HEDS Discussion Paper 10/07

Disclaimer:

This is a Discussion Paper produced and published by the Health Economics and Decision Science (HEDS) Section at the School of Health and Related Research (ScHARR), University of Sheffield. HEDS Discussion Papers are intended to provide information and encourage discussion on a topic in advance of formal publication. They represent only the views of the authors, and do not necessarily reflect the views or approval of the sponsors.

Published paper

None.

ScHARR

School of Health and Related Research

Valuation of the Child Health Utility Index 9D (CHU9D)

Katherine Stevens

Health Economics and Decision Science (HEDS), ScHARR,
The University of Sheffield, UK.

Address for correspondence:

Katherine Stevens
Health Economics and Decision Science
School of Health and Related Research, University of Sheffield,
Regent Court, 30 Regent Street
Sheffield S1 4DA, UK
Telephone: 0114 222 0841
Fax: 0114 272 4095
Email: K.Stevens@Sheffield.ac.uk

Key Words: Economic evaluation, quality of life, paediatric, preference based measures, health state valuation

Abstract

Objectives

The aim of this study was to test the feasibility of estimating preference weights for all health states defined by the Child Health Utility 9D (CHU9D), a new generic measure of health related quality of life for children. This will allow the calculation of quality adjusted life years (QALYs) for use in paediatric economic evaluation.

Methods

Valuation interviews were undertaken with 300 members of the UK adult general population using standard gamble and ranking valuation methods. Regression modelling was undertaken to estimate models that could predict a value for every health state defined by the CHU9D. A range of models were tested and evaluated based on their predictive performance.

Results

Models estimated on the standard gamble data performed better than the rank model. All models had a few inconsistencies or insignificant levels and so further modelling was done to estimate a parsimonious consistent regression model, by combining inconsistent levels and removing non significant levels. The final preferred model was an OLS model where all coefficients were significant, there were no inconsistencies and the model had the best predictive performance.

Conclusion

This research has demonstrated it is feasible to value the CHU9D descriptive system and preference weights for each health state can be generated to allow the calculation of QALYs. The CHU9D can now be used in the economic evaluation of paediatric health care interventions. Further research is needed to investigate the impact of children’s preferences for the health states and what methods could be used to obtain these preferences.

Introduction

Recent research by Stevens [1,2,3] developed a descriptive system for a new generic health related quality of life measure, the Child Health Utility 9D (CHU9D). The descriptive system contains 9 dimensions (worried, sad, pain, tired, annoyed, school work, sleep, daily routine and activities), each with 5 levels. The measure was developed with the intention of being preference based and therefore suitable for use in paediatric economic evaluation. To meet this demand, preference weights need to be developed for each health state defined by the descriptive system. The objectives of this research were therefore to test the feasibility of valuing this descriptive system and to generate preference weights for every health state defined by the CHU9D.

Methods

Valuation technique

Currently, the recommendation is to obtain health state preferences using either the Standard Gamble (SG) or Time Trade Off (TTO) elicitation methods [4][5]. In addition, current guidance from the UK National Institute for Health and Clinical Excellence [6] recommends for its reference case that a choice based method be used, such as the SG or TTO.

NICE also recommends that preferences are obtained from a representative sample of the public. [6] In order to be consistent with NICE and due to its successful application in the valuation of other preference based instruments, including the SF-6D [7], the Health Utilities Index 2 [8] and the ADQoL [9] the SG method was chosen using a sample of the UK adult general population. In addition to SG, ranking was also used as an alternative method of valuation. There has been recent interest in using this technique for health state valuation [10] and it may have potential for use in future valuation work to try and obtain preferences from children as it is thought to be less cognitively demanding and so the feasibility of using ranking methods to value the descriptive system was also tested.

The standard gamble approach asks respondents to make a choice between a certain intermediate outcome and the uncertainty of a gamble with two possible outcomes:

- Choice A: 100% chance of the health state being valued
- Choice B: A chance of perfect health with probability p and a chance of dead with probability 1-p.

The value of p is varied until the respondent is indifferent between the two alternatives A and B. The point of indifference is the utility value of the health state.

Sample

Due to resource constraints, there was a limit to the number of interviews that could be undertaken and a trade off had to be made between the number of interviews achieved and geographical spread. As previous research has demonstrated that geographical location does not make a difference to health state values [11], local interviews were undertaken as this meant a much larger sample size, which was felt to be more important than achieving geographical spread. A spread across age, gender, ethnicity and social class could still be achieved however. A sample size of 300 was possible and compares favourably with the 200 used in the UK valuation of the HUI2 [12] and the ADQoL [9], although approximately half of the sample size used in the SF-6D valuation [7].

An interview team of three people was contracted to undertake the interviews from the Centre for Research and Evaluation (CRE) at Sheffield Hallam University. The interview team were very experienced in this type of work, particularly in the use of SG methods. CRE also undertook the sampling, management of the interviews and entered the data.

A random street sample was selected from addresses in Sheffield and Huddersfield using AFD Names and Numbers software which provides access to UK names and addresses for over 39 million people. The sampled households were all posted a letter, inviting them to take part in the research. When interviewers called at the household, participants were given an information leaflet to read with further information. After reading this and asking any questions, participants who agreed to take part were asked to sign a consent form and an interview was arranged in their own homes at their convenience.

Selection of health states

As there are 1,953,125 unique health states (levels^$^{\text{dimensions}}$ =5^^{9}) defined by the CHU9D descriptive system, it was infeasible to value them all. Instead, a sample was selected to be valued and a model estimated using these values, to predict a value for every health state defined by the system. This approach has been used successfully in a number of previous valuation surveys [13], [7], [12]. Previous experience has shown that respondents can manage about 9 valuation tasks in an interview, [7], [13] so with 300 interviews, this gave 2700 potential observations.

An orthogonal array of health states was generated using the Orthoplan feature of SPSS (Version 12) which generates a design for an additive model using the minimum number of health states required to estimate a model to predict all health states in the descriptive system. This found the minimum number of states required for a 9 dimensional system with 5 levels per dimension was 64. The design generated included 2 duplicate states and also the best state in the descriptive system (level 1 on each dimension, termed state 111111111) twice. As the design of the SG method assumes the best state to be equal to 1, it was not possible to use this as an intermediate state for valuation, therefore two substitute states were created to replace these best states, with all dimension levels at the top apart from 1 dimension, so as to keep the replacement health states as close as possible to the top of the descriptive system, keeping as close to the orthoplan design as possible. It is common to get duplicate states in this size of descriptive system and so more observations were obtained on the duplicate states, rather than substituting more states. Each health state can be represented by a 9 digit number, with each digit representing the level on each of the dimensions. The 64 states were divided into 8 sets of 8, trying to balance the severity of states in each set (by looking at the levels on each dimension) and making sure the 2 duplicate states were separated. The worst health state (Pits, 555555555) was added to each set, giving a total of 9 health states in each set. The interviewers rotated round the sets so that each state got an equal number of observations and each respondent only had 9 SG valuation tasks to do.

As adults were valuing the health states, the schoolwork/homework dimension was changed to work. This meant that the health states that the adults were considering for valuation made sense as adult health states. The descriptive system is given in Figure 3 and an example health state is shown in Figure 4.

Valuation interviews

The respondent was first asked to complete the descriptive system in order to familiarise themselves with it and also to understand the range of the levels within each dimension. Respondents were then asked to rank the 9 health states from the set being used, plus the best state in the descriptive system (111111111) and dead. This was followed by SG valuations of all 9 health states. The version of the SG script used was the same as that used in the original and UK HUI2 valuations [12], [8] which is the ping pong version. A Chance Board prop was used, which displays the probabilities numerically and in a pie chart. This method uses increments of 10% except at the top and bottom ends (between 90 and 100 and 0 and 10) where it uses increments of 5%. The perspective the respondent was asked to imagine was that they would be in the health state described for the rest of their life. The health state was valued against perfect health (as state 1,1,1,1,1,1,1,1,1 in this descriptive system is assumed to be equivalent to perfect health) and dead. If respondents rated a health state worse than dead in the ranking exercise, they did a worse than dead form of SG valuation where a certain choice of dead was offered as a choice against perfect health with probability p and the health state with probability (1-p). The utility value is –p at the point of indifference. The methodology is that undertaken in the valuation of the SF-6D [7]and the UK HUI2 [12] valuation and is based on the transformation by Patrick et al [14].

Finally, basic socio demographic information was collected and questions asking respondents how difficult they found the tasks. After completion of the interview, the interviewer was asked to assess how well they felt the respondent had concentrated and understood the tasks. The start and end times of the tasks were recorded. To obtain information on the socio-economic characteristics of respondents, the database supplying the names and addresses (AFD) [15] also provides censation data on the households (based on the latest census data), including affluence, which categorizes people into five categories (wealthy, prosperous, comfortable, striving and struggling). The figures for the sample were compared to the UK population as a whole.

Exclusion criteria

Based on the principles used in previous valuation work, for the SG modelling, respondents were excluded if they valued all health states the same, as this was taken to be an indication of misunderstanding the task [13], [9], [12] and observations were excluded if the data was unusable. Whilst imposing restrictions on the data, this is common practice in this type of research [7], [12].

Modelling

The aim was to estimate a model for predicting health state values for every health state defined by the descriptive system. The approach taken in previous valuation work of generic measures was followed [12], [7].

The basic model structure for the model was:

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

Where:

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

j= 1,2,…..,m represents individual respondents;

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 of each dimension in the descriptive system.

εij = the error term.

Personal characteristics were not included in the model as the model is intended to be used as a societal model of preferences and not adjusted for individual characteristics. Whilst there may be personal characteristics that prove important and can be estimated as respondent level covariates, these will not be used when applying the algorithm in practice as the aim was to estimate a utility function for the UK population as a whole.

To estimate the model, dummy variables were created for each level on each dimension with level 1 acting as the baseline for each dimension. The dummy variables take a value of 1 if the health state has the dimension at the level the dummy is representing and 0 otherwise. In a simple linear model, the intercept represents the estimated value of the best state (1,1,1,1,1,1,1,1,1) and the value for all other health states are derived by summing the coefficients of the appropriate dummies. Models estimated in this way have utility as the dependent variable. An alternative is to assume that because this is a generic measure, the best state can be assumed to be perfect health and so takes a value of 1. To achieve this, the constant can be forced to unity by estimating the model with no intercept and the dependent variable becomes disutility (U_{ij} -1). The value of a health state then becomes 1 plus the sum of the coefficients on the relevant dummy variables. Both the utility and disutility forms of model were estimated.

There are numerous possible interaction terms in the descriptive system and modelling them all would require a much larger dataset to prevent the risk of finding statistical significance due to chance [12]. In addition, previous valuation work has found that interactions do not improve the models and often increase the number of inconsistencies [11], [5], [8]. As the orthogonal design of the survey was for main effects only, only very basic interactions were estimated. Therefore two interaction terms were added, following the previous methods of the UKHUI2 and SF-6D valuations. A dummy variable (MOST) taking the value of 1 if a health state had any level at level 1 otherwise 0 and a dummy variable (LEAST) taking the value of 1 if a health state had any level at level 5, otherwise 0 were added.

Other specifications and transformations such as Tobit models could be considered, however previous valuation work has shown that these do not improve the modelling at all [12], [7]. Another recent approach that has been applied is that of Bayesian non-parametric modelling [16], [17]. This is a complex approach and requires a balanced design approach to the sampling of the health states which is different from the orthogonal array used here and so was not considered.

Specification of the models

The choice of model specification depends upon the type of data used. Standard OLS regression assumes a zero mean, constant variance error structure, with independent error terms, i.e. $\mathrm{cov}(\varepsilon ij,\varepsilon i'j)=0$, $i\neq i$` [18]. This assumption means that the 2700 observations from 300 respondents are treated as though 2700 respondents provided them.

An alternative specification is the Random Effects (RE) model, which allows for the fact that the error term may not be independent of the respondent, and separates out within and between respondent error terms.

The RE model also assumes that the allocation of health states to respondents is random i.e. cov(uj, eij)=0 [12].

A different specification is the fixed effects model, which also allows for the importance of individual effects but does not assume these are random, instead the respondent specific variation is estimated along with the coefficients on the explanatory variables.

To test whether individual effects were important the Breusch-Pagan test was used and if they were, the Hausman test was used to determine whether fixed or random effects were appropriate [7].

In addition to the individual level models described above, mean and median aggregate level models were estimated using the mean and median values for each health state.

One further type of modelling was undertaken, which made use of the ordinal preference data obtained from the ranking exercise. This type of rank modelling is a more recent development in health state valuation modelling and has been successfully applied in the major generic PBMs [10]. The basic foundation for this type of modelling (estimating cardinal values from ordinal data) is based on Thurstone's law of comparative judgement [10]. The modelling process followed that undertaken by McCabe in estimating a rank model for the HUI2 [19]. The ranking task was designed to include the state 'dead' so that the modelling could be normalised to produce a utility of 0 for dead. By including dead in the regression model, the estimated coefficient can be used to rescale the results onto a scale with dead as 0 [10].

The model is a rank-ordered logit model with a vector of dummy explanatory variables for each level of each dimension. A dummy variable for dead was included which took the value of 1 for this state and 0 otherwise. Perfect health is constrained to equal 1 and the value of a health state is calculated by subtracting the sum of the coefficients for each of the dummy variables from 1. As the model is not directly estimating utility on the 0-1 (dead- perfect health) scale required for health state valuation, the coefficients have to be rescaled using the formula $\beta rij=\beta ij/\theta$; where βrij is the rescaled coefficient and θ is the coefficient for dead. By rescaling, the model produces values on the dead (=0) - perfect health (=1) scale.

Assessment of models

Several measures were used to assess model performance. Firstly, coefficients of the models were examined to see if they were significant and had the expected negative (for utility) or positive (for disutility) sign. As the dummies represent progressively worse problems on each dimension starting from a baseline of no problems, the coefficients were expected to be increasing in absolute size.

Logical inconsistencies in the coefficient values were looked for and the adjusted R^2 for each model was also reported (where appropriate).

Models were also assessed on the basis of their predictive performance, i.e. how well they predicted observed mean values. To do this, a number of measures were used. The mean absolute error (MAE) and root mean square error (RMSE) were calculated, which are both summary measures giving an indication of the prediction errors of a model, with the RMSE giving more weight to larger errors. In addition, the percentage of health states predicted to within 0.1 and 0.05 (absolute value) of the observed mean value are reported. The value of 0.1 was chosen as it is the value used in previous valuation studies [12] and the value of 0.05 was chosen as it has been considered an important difference in many contexts [20].

Finally, the predicted health state values were plotted against the observed health state values to look for any patterns in the errors. Both the SG and the rank models were tested against the observed mean SG values after the exclusion criteria were applied. A test of the null hypothesis that the mean prediction error was 0 was undertaken for each of the models in order to determine whether there was any bias in the predictions. A Ljung box test was also carried out to test whether there was any non randomness in the prediction errors, i.e. if the error was systematically related to the severity of the health state [21]. Errors were ordered by actual mean health state valuation.

The rank model was tested for a key assumption of this type of modelling which is the independence from irrelevant alternatives. All modelling and analysis was carried out using STATA version 10.

Results

1245 addresses were mailed to and of these, 1195 were approached in person at the door. Out of those approached, 534 (45%) were not in/no contact was made, 320 (27%) refused and 300 (25%) agreed to be interviewed. Therefore the response rate was 25%. In total 300 interviews were carried out. For the SG modelling, 52 observations were excluded as unusable and 17 respondents were excluded as they valued all health states the same. In addition, 1 respondent was excluded as they valued the Pits state as 1 and all other health states at 0.95. This led to a dataset with 2478 observations from 282 respondents (6% of respondents were excluded). For the rank modelling, no exclusions were made.

The characteristics of the included and excluded populations are shown in Table 2. Compared to the included population, the excluded population had a higher percentage of men, more left school at 16, more found the SG exercise very difficult and the interviewer rated the understanding and concentration in the ranking and SG tasks lower.

The socio economic characteristics of the whole sample were as follows (data was missing for 1 person) and the UK figures are also given.

Category	Sample %	UK%
A, Wealthy	32.8	23.4
B, Prosperous	7.7	20.9
C, Comfortable	19.1	19.7
D, Striving	33.8	21.3
E, Struggling	6.7	13.6

Descriptive statistics for the health states from the included respondents are shown in Table 3. Each state was valued 35 times on average (minimum 32, maximum 39), apart from the 2 duplicate states (222222212 and 333333313) which were valued 68 and 72 times respectively. In addition, the Pits state (555555555) was valued 235 times.

The mean health state values range from 0.387857 to 0.931579. The median mostly exceeds the mean (66.7% of cases). There were 23 negative valuations (0.93%).

The interaction terms made things worse and did not improve the modelling as they increased the number of inconsistencies and decreased the number of significant coefficients and are not reported here. Overall, the disutility models were much better than the utility models in terms of the number of significant coefficients (higher) and the number of inconsistencies (lower), therefore only the disutility models are reported. The Breusch Pagan test suggested that individual effects were present in the data (chi2= 2388.23, p=0.00). The Hausman test did not work as the model fitted failed to meet the asymptotic assumptions of the Hausman test, probably due to a misspecification problem, in that the random effects model was not efficient for the data. A more general test was tried but this also failed. Therefore both random and fixed effects models were estimated and judged on the basis of their predictive performance.

Going on the number of significant coefficients and the number of inconsistencies, the best three models from all the disutility models were the OLS, RE and mean disutility models with the constant restricted to 1. They are summarized in Table 4.

All coefficients are significant in the rank model, have the expected sign and there are 8 inconsistencies. It is shown alongside the three best restricted disutility models in Table 4. The results of the Hausman test for the independence from irrelevant alternatives are shown in Table 1. Significant results are shown in bold. It was not possible to test the models estimated without health states ranked second and tenth as the model violated the assumptions. A more general test did not work either. The models are sensitive to excluding those health states at the top and towards the bottom of the rankings as those ranked first, eight and ninth and significant, hence we reject the equality and there is some evidence that the assumption does not hold.

For the OLS model, all the coefficients have the expected positive sign and there are 30 out of 36 coefficients significant at the 0.1 level. There are 14 inconsistencies which reduces to 10 if you remove those that are not significant at the 0.1 level. The RE model also has all coefficients with the expected positive sign and there are 33 out of 36 significant at the 0.1 level. There are 11 inconsistencies which reduces to 10 if you remove the one not significant at the 0.1 level (tired 5). The mean model has all coefficients with the expected positive sign and 28 out of 36 are significant. There are 14 inconsistencies which reduces to 8 removing those that are not significant at the 0.1 level. Finally, the rank model has all 36 coefficients significant at the 0.1 level and they are all the expected positive sign. There are 8 inconsistencies.

In terms of predictive performance, the OLS and mean models perform best with 100% of errors within +/-0.1, whilst the RE model is still high at 98.4 and the rank model is the worst, at 90.5. When the accuracy is increased to within 0.05, the mean model performs best, at 98.4% with the OLS next at 90.5%. The RE model is much lower at 77.8% and the rank model is the worst at 65.1%.

The MAE is lowest for the OLS model at 0.0261, closely followed by the mean model at 0.0263. The MAE of the RE model is higher at 0.0313 and the rank model has the highest at 0.0461. The RMSE is lowest for the mean model, then the OLS, then the RE and finally is highest for the rank model at 0.0573.

There did not appear to be any systematic pattern in the errors apart from the rank model which under predicted at the higher end (i.e. the health states with a higher observed mean value). The results of the Ljung box tests for each of the models (1 to 4) are shown below.

	OLS(1)	RE(2)	Mean(3)	Rank(4)
Test statistic	4.9116	8.9025	5.3772	7.9305
Prob>Chi2(8*)	0.7670	0.3506	0.7166	0.4403

* The number of lags is the square root of n which is conventional for this type of test.

None of the test statistics are significant, therefore none of the models show evidence of autocorrelation in the prediction errors.

The RE model appears to be the only model that gives biased predictions, as indicated by the t test of the null hypothesis that the mean prediction error is 0.

Further modelling

Despite the very good predictive performance, the models still have inconsistencies in them and some coefficients are not significant. Therefore further modelling was undertaken to estimate a parsimonious consistent regression model using the general to specific approach. This approach was used in the valuation of the SF12 and later SF-36 models [22]. These models were constructed by combining levels where inconsistencies were present and removing levels not significant at p<0.1. This was done on the two best performing models (the mean and OLS restricted). These two models are shown in Table 5 and graphs of their predictive performance are shown in Figures 1 and 2. All coefficients are significant at p<0.1 and all but 1 coefficient in both models are significant at p<0.05. There are no inconsistencies in these models. The models are also consistent in that the same levels had to be combined, apart from the OLS model which still has sleep4, whereas the mean model has sleep234 combined. The dimensions worry, annoyed and tired all had all levels (except level 1) combined. Levels 4 and 5 were combined for sad, levels 2 and 3 for pain, levels 2 and 3 for work and 4 and 5 for work and levels 2, 3 and 4 for activities. The predictive performance is not as good as the full models where levels were not combined. The MAE for the OLS model is 0.0343 and similar at 0.0349 for the mean model. These are higher than the full models. The RMSE are also similar, at 0.0426 for the OLS model and 0.0431 for the mean model. Both models predict well at 98.41% of predicted values within 0.1 of the observed mean, and the mean model is slightly better at 76.19% of predicted values within 0.05 of the observed mean compared to the OLS model, which predicts 73.02%. Neither model had biases in the prediction errors as indicated by the t test of the null hypothesis that the mean prediction error is 0. There do not appear to be any patterns in the prediction errors either from looking at the graphs in Figures 1 and 2. The results of the Ljung box test are shown below for both models. Neither model shows evidence of autocorrelation in the prediction errors.

	OLS reduced (5)	Mean reduced (6)
Test statistic	6.5165	6.7737
Prob>Chi2(8*)	0.5896	0.5612

* The number of lags is the square root of n which is conventional for this type of test.

Discussion

Health state values were successfully generated for the health states in the survey and a reasonable range of values was produced, although the mean value for the Pits state (0.337) was perhaps higher than what was expected. This compares with other generic descriptive systems with mean health state values which ranged from -0.543 to 0.878 (EQ-5D [13]), 0.10 to 0.99 (SF-6D [7]) and -0.07 to

0.79 (UKHUI2) [12]. The Pits state could be low due to the language used due to the nature of it being a paediatric descriptive system. Adults did not have any knowledge that the states being valued were child health states when undertaking the valuation tasks. Hence, when reading the descriptions, they may have placed less weight on the severity of the levels. For example, the level "I feel very worried today" may be seen by a child as really severe, however for an adult, who is perhaps thinking in terms of stronger language such as anxiety, this might not seem so severe, as they can imagine much worse levels, for example" I feel really anxious". Changing schoolwork/homework to work for adults also changes one of the dimensions of the descriptive system. Whilst this is perhaps the closest in meaning that this can be for an adult, there are differences in how health may affect work and so the implications of this mean that the health state being valued is not quite what is intended. This is something that can be tested in future valuation work, for example valuation work with adults that considers the actual child state from the perspective of the child, instead of a non child state from their own perspective.

Generally, the modelling was successful and overall the disutility models performed much better than the utility models and the best performing of these were the models where the constant was restricted to equal 1. This fits in well with the practical application that is required of these models in calculating QALYs, in that a scale with perfect health =1 and dead =0 is required and there are strong theoretical arguments for restricting the intercept to unity [7]. This model assumes that the best health state in the descriptive system (111111111) has a value of 1 and dead has a value of 0.

Overall, the mean model was the best in terms of predictive performance as it is the most accurate at predicting observed mean values, with the highest percentage predicted within +/-0.05 for all models, the lowest RMSE and a low MAE (nearly equivalent to OLS MAE which is the lowest). The mean model also has one of the lowest number of inconsistencies (8), the same as the rank model. The rank model was the worst in terms of predictive performance, being the worst on each measure of performance, however it is the only model with all coefficients significant at the 0.1 level and has the same number of inconsistencies as the mean model. None of the models had any problems with autocorrelation in the prediction errors.

It should be noted that the SG models are being tested against the data they were estimated on, whereas the rank models are not, although the data comes from the same respondents. Therefore it is perhaps not surprising that the rank model is outperformed by all the SG models. However, the rank model is still a reasonably good model and there are similarities with the inconsistencies in the SG models, for example, sad5, tired4, tired5, pain3 and activities4. It is notable that the rank model performs well and is similar to the SG models in terms of what levels are inconsistent and this gives encouraging results for using this type of valuation technique in the future to access children's valuations.

Other studies that have used rank models have found that the results are not dissimilar to the SG models. The UK valuation of the HUI2 found the rank model increased the inconsistencies by one [19] and found the best SG model performed better on all tests, but was remarkably similar. The SF-6D found the rank model quite different to the best performing SG model, as the number of inconsistencies decreased however the predictive performance of the rank model was only slightly worse [19].

The inconsistencies were similar across the different model specifications, the most common were sad5, annoyed3, annoyed4, pain3, sleep3, work3, work5 and activities4. The exception was the rank model which was the only model to be consistent for the work dimension.

Estimating parsimonious consistent models from the two best performing full models worked, although several levels had to be collapsed. The results of this were similar across the two models which is reassuring. Part of the collapsing may be due to the fact that adults were valuing these health states and not children. For example the dimension *worry* is perhaps not seen as very strong by an adult in contrast to the similar concept usually used in adult measures, anxiety. Similarly, it may be that adults see being *annoyed* as nothing particularly bad and so this dimension also had collapsed levels. Perhaps the most surprising dimension that had to combine levels was *tired*, however it may be that because there is also a *sleep* dimension, the adults valuing these health state focused in on that. Alternatively, it may be that the descriptive system is too big with 9 dimensions or perhaps there are too many levels and adults are employing simplifying heuristics when valuing the health states. Larger descriptive systems are more likely to result in doing this, such as just focusing in on key dimensions [23]. Undertaking a large valuation survey with children valuing the health states would provide more information on this issue and also using ‘think aloud’ techniques when people are valuing health states to gain a better understanding of what they are focusing on and whether they use any heuristics.

The preferred overall model is model 5, the OLS parsimonious consistent model. This model has all coefficients significant and has no inconsistencies. It is slightly better on predictive performance that the mean model.

The mean absolute error of the best model (model 5) is 0.0343 and this amount is unlikely to be considered meaningful in many contexts [8]. As the aim of the model is to predict mean health values across patients in many different states and the error is random, this is an acceptable error when using the model in practice. Research by Walters and Brazier [24] also found that the minimally important difference in utility score for the SF-6D was 0.041 (mean) and 0.074 (mean) for the EQ-5D.

The proportion of health states valued out of the entire descriptive system was very small at 0.003% (63/1,953,125). This compares with 1.4% for the SF-6D valuation and 0.64% for the UK HUI2 valuation [12]. It may be that with a larger dataset and a larger number of health states being valued, some of these problems may be overcome.

One of the most important factors in this valuation study is that the population valuing the health states is adults, in contrast to the descriptive system which is for children. In addition, when valuing the health states, adults were not aware that these states were child health states and were asked to imagine themselves as they are now (as an adult) in this health state for the rest of their lives. Adults were chosen for the valuation survey as using children to undertake valuations is something that has not been done before. This does not mean it is not possible, rather that further research is needed to investigate whether it is feasible to obtain valuations from children. The SG and TTO methods are cognitively demanding and it is uncertain whether children would be able to manage these tasks. There are also ethical issues that would be raised by asking to children to think about scenarios that involved a risk of death. In recent years, the use of ranking/ordinal methods to value health states has increased and this is perhaps a method that would be more appropriate and feasible to undertake with children. Perhaps the simplest way would be to present health states in pair wise comparisons as ranking many health states at once can be just as cognitively demanding. Work has been done to value descriptive systems in this way using ordinal techniques, including the use of discrete choice experiment (DCE) techniques for estimating preference weights for a sexual quality of life questionnaire [25] and for an asthma quality of life questionnaire [26]. It would also be interesting to undertake preference elicitation work where the adult valuing the health state knew this was a child health state and see if this makes any difference to the values. This was the approach taken in the valuation of the HUI2 and the ADQoL [8], [9] . Whether children’s valuations should be used is a normative issue and there are arguments either way. It can be argued that children are not rational, informed and autonomous individuals (an ideal for health state valuation) and therefore should not undertake valuation tasks. However, it may be that some adults also do not fulfil this criteria and previous valuation work has demonstrated some evidence of this as respondents have been excluded on the grounds of irrational or inconsistent responses [7], [12]. Perhaps more importantly, some people may argue that society does not see children as legal agents, in that before the age of 18 they are not allowed to vote and hence not viewed as decision makers in society.

What is unknown is how children’s valuations may differ (if at all) from adult valuations. If there is very little difference in values, then it perhaps becomes irrelevant whose values are used. However, if there are differences, then a decision would have to be made over which values are more appropriate. This is something that can only be determined empirically and what would be most interesting is perhaps the potential differences in strength of preference for the different dimensions of health. These important questions should be the subject of future research.

Conclusions

This research has demonstrated that it is feasible to value the CHU9D descriptive system and preference weights have been generated for all health states defined by the system. A number of models have been estimated using both the SG and ordinal (rank) data. The best performing models were restricted disutility models, which restrict the constant term to 1 and have stronger theoretical arguments. The model recommended for use in assigning preference weights for the health states defined by the CHU9D is the OLS parsimonious model (model 5). The CHU9D is now able to be used to generate quality adjusted life years (QALYS) by using the system and combining it with length of life. The CHU9D offers an alternative to the HUI2 and can be used in the economic evaluation of paediatric health care interventions. Further research is needed to investigate whether it is feasible to obtain children’s preferences for the health states and whether this is desirable.

Table 1: Hausman test for the independence from irrelevant alternatives

Alternative dropped	Hausman	Prob>chi2
1	211.04	**0.0000**
2	-	-
3	38.60	0.3970
4	20.80	0.9854
5	15.56	0.9992
6	22.20	0.9741
7	25.65	0.9201
8	87.82	**0.000**
9	196.00	**0.000**
10	-	-
11	4.40	1.000

Table 2: Characteristics of the population (full sample n=300)

		Total (n=300)	Included (n=282)	Excluded (n=18)
Age in years (mean)		49.01	48.98	50.72
% Male		40.8	40.21	50
Employment (%)	employment or self-employment	51.33	52.48	33.33
	retired	29	28.37	38.89
	housework	7.33	6.74	16.67
	student	4	3.55	11.11
	seeking work	1.67	1.77	-
	unemployed	1.67	1.77	-
	long-term sick	3.67	3.9	-
	other	1.33	1.42	-
Highest level of education (%)	secondary school (left school at 16 or before)	51.01	50.36	61.11
	further education (left school at 18)	16.11	16.79	5.56
	higher education (university or college)	28.19	27.86	33.33
	post-graduate education	4.7	5	-
Ethnicity (%)	White	97.99	98.22	94.4
	Mixed/dual heritage	1.34	1.42	-
	Asian or Asian British	0.67	0.36	5.6
Difficulty with ranking exercise (%)	very difficult	14.48	14.7	11.11
	quite difficult	31.31	31.18	33.33
	neither difficult or easy	21.21	21.15	22.22
	fairly easy	26.94	26.88	27.78
	very easy	6.06	6.09	5.56
Difficulty with standard gamble exercise (%)	very difficult	6.35	6.05	11.11
	quite difficult	21.07	22.06	5.56
	neither difficult or easy	19.73	20.28	11.11
	fairly easy	43.48	43.42	44.44
	very easy	9.36	8.19	27.78
Understanding on ranking exercise (%)	fully understood the task	80.94	80.07	94.44
	partially understood the task	18.06	18.86	5.56
	did not really understand the task	1	1.07	-
Understanding	fully understood the task	82.61	82.21	88.89

on SG task (%)	partially understood the task	16.39	17.08	5.56
	did not really understand the task	1	0.71	5.56
Interviewer rating of respondents understanding of ranking task (%)	understood and performed tasks easily	65.65	66.67	50
	some problems but seemed to understand	28.91	28.99	27.78
	doubtful whether the respondent understood	5.44	4.35	22.22
Interviewer rating of respondents understanding of standard gamble task (%)	understood and performed tasks easily	71.43	72.46	55.56
	some problems but seemed to understand	22.79	22.46	27.78
	doubtful whether the respondent understood	5.78	5.07	16.67
Effort and concentration of respondent on ranking (interviewer assessed %)	Concentrated very hard and put a great deal of effort into it	40.82	42.39	16.67
	Concentrated fairly hard and put some effort into it	50	50.72	38.89
	Didn't concentrate very hard and put little effort into it	8.5	6.52	38.89
	Concentrated at the beginning but lost interest/concentration before reaching the end	0.68	0.36	5.56
Effort and concentration of respondent on standard gamble (interviewer assessed %)	Concentrated very hard and put a great deal of effort into it	41.5	43.12	16.67
	Concentrated fairly hard and put some effort into it	52.04	52.17	50
	Didn't concentrate very hard and put little effort into it	6.46	4.71	33.33

Table 3: Descriptive statistics for health states

State	Observations	Mean	Median	Std. Dev.	Min	Max
111111112	38	0.9316	0.975	0.1039	0.55	1
111121111	34	0.9206	0.95	0.1027	0.55	1
112254323	38	0.4888	0.45	0.2669	0.05	0.975
115422323	37	0.7426	0.75	0.2378	0.1	0.975
122531334	32	0.7195	0.75	0.2130	0.1	0.975
123342125	34	0.5757	0.65	0.2422	0.05	0.975
123523451	34	0.6838	0.75	0.2198	0.1	0.975
124315213	39	0.6654	0.65	0.2585	0.05	1
131232435	34	0.6103	0.65	0.2417	0.05	0.975
131435232	33	0.6538	0.65	0.2391	0.1	1
132143542	36	0.6028	0.65	0.2547	0.05	0.975
135123244	34	0.7331	0.8	0.2431	0.1	1
142312513	34	0.5904	0.55	0.2946	0	0.975
143253251	33	0.6061	0.65	0.2373	0	0.95
153324122	34	0.7404	0.8	0.2318	0.1	1
154231332	34	0.7441	0.75	0.1868	0.35	0.975
211543312	37	0.6946	0.75	0.2359	0.25	0.975
212335154	37	0.7291	0.75	0.2200	0.1	0.975
213125433	34	0.6191	0.55	0.2224	0.05	0.975
214233521	37	0.7264	0.85	0.2498	0.1	0.975
222222212	68	0.7699	0.85	0.1873	0.15	1
224153133	37	0.6649	0.75	0.2726	0.1	0.975
225341331	34	0.6846	0.675	0.1818	0.15	0.975
231324531	34	0.6993	0.75	0.2234	0.15	0.975
232114355	34	0.6125	0.65	0.2372	0.05	0.975
233451112	37	0.7264	0.75	0.2516	0.05	0.975
235211423	33	0.6568	0.75	0.2946	0.1	1
241532143	34	0.7051	0.65	0.2099	0.1	0.975
243131225	37	0.6764	0.75	0.2552	0.05	0.975
251313224	34	0.7551	0.75	0.2196	0.1	1
253412341	34	0.6221	0.775	0.2309	0.15	0.975
312351442	32	0.6047	0.65	0.1981	0.25	0.95
313212534	36	0.6535	0.55	0.2794	0.05	0.975
313514232	33	0.7811	0.675	0.1918	0.15	1
314321245	33	0.5879	0.8	0.2509	0.05	0.95
321152324	34	0.5537	0.65	0.2481	0.05	0.975

321421553	37	0.6811	0.55	0.2539	0.1	0.975
322413135	34	0.6787	0.7	0.2497	0.05	0.975
325134211	34	0.7625	0.7	0.1933	0.15	0.975
332245121	34	0.5559	0.75	0.2327	0.05	0.975
333333313	72	0.7236	0.575	0.2315	0.05	0.975
334522121	38	0.7105	0.75	0.2378	0.05	1
341125322	37	0.6608	0.75	0.2605	0.1	1
345213132	35	0.7021	0.75	0.2372	0.1	1
351241253	38	0.5105	0.75	0.2615	-0.1	0.975
352132411	37	0.7338	0.55	0.2491	0.05	0.975
411223315	34	0.6890	0.85	0.2479	0.05	1
415332152	37	0.7561	0.725	0.2310	0.1	0.975
423131522	34	0.6662	0.85	0.2309	0.05	0.975
423215341	34	0.6221	0.65	0.2639	0.05	0.975
431352231	33	0.6280	0.65	0.2728	0.05	1
432511223	34	0.7419	0.65	0.2279	0.05	1
444444414	37	0.5824	0.75	0.2629	0.05	0.975
452123133	37	0.7081	0.65	0.2374	0.1	1
512433221	34	0.7640	34	0.1652	0.35	0.975
513142233	34	0.6699	34	0.2334	0.05	0.975
521234143	34	0.6316	0.75	0.2416	0	0.975
521313422	34	0.7750	0.65	0.1526	0.35	0.975
533221114	34	0.6985	0.75	0.2409	0.05	0.975
534112352	37	0.7716	0.75	0.2419	0.1	0.975
542321331	38	0.6283	0.85	0.2818	0	0.975
555555515	35	0.3879	0.65	0.3070	-0.1	1
555555555	235	0.3368	0.45	0.3154	-0.75	1

Table 4: Models

	Model			
Coefficient	OLS (1)	Random Effects (2)	Mean (3)	Rank (4)[+]
Constant	1	1	1	1
worry2	0.0058	0.0117	0.0082	0.0206**
worry3	0.0363**	0.0292**	0.0380**	0.0342**
worry4	**0.0261**	0.0313**	**0.0250**	0.0417**
worry5	**0.0312***	0.0344**	**0.0324**	0.0964**
sad2	0.0405**	0.0335**	0.0430**	0.0457**
sad3	0.0435**	0.0377**	0.0458**	**0.0386****
sad4	0.0780**	0.0677**	0.0772**	0.0717**
sad5	**0.0688****	**0.0677****	**0.0699****	**0.0613****
annoy2	0.0380**	0.0271**	0.0398**	0.0377**
annoy3	**0.0316****	**0.0265****	**0.0334***	0.0382**
annoy4	**0.0248**	**0.0217***	**0.0233**	**0.0372****
annoy5	**0.0243**	0.0335**	**0.0257**	0.0572**
tired2	0.0668**	0.0390**	0.0679**	0.0377**
tired3	**0.0397****	**0.0276****	**0.0402****	0.0380**
tired4	**0.0355****	**0.0271****	**0.0353***	**0.0304****
tired5	**0.0380****	**0.0199**	**0.0376***	**0.0287****
pain2	0.0394**	0.0434**	0.0418**	0.0637**
pain3	**0.0241***	**0.0285****	**0.0259**	**0.0409****
pain4	0.1236**	0.1301**	0.1216**	0.1035**
pain5	0.1471**	0.1504**	0.1475**	0.1135**
sleep2	0.0319**	0.0248**	0.0344**	0.0315**
sleep3	**0.0091**	**0.0176***	**0.0107**	0.0330**
sleep4	0.0489**	0.0543**	0.0476**	0.0678**
sleep5	0.0955**	0.0910**	0.0971**	0.0699**
daily2	0.03525**	0.0411**	0.0372**	0.0382**
daily3	0.0595**	0.0592**	0.0610**	**0.0358****
daily4	0.0685**	0.0803**	0.0677**	0.0620**
daily5	0.0969**	0.1022**	0.0990**	0.0963**
work2	0.0485**	0.0519**	0.0413**	0.0443**
work3	**0.0454****	**0.0457****	**0.0379****	0.0523**
work4	0.0842**	0.0801**	0.0770**	0.0756**
work5	**0.0507****	**0.0578****	**0.0458****	0.1039**
activ2	0.0115	0.0122	0.0128	0.0314**
activ3	0.0634**	0.0535**	0.0646**	0.0484**

activ4	**0.0422****	**0.0336****	**0.0415****	**0.0396****
activ5	0.1148**	0.1018**	0.1163**	0.0766**
Dead	-	-	-	1**
N	2478	2478	63	3000
Inconsistencies (after removing insignificant ones)	14 (10)	11 (10)	14(8)	8(8)
% within +/-0.1	100	98.4	100	90.5
% within +/-0.05	90.5	77.8	98.4	65.1
MAE	0.0261	0.0313	0.0263	0.0461
RMSE	0.0312	0.0397	0.0309	0.0573
T test	-0.944	-4.522**	-0.505	1.660

+ rescaled coefficients

**significant at p<0.05

*significant at p<0.1

Inconsistencies are shown in bold type

Table 5: Parsimonious consistent models

	Model				
Coefficient	OLS (5)	P>t	Coefficient	Mean (6)	P>t
Constant	1		Constant	1	
worry2345	0.0227	0.047	worry2345	0.0251	0.082
sad2	0.0420	0.003	sad2	0.0438	0.018
sad3	0.0445	0.002	sad3	0.0460	0.013
sad45	0.0722	0	sad45	0.0728	0
annoy2345	0.0313	0.006	annoy2345	0.0326	0.025
tired2345	0.0479	0	tired2345	0.0482	0.001
pain23	0.0332	0.004	pain23	0.0349	0.02
Pain4	0.1245	0	Pain4	0.1225	0
Pain5	0.1426	0	Pain5	0.1461	0
sleep23	0.0212	0.08	sleep234	0.0280	0.059
sleep4	0.0506	0.004			
sleep5	0.0907	0	sleep5	0.0952	0
daily2	0.0371	0.009	daily2	0.0379	0.039
daily3	0.0612	0	daily3	0.0612	0.001
daily4	0.0699	0	daily4	0.0682	0.003
daily5	0.0930	0	daily5	0.0971	0
work23	0.0487	0	work23	0.0403	0.016
work45	0.0656	0	work45	0.0609	0.002
activ234	0.0368	0.001	activ234	0.0376	0.01
activ5	0.1079	0	activ5	0.1129	0
N	2478			63	
Inconsistencies	0			0	
% within +/-0.1	98.41			98.41	
% within +/-0.05	73.02			76.19	
MAE	0.0343			0.0349	
RMSE	0.0426			0.0431	
T test	-0.770			-0.336	

Figure 1: Observed and predicted values (OLS parsimonious model)

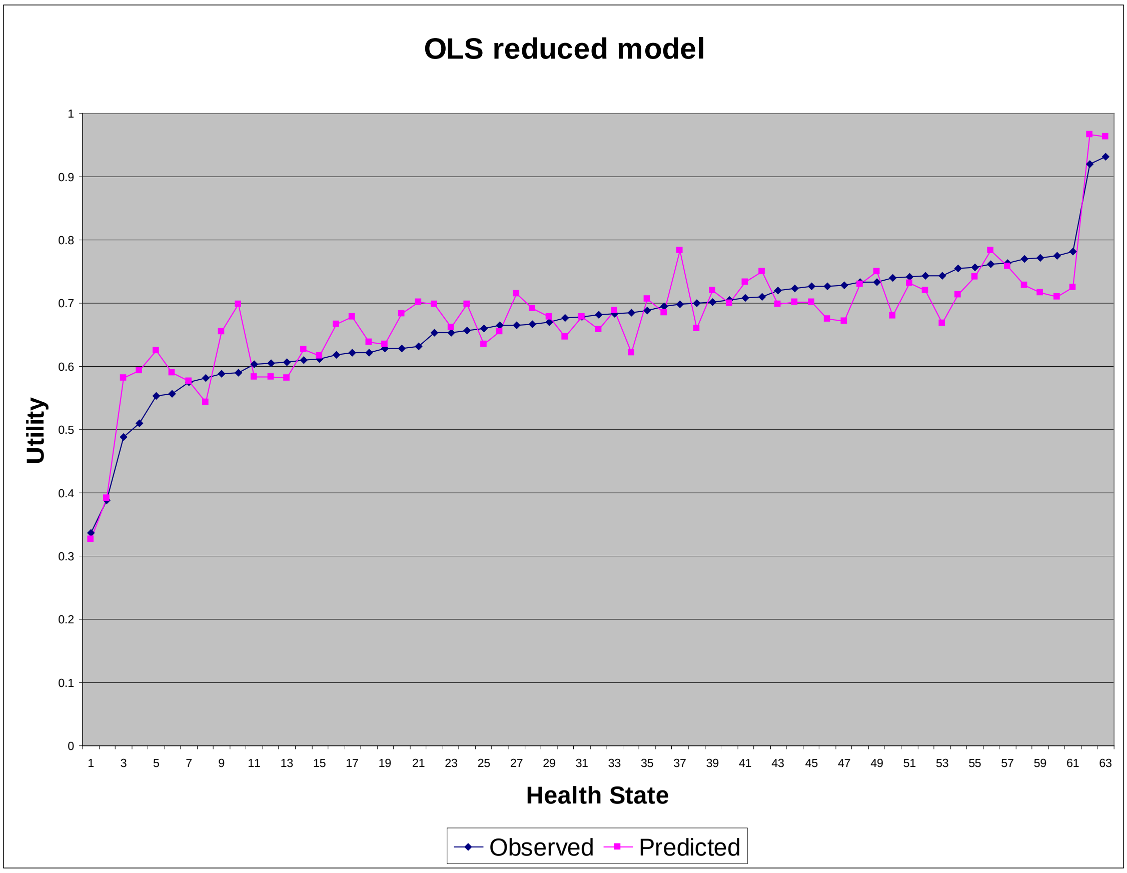

Figure 2: Observed and predicted values (Mean parsimonious model)

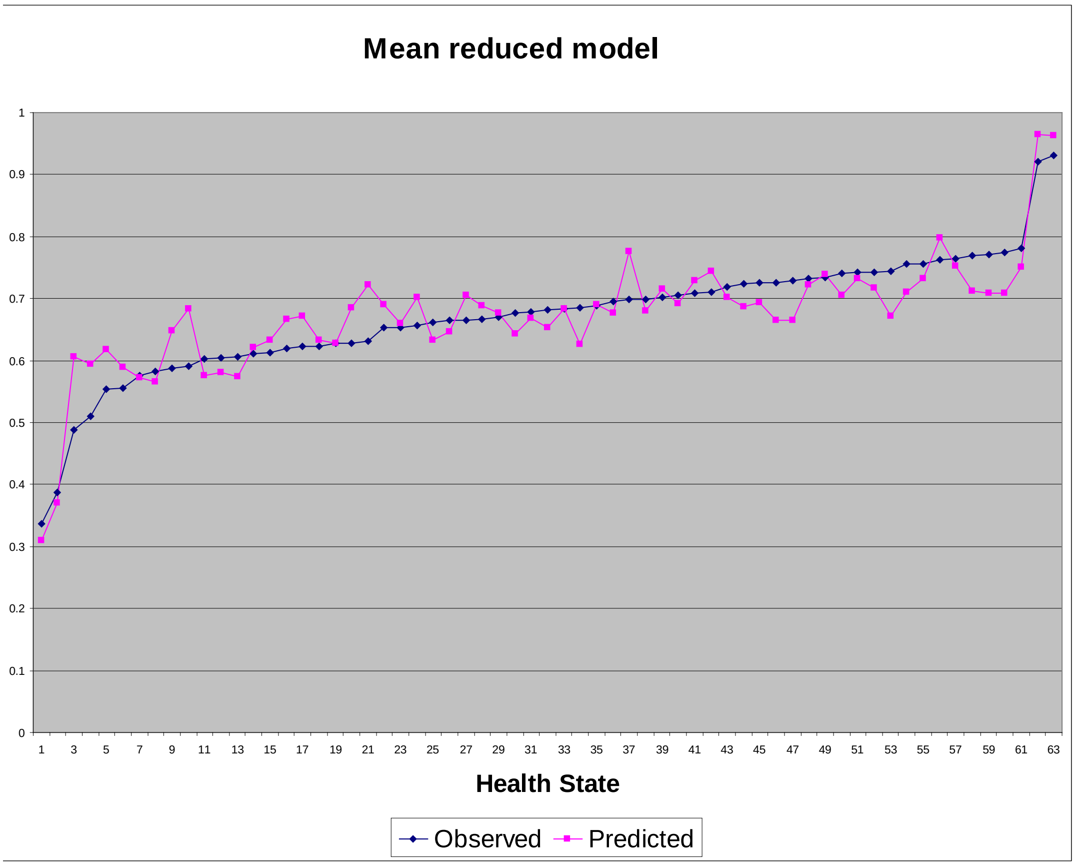

Figure 3 Descriptive system

Dimension	Level	Description
Worried	1	I don’t feel worried today
	2	I feel a little bit worried today
	3	I feel a bit worried today
	4	I feel quite worried today
	5	I feel very worried today
Sad	1	I don’t feel sad today
	2	I feel a little bit sad today
	3	I feel a bit sad today
	4	I feel quite sad today
	5	I feel very sad today
Annoyed	1	I don’t feel annoyed today
	2	I feel a little bit annoyed today
	3	I feel a bit annoyed today
	4	I feel quite annoyed today
	5	I feel very annoyed today
Tired	1	I don’t feel tired today
	2	I feel a little bit tired today
	3	I feel a bit tired today
	4	I feel quite tired today
	5	I feel very tired today
Pain	1	I don’t have any pain today
	2	I have a little bit of pain today
	3	I have a bit of pain today
	4	I have quite a lot of pain today
	5	I have a lot of pain today
Sleep	1	Last night I had no problems sleeping
	2	Last night I had a few problems sleeping
	3	Last night I had some problems sleeping
	4	Last night I had many problems sleeping
	5	Last night I couldn’t sleep at all
Daily routine	1	I have no problems with my daily routine today
	2	I have a few problems with my daily routine today
	3	I have some problems with my daily routine today
	4	I have many problems with my daily routine today
	5	I can’t do my daily routine today
Work	1	I have no problems with my work today
	2	I have a few problems with my work today
	3	I have some problems with my work today
	4	I have many problems with my work today
	5	I can’t do my work today
Able to join in activities	1	I can join in with any activities today
	2	I can join in with most activities today
	3	I can join in with some activities today
	4	I can join in with a few activities today
	5	I can join in with no activities today

Figure 4: Example Health State

Health State 153324122

I don’t feel worried

I feel very sad

I feel a bit annoyed

I feel a bit tired

I have a little bit of pain

I have many problems sleeping

I have no problems with my daily routine

I have a few problems with my work

I can join in with most activities

Acknowledgements

This research was funded as part of a UK Medical Research Council Special Training Fellowship in Health Services and Health of the Public Research. Thanks go to the Centre for Research and Evaluation at Sheffield Hallam University and to all the participants who took part.

References

1. Stevens, K J. Developing a descriptive system for a new preference-based measure of health related quality of life for children. Quality of Life Research 2009; 18 (8): 1105-1113

2. Stevens, K J. Working With Children to Develop Dimensions for a Preference-Based, Generic, Pediatric Health-Related Quality-of-Life Measure. Qualitative Health Research 2010a; vol. 20: 340 - 351

3. Stevens, K J. Assessing the performance of a new generic measure of health related quality of life for children and refining it for use in health state valuation. Health Economics and Decision Science Discussion Paper 10/02. 2010b. Available from http://eprints.whiterose.ac.uk/view/iau/Sheffield=2EHCM=2EWP.html

4. Gold, M. R., Siegel, J. E., Russell, L.B. et al. Cost Effectiveness in Health and Medicine. 1996. Oxford University Press. Oxford
5. Brazier, J.E., Deverill, M., Harper, R., et al. A review of the use of Health Status measures in economic evaluation. Health Technology Assessment. 1999. 3(9)
6. National Institute for Clinical Excellence. Guide to the Methods of Technology Appraisal. April 2004. NICE.
7. Brazier, J.E., 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.
8. Torrance, G., Feeny, D., Furling, W., Barr, R., Zhang, Y. & Wang, Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Medical Care 1996; 34, 702-22.
9. Stevens, K. J., Brazier, J. E., McKenna, S. P., Doward, L. C. & Cork, M. J.. The development of a preference-based measure of health in children with atopic dermatitis. British Journal of Dermatology. 2005; 153, 372-377
10. Brazier, J.E., Ratcliffe, J., Salomon, J. & Tsuchiya A. Measuring and Valuing Health Benefits for Economic Evaluation. 2007. Oxford University Press.
11. The MVH Group, Dolan, P., Gudex, C., Kind, P. & Williams, A. May The Measurement and Valuation of Health. First Report on the Main Survey. 2004. Centre for Health Economics, University of York.
12. McCabe, C., Stevens, K., Roberts, J. & Brazier, J.E. Health State Values for the Health Utilities Index Mark 2 descriptive system: results from a UK valuation survey. Health Economics 2005 14; (3) 231-44.
13. Dolan, P. Modelling valuations for EuroQol Health States. Medical Care 1997; 35(11), 1095-1108.
14. Patrick, D.L., Starks, H.E., Cain, K.C. et al. Measuring preferences for health states worse than death. Medical Decision Making; 1994. 14, 9-18.
15. Names and numbers manual. http://www.afd.co.uk/manuals/namesandnumbers/contents.htm Accessed 03/07/2008

16. Kharroubi SA, Brazier JE, Roberts J, O'Hagan A. Modelling SF-6D health state preference data using a nonparametric Bayesian method. Journal of Health Economics 2007; 26**:** 597-612**.**
17. Kharroubi S, McCabe C. Modelling HUI 2 health state preference data using a nonparametric Bayesian method. Medical Decision Making. 2008; 28: 875-887.
18. Gujarati, D.N. Basic Econometrics. Chapter 3. 1995. McGraw-Hill, Inc. Third Edition.
19. McCabe, C., Brazier, J.E., Gilks, P., Tsuchiya, A., Roberts, J., O'Hagan, A. & Stevens, K. Using rank data to estimate health state utility models. Journal of Health Economics 2006; 25, 418-431.
20. O'Brien, B. J., Drummond, M. F. Statistical versus quantitative significance in the socioeconomic evaluation of medicines. Pharmacoeconomics 1994; 5, 389.
21. Ljung, G., Box, G. On a measure of lack of fit in time series models. Biometrika 1979; 66, 265-270.
22. Brazier, J. E. & Roberts, J. R. The Estimation of a Preference-Based Measure of Health from the SF-12. Medical Care 2004; 42(9).
23. Lloyd, A. Threats to the estimation of benefit: are preference estimation methods accurate? Health Economics 2003; 12, 393-402.
24. Walters S.J., & Brazier J.E. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Quality of Life Research 2005; 14, 1523-1532.
25. Ratcliffe J, Brazier JE, Tsuchiya A, Symonds T, Brown M. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Economics 2009;18 (11), 1261-1276
26. John Brazier, Donna Rowen, Yaling Yang and Aki Tsuchiya. Using rank and discrete choice data to estimate health state utility values on the QALY scale. Health Economics and Decision Science Discussion Paper 09/10, 2009. The University of Sheffield. Available from http://eprints.whiterose.ac.uk/view/iau/Sheffield=2EHCM=2EWP.html