the two coefficients is not statistically significant. It may be that respondents saw little meaningful difference between level 4; “Requires the help of another person to walk or get around and requires mechanical equipment” and Level 5; “Unable to control or use arms or legs”; in both circumstances an individual is completely dependent upon others to move around.

The poor prediction of two of the health states from the estimation sample remains a cause for concern. It may be that health state (3,1,3,3,3,1) was difficult for respondents to value as it includes level one functioning on mobility; "Able to walk, bend, lift, jump and run normally for age" and level 3 self care; "Requires mechanical equipment to eat, bathe, dress, or use the toilet independently". However, the health state (2,1,3,3,2,1) does not appear to be implausible and no convincing explanation for the poor prediction has been identified.

Unlike the EQ-5D, in which the dimensions levels are unambiguously different, some dimension levels within the HUI2 classification system could be seen as equivalent. Indeed plausible arguments can be made for the reversal of certain dimension orderings. For example, level 4 of self care may be preferred to level 3, if the respondent assumes that the description 'requires help' means that help will be provided. It may be that classification systems with more dimension levels, such as the HUI2 and SF-6D, run a greater risk of this type of ambiguity as they attempt to define smaller decrements in health. However, simpler descriptive systems such as the EQ-5D run the risk of grouping together health states that people would value as very different, if the descriptive system could differentiate them. There is clearly a balance to be struck between sensitivity and ambiguity in the descriptive systems.

None of the models with interaction terms represented an improvement on the main effects models presented in Table 3. However, this is unsurprising given the design of the valuation study. There is some risk that if interactions are present in individual's underlying utility function then these are being 'loaded' on to the main effects. We

cannot test for this with the existing data. We would suggest that good predictive performance of the model indicates that such loading, if present, is small. This said, future work focusing on the potential role of interactions in preferences within the HUI2 health state classification system is warranted. Some insight into the importance of such interactions may be obtained from the companion study to the work presented here, which is designed to construct a multiplicative multi-attribute utility function for the revised HUI2 classification.

The valuation dataset was small in terms of both the number of health states values and the number of respondents. The EQ-5D health state value model was estimated on a sample of 3,325 respondents, valuing 17.3% of the health states in the descriptive system. The SF-6D was estimated on a sample of 611 respondents valuing 249 health states 1.4% of the possible health states. This study used 176 respondents valuing 0.64% of the possible health states. It seems plausible that a larger dataset, both in terms of the number of observations per health state and the number of health states valued, could provide an improved predictive model. However, finding non-systematic and unbiased prediction errors may be considered a good performance from such a limited dataset.

The range of mean health state values in the estimation dataset is another potential cause for concern. The highest valued health state has a mean value of 0.79. It would have been desirable to have more highly valued health states in the estimation sample. Further work must include the valuation of less severe health states. For comparative purposes it is useful to note that the original MAUF uses measured person mean utility data in the range 0.51 to 0.88, to estimate the VAS-SG power curve; and

calculated utility data in the range 0.46 to 0.89 to estimate the MAUF. The functional form of both our preferred model and the McMaster MAUF plays a significant role in the predicted values for health states in certain ranges of the classification.

In common with other population health state valuation studies, we have not reported models with parameters for socio-economic characteristics. The socio-economic data collected as part of the interview are not necessarily the criteria that one would wish to give explicit weight to, in modelling population health state preferences. Indeed, they were chosen for comparability with other health state valuations studies, and not for their importance in social health state preferences. The incorporation of socio-economic characteristics into health state preference modelling is an important issue, but it is not the focus of the research reported here, nor was the study designed to inform such research.

The form of the Standard Gamble question used in both the original HUI2 valuation survey, and the surveys reported in this paper has been criticised. The question asks the respondent to assume that they are ten years old, and will spend 60 years in the final health state. It is unclear from the questions whether the respondent is meant to attempt to adopt the attitudes of a ten year old child or apply their own attitudes to choices which will have extremely long term effects. The data is not available to say with any confidence what the respondents in this survey did. Detailed follow-up interviews examining their thought processes whilst highly desirable were not possible within the available resources.

The variant of the standard gamble question used in the surveys reported here and the original HUI2 valuation survey, included the chance board prop.[14] Dolan and Sutton showed that the variant of the technique used to elicit health state preferences can have a greater impact upon the values obtained than the primary technique.[11] We have no data to assess whether a different variant of the SG technique would have obtained similar data. However, this issue does not affect comparisons between our valuation and validation survey, and the original HUI2 valuation survey.

In common with other health state valuation exercises [15][16] we have excluded data from respondents who provided the same value for all health states or who valued the PITS state higher than another state. The rationale for this action being that both of these reflect a failure to understand the nature It can be argued that this imposes the analyst's preferences upon the dataset. Future work should consider feeding back the responses to the individual to obtain confirmation that these reflect their preferences.

The preferred model stands comparison with those developed for the recently published SF-6D [15] [16] and the EQ-5D. It is applied to a potentially richer descriptive framework than the EQ-5D, 8000 states compared to 243, and one which can describe more severe health states than the SF-6D. Also, whilst the range of health state values does span zero,[16] the value attached the worst health state is not as extreme as in the EQ-5D.[15] The valuation model avoids the concerns associated with the utilisation of VAS data for collecting health state preferences. Whether these potential advantages are realised can only be assessed through head-to-head comparisons of the instruments. Fortunately the new algorithm can be applied to

existing Health Utilities Index data and it should be possible to undertake these head-to-head studies soon.

Pragmatically, the value of this alternative algorithm depends upon whether the current concerns about VAS health state data are resolved in its favour or not. Other considerations include whether the new algorithm performs better in predicting health state valuations than (a) the existing HUI2 multi-attribute utility function, (MAUF) and (b) a multi-attribute utility function based on a UK valuation survey of the revised HUI2 classification system, utilising the McMaster Valuation Framework. The valuation survey for the latter piece of work has been completed and the MAUF has been completed and is being prepared for publication. The external validation survey utilised in this paper will be used to compare the predictive performance of the two valuation algorithms. Although the analysis has not been presented in this paper, we can report that the Mean Absolute Error for the McMaster MAUF predictions for the 51 health states in the valuation dataset is >0.1.

This work has assumed that standard gamble data measure utilities. Whilst recognising the debate around the relative strengths and weaknesses of alternative valuations techniques, it is not the purpose of this paper to enter into this debate. The current consensus suggests that whilst there are (different) reasons to doubt whether the Standard Gamble or Time Trade Off techniques provide unbiased measures of preferences over health, it is not possible to identify either as the superior technique. [17] [20] Given the context, where data from SG, TTO and VAS based valuation exercises are being used to inform health care resource allocation decisions, we

believe the use of Standard Gamble data to explore possibilities of improving the data used in these decisions is defensible.

## Summary

This study has demonstrated that it is possible to estimate a health state valuation model for the revised HUI2 descriptive system, using direct health state preference data, obtained using the standard gamble technique. Our preferred model for the estimation of health state utilities for the revised HUI 2 descriptive system is the OLS main effects model with the constant restricted to unity. Given the current concerns over the role of VAS data in the measurement of health state preferences, this model represents an alternative source of preference based quality of life weights, when health status data has been collected using the Health Utilities Index questionnaire.

## Acknowledgements

We would like to thank the UK Medical Research Council for funding the UK PICOS study, of which this is one part. In addition, we would like to thank the staff at SSRC, Sheffield Hallam University and, of course, the individuals who agreed to be interviewed for this study. We would also like to thank two anonymous referees for their valuable and constructive comments. The usual disclaimer applies.

## Table 1: Descriptive statistics for health states in the valuation sample

| | Mean | Minimum | Maximum | State | Mean | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| **2,2,1,2,2,1** | .34 | -.75 | .98 | **4,2,3,1,2,2** | .73 | .08 | .98 |
| **1,4,1,3,4,1** | .43 | .03 | .98 | **1,2,4,1,3,4** | .79 | .40 | .98 |
| **1,1,2,2,2,2** | .50 | -.03 | .93 | **4,1,2,4,3,1** | .41 | -.93 | .98 |
| **1,1,2,1,2,3** | .57 | -.03 | .98 | **4,3,1,3,2,2** | .39 | -.65 | .98 |
| **3,1,3,3,3,1** | .77 | .03 | .98 | **2,1,3,3,2,1** | .54 | .03 | .95 |
| **1,1,1,2,3,2** | .42 | .03 | .93 | **3,2,2,4,1,2** | .41 | .03 | .98 |
| **3,2,2,2,2,1** | .46 | .03 | .98 | **3,3,1,2,3,3** | .52 | .03 | .93 |
| **1,2,1,1,3,2** | .71 | .03 | .98 | **3,2,1,3,4,5** | .61 | .03 | .98 |
| **1,3,3,2,1,3** | .62 | .03 | .98 | **2,2,2,3,3,3** | .64 | .03 | .98 |
| **1,2,5,2,1,1** | .64 | .08 | .98 | **2,5,5,3,3,2** | .54 | -.15 | .98 |
| **1,4,2,3,1,1** | .27 | -.93 | .93 | **2,1,4,2,4,2** | .64 | .03 | .98 |
| **2,2,1,2,1,4** | .43 | .03 | .93 | **3,1,4,4,3,1** | .46 | .03 | .98 |
| **1,2,2,2,2,2** | .39 | -.65 | .85 | **3,1,5,3,1,2** | .61 | -.03 | .98 |
| **2,3,4,1,1,1** | .33 | -.65 | .85 | **1,3,2,3,3,2** | .25 | -.93 | .85 |
| **3,3,1,1,3,1** | .52 | .03 | .98 | **3,4,4,2,2,2** | .48 | .03 | .98 |
| **2,3,5,1,2,1** | .44 | -.35 | .93 | **2,2,3,2,3,5** | .19 | -.65 | .85 |
| **2,2,2,1,4,2** | .43 | -.93 | .98 | **4,2,4,3,1,3** | .25 | -.75 | .98 |
| **3,4,2,1,2,4** | .41 | -.08 | .93 | **4,5,2,2,4,1** | .51 | .03 | .98 |
| **2,3,1,4,1,2** | .34 | -.93 | .85 | **1,3,3,4,4,4** | .34 | -.85 | .98 |
| **3,4,3,1,1,2** | .38 | -.15 | .98 | **2,4,1,4,2,3** | .47 | .03 | .93 |
| **3,1,1,3,2,4** | .51 | -.90 | .98 | **3,3,2,2,2,5** | .63 | .03 | .93 |
| **2,1,2,3,1,4** | .50 | -.65 | .98 | **1,2,5,4,2,5** | .40 | .03 | .93 |
| **3,2,3,3,3,1** | .47 | .03 | .98 | **2,4,2,1,3,5** | .60 | .03 | .98 |
| **4,2,1,1,1,4** | .46 | -.93 | .98 | **1,4,4,3,2,5** | .62 | .03 | .93 |
| **3,1,5,1,4,3** | .44 | -.40 | .93 | **4,4,5,2,3,4** | .70 | .35 | .93 |
| | | | | **4,5,5,4,4,5** | -.07 | -.98 | .85 |

## Table 2: Characteristics of excluded and included respondents

| | **Included (n=175)** | **Excluded (n=23)** |
|---|---|---|
| Age (mean s.d.) | 53 (15) | 62 (15) |
| *%* | | |
| Female | 72.7 | 60.8 |
| Married | 69.9 | 59.1 |
| With children <16 | 17.8 | 13.6 |
| Renting property | 9.7 | 27.3 |
| In FT employment | 57.7 | 77.3 |
| Highest Qualification | | |
| Degree | 20.5 | 4.5 |
| GCSE | 22.7 | 16.5 |
| No qualifications | 22.2 | 18.2 |
| | | |
| Found valuation task difficult[1] | 9.1 | 9.1 |
| Poor understanding of valuation task[2] | 1.7 | 4.5 |

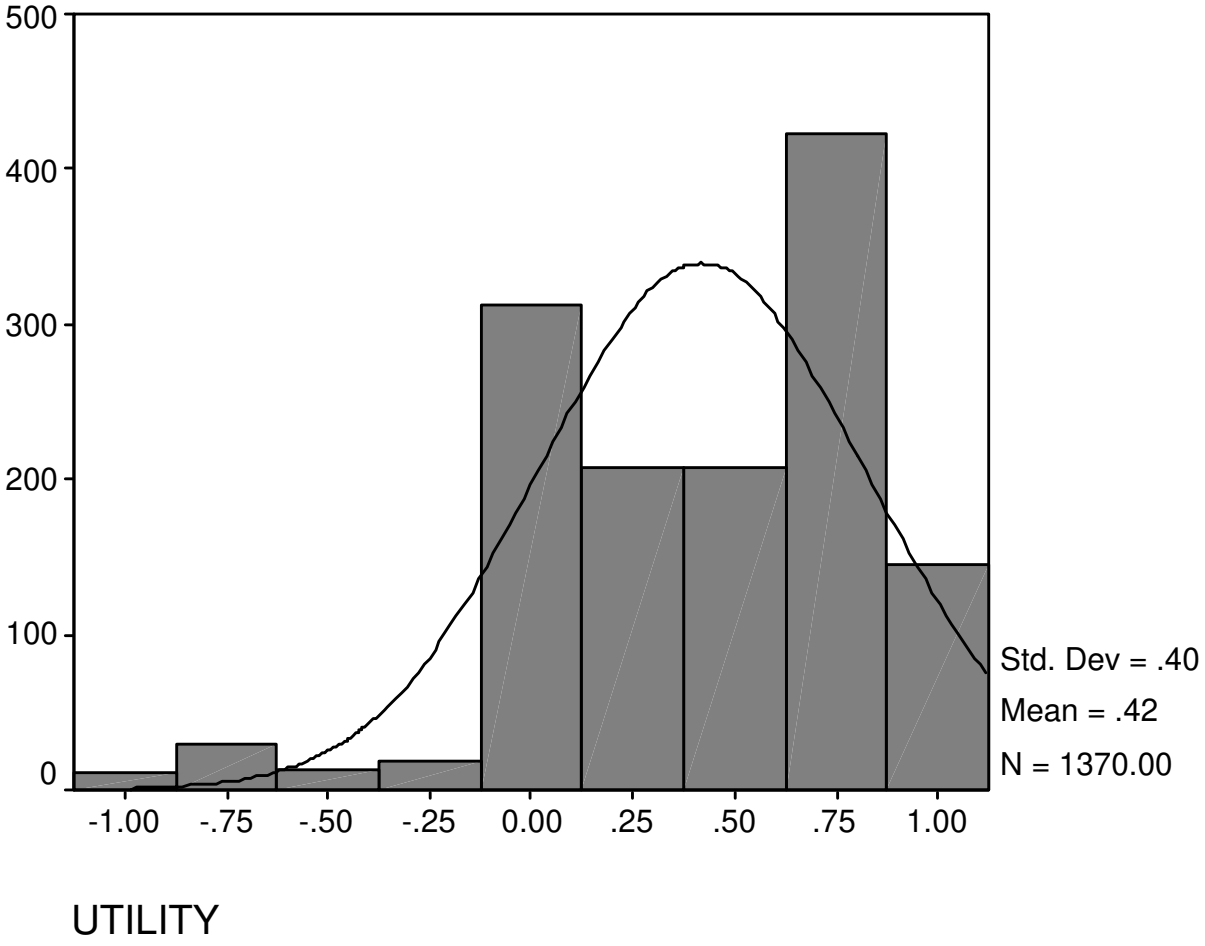
[1] judged by respondent, [2] judged by interviewer

## Table 3: Main Effects Valuation Models

| | Model Number | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| | OLS | RE | Mean |
| C | **1** | **1** | **1** |
| Sens2 | **-0.114** | **-0.114** | **-0.115** |
| Sens3 | **-0.123** | **-0.123** | **-0.120** |
| Sens4 | **-0.225** | **-0.225** | **-0.227** |
| Mobil2 | **-0.051** | **-0.051** | **-0.057** |
| Mobil3 | **-0.122** | **-0.122** | **-0.129** |
| Mobil4 | **-0.131** | **-0.131** | **-0.140** |
| Mobil5 | **-0.113** | **-0.113** | -0.098 |
| Emot2 | **-0.094** | **-0.094** | **-0.095** |
| Emot3 | **-0.112** | **-0.112** | **-0.100** |
| Emot4 | **-0.181** | **-0.181** | **-0.179** |
| Emot5 | **-0.184** | **-0.184** | **-0.177** |
| Cogn2 | **-0.055** | **-0.055** | **-0.052** |
| Cogn3 | **-0.096** | **-0.096** | **-0.092** |
| Cogn4 | **-0.168** | **-0.168** | **-0.163** |
| S_care2 | **-0.052** | **-0.052** | **-0.055** |
| S_care3 | **-0.114** | **-0.114** | **-0.111** |
| S_care4 | **-0.117** | **-0.117** | **-0.122** |
| Pain2 | **-0.110** | **-0.110** | **-0.109** |
| Pain3 | **-0.116** | **-0.116** | **-0.116** |
| Pain4 | **-0.161** | **-0.161** | **-0.163** |
| Pain5 | **-0.255** | **-0.255** | **-0.248** |
| N | 1370 | 1370 | 51 |
| (Adj) $R^2$ | 0.77 | n/a | 0.99 |
| Incons. | 1 | 2 | 2 |
| No > +/- 0.1 Within sample (n=51) | 2 | 5 | 2 |
| No > +/- 0.1 validation sample (n=15) | 1 | 0 | 1 |
| RMSE | 0.059 | 0.066 | 0.059 |
| MAE | 0.047 | 0.053 | 0.048 |
| t-test | -0.157 | **-2.775** | 0.493 |
| RESET test (F-test) | 2.93 | n/a | 2.71 |
| LB | 6.364 | 5.04 | 4.526 |

Estimates shown in bold are significant at the t=0.1, models are estimated with heteroskedasticity consistent standard errors.

# Figure 1: Health state utility values



## Appendix 1: Dimension and Level Descriptions for the Health Utilities Index Mark 2

| Dimension &Levels | Description | Dimension &Levels | Description |
|---|---|---|---|
| **Sensation** Level 1 | Able to see, hear and speak normally for age | **Self Care** Level 1 | Eats, bathes, dresses and uses the toilet normally for age |
| Level 2 | Requires equipment to see or hear or speak | Level 2 | Eats, bathes, dresses or uses the toilet independently with difficulty |
| Level 3 | Sees, hears, or speaks with limitations even with equipment | Level 3 | Requires mechanical equipment to eat, bathe, dress, or use the toilet independently |
| Level 4 | Blind, deaf, or mute | Level 4 | Requires the help of another person to eat, bathe, dress or use the toilet |
| **Mobility** Level 1 | Able to walk, bend, lift, jump and run normally for age | **Cognition** Level 1 | Learns and remembers schoolwork normally for age |
| Level 2 | Walks, bends, lifts, jumps or runs with difficulty but does not require help | Level 2 | Learns and remembers schoolwork more slowly than classmates as judged by parents and/or teachers |
| Level 3 | Requires mechanical equipment (such as canes, crutches, braces or a wheelchair) to walk or get around independently | Level 3 | Learns and remembers very slowly and usually requires special educational assistance |
| Level 4 | Requires the help of another person to walk or get around and requires mechanical equipment | Level 4 | Unable to learn and remember |
| Level 5 | Unable to control or use arms or legs | **Pain** Level 1 | Free of pain and discomfort |
| **Emotion** Level 1 | Generally happy and free from worry | Level 2 | Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities |
| Level 2 | Occasionally fretful, angry, irritable, anxious depressed or suffering from “night terrors” | Level 3 | Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities |
| Level 3 | Often fretful, angry, irritable, anxious depressed or suffering from “night terrors” | Level 4 | Frequent pain. Frequent disruption of normal activities. Discomfort requires prescription narcotics for relief |
| Level 4 | Almost always fretful, angry, irritable, anxious, depressed | Level 5 | Severe pain. Pain not relieved by drugs and constantly disrupts normal activities. |
| Level 5 | Extremely fretful, angry, irritable, anxious or depressed usually requiring hospitalisation usually requiring hospitalisation or psychiatric institutional care | **Fertility** Level 1 | Able to have children with a fertile spouse |
| | | Level 2 | Difficulty in having children with a fertile spouse |
| | | Level 3 | Unable to have children with a fertile spouse |

## Appendix 2: Example health state descriptions from valuation survey

Able to see, hear, and speak normally for age

Walks, bends, lifts, jumps, or runs with some limitations but does not require help

Extremely fretful, angry, irritable, anxious, or depressed
Usually requiring hospitalisation or psychiatric institutional care

Unable to learn and remember

Eats, bathes, dresses, or uses the toilet independently with difficulty

Severe pain. Pain not relieved by drugs and constantly disrupts normal activities. **STATE**
**1,2,5,4,2,5**

Requires equipment to see, or hear, or speak

Walks, bends, lifts, jumps or runs with some limitations but does not require help

Often fretful, angry, irritable, anxious, depressed or suffering "night terrors"

Learns and remembers school work more slowly than classmates as judged by parents and/or teachers

Requires mechanical equipment to eat, bathe, dress, or use the toilet independently

Severe pain. Pain not relieved by drugs and constantly disrupts normal activities. **STATE**
**2,2,3,2,3,5**

# REFERENCES


[1] Feeny D. Furlong W. Barr R.D. Torrance G.W. Rosenbaum P. Weitzman S. A comprehensive multi-attribute system for classifying the health status of survivors of childhood cancer Journal of Clinical Oncology 1992; 10(6):923-928

[2] Cadman D. Goldsmith C. Torrance G. Boyle M. Furlong W. Development of a health status index for Ontario Children: Final Report to the Ontario Ministry of Health Grant Research DM648 (00633) McMaster University Hamilton Ontario December 1986

[3] Torrance G.W. Furlong W. Feeny D. et al. Multi-attribute preference functions: health utilities index. Pharmacoeconomics 1995;7(6):503-520

[4] Torrance G.W. Feeny D.H. Furlong W.J. Barr R.D. Zhang Y. Wang Q. A. Multi-attribute utility function for a comprehensive health status classification system: Health Utilities Mark 2. Medical Care 1996;34(7):702-722

[5] Nord E. Validity of the Visual analogue scale in determining social utility weights for health states International Journal of Health Planning and Management 6:234-242

[6] Robinson A. Dolan P. Williams A. Valuing health states using VAS and TTO; what lies behind the numbers? Soc Sci Med 1997;45:1289-97

[7] Morris J. Durand A. Category rating methods: numerical and verbal scales 1989 University of York: Centre for Health Economics

[8] Schwartz A. Rating scales in context Med Decision Making 1998; 18:236

[9] Robinson A. Loomes G. Jones-Lee M. Visual analogue scales, standard gambles and relative risk aversion Med Dec Making 2001; 21:17-27

[10] Bleichrodt H. Johanneson M. An experimental test of a theoretical foundation for rating-scale valuations Med Dec Making 1997;17:208-216

[11] Dolan P. Sutton M. Mapping visual analogue scale scores on to time trade off and standard gamble utilities. Soc Sci Med 1997; 44(10):1519-1530

[12] van Busschback J. The validity of QALYs PhD Thesis; (Erasmus University, Rotterdam) 1994 cited in Dolan P. The measurement of health related quality of life for use in resource allocation decisions in health care in Culyer A.J. Newhouse J.P. (eds) Handbook of Health Economics Elsevier Science BV 2000

[13] Gold M.R. Siegel J.E. Russell L.B. Weinstein M.C. Cost effectiveness in Health Care and Medicine Oxford University Press 1996 Oxford

[14] Furlong W. Feeny D. Torrance G.W. Barr R. Horsman J. Guide to design and development of health state utility instrumentation CHEPA Working Paper #90-9 McMaster University Hamilton, Ontario 1990

[15] Dolan P. Modelling valuations for EuroQol Health States. Medical Care 1997;35(11):1095-1108

[16] Brazier J. Roberts J. Deverill M. The estimation of a preference based measure of health from the SF-36. Journal of Health Economics, 2002; 21(2):271-292

[17] Dolan P. The measurement of health related quality of life for use in resource allocation decisions in health care in Culyer A.J. Newhouse J.P. Handbook of Health Economics Elsevier Science BV 2000

[18] Ljung G. and Box G. On a measure of lack of fit in time series models Biometrika 1979;66:265-270

[19] Feeny D. Furlong W. Torrance G.W. Goldsmith C.H. Zhu Z. DePauw S. Denton M. Boyle M. Multiattribute and single attribute utility functions for the Health Utilities Index Mark 3 system Medical Care 2002;40(2):113-128

[20] Brazier J.E. Deverill M. Green C. Harper R. Booth A. A review of the use of health status measures in economic evaluation Health Technology Assessment 1999;3(9)