Article:
Ratcliffe, J., Huynh, E., Chen, G. et al. (2016) Valuing the Child Health Utility 9D: Using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm. Social Science and Medicine, 157. pp. 48-59. ISSN: 0277-9536

https://doi.org/10.1016/j.socscimed.2016.03.042

Manuscript Title: Valuing the Child Health Utility 9D: Using Profile Case Best Worst Scaling Methods to Develop a New Adolescent Specific Scoring Algorithm

Social Science and Medicine

Ratcliffe J, Huynh E, Gang C, Stevens K, Swait J, Brazier J, Sawyer M, Roberts R,

Flynn T.

Abstract

In contrast to the recent proliferation of studies incorporating ordinal methods to generate health state values from adults, to date relatively few studies have utilized ordinal methods to generate health state values from adolescents. This paper reports upon a study to apply profile case best worst scaling methods to derive a new adolescent specific scoring algorithm for the Child Health Utility 9D (CHU9D), a generic preference based instrument that has been specifically designed for the estimation of quality adjusted life years for the economic evaluation of health care treatment and preventive programs targeted at young people. A survey was developed for administration in an on-line format in which consenting community based Australian adolescents aged 11 to 17 years (N=1982) indicated the best and worst features of a series of 10 health states derived from the CHU9D descriptive system. The data were analyzed using latent class conditional logit models to estimate values (part worth utilities) for each level of the nine attributes relating to the CHU9D. A marginal utility matrix was then estimated to generate an adolescent-specific scoring algorithm on the full health = 1 and dead = 0 scale required for the calculation of QALYs. It was evident that different decision processes were being used in the best and worst choices. Whilst respondents appeared readily able to choose 'best' attribute levels for the CHU9D health states, a large amount of random variability and indeed different decision rules were evident for the choice of 'worst' attribute levels, to the extent that the best and worst data should not be pooled from the statistical perspective. The optimal adolescent-specific scoring algorithm was therefore derived using data obtained from the best choices only. The study provides important insights into the use of profile case best worst scaling methods to generate health state values with adolescent populations.

1. Background

Adolescence is a key transitional stage of physical and mental human development that is associated with more biological, psychological and social role changes than any other stage of life except infancy (Williams et al.., 2002). Adolescence generally occurs between the ages of 11 and 17 years, commencing at the onset of puberty and terminating at legal adulthood. It represents a period of life when individuals become increasingly responsible for their own health and health care and when several health risk behaviours start to become prevalent, e.g. alcohol use, cigarette smoking and illicit drug use. Adolescence is therefore an important life phase where the introduction of targeted educational and preventative efforts has the potential to impact positively upon both short and long term health status and health related quality of life outcomes (Kleinet, 2007). In 2009, the National Health and Hospitals Reform Commission for Australia produced an influential report highlighting the need for more information in relation to adolescents' attitudes about their own health status and the need to incorporate adolescents' views and preferences into health and public health programmes targeted to meet their needs (National Health and Hospitals Reform Commission, 2009). Such information is an essential prerequisite for the planning and development of preventive strategies and clinical treatment programs designed to improve adolescent health. Despite the production of this report, an acute awareness of the importance of health and public health programmes targeted for this age group and the acknowledgement that (both individually and collectively) adolescents are important consumers of health care in their own right, adolescent health continues to remain a neglected and poorly resourced area of research.

Traditionally within the framework of economic evaluation, health economists and health service researchers have principally sought the views and preferences of adults (aged 18 years and over) to provide information about the relative benefits of competing health care programmes, including those targeted for adolescents (Chen and Ratcliffe, 2015). The most prevalent form of economic evaluation is cost utility analysis (CUA) whereby the benefits of health and public health

programmes are captured through the estimation of quality adjusted life years (QALYs). The QALY recognises the key importance of quality of life in addition to length of life as a defining outcome for assessing the cost effectiveness of health and public health programmes.. As a generic (as opposed to a condition specific) measure, the QALY enables comparisons of the benefit generated from disparate treatment and service programs. Health state values for the calculation of QALY's are generated on a common scale where the endpoint "0" is defined as a state equivalent to being dead and the endpoint "1" is defined as a state equivalent to full health Negative values are also possible for states considered to be worse than being dead. (Brazier et al., 2007). Despite the term CUA the majority of elicitation approaches utilised to derive health state preferences generate values and not utilities. Strictly, only the standard gamble method generates utilities as it incorporates a preference for risk and therefore satisfies the axions of von Neumann-Morgenstern expected utility theory (Mehrez and Gafni 1993).

In recent years, generic preference based instruments have become the most popular mechanisms for the estimation of QALYs within CUA. A recent review of generic preference based instruments utilised in published studies between 2005 and 2010 identified the adult version of the EQ-5D as the most prevalent and widely used internationally, having been translated into 150 languages and applied in 63% of the studies identified (Richardson et al., 2014). Other popular generic preference based instruments applied internationally include the Short Form 6 Dimensions (SF-6D) (Brazier et al., 2002) and the Health Utilities Index (HUI) (Torrance et al., 1996). Whilst these instruments all differ in the way that they describe health and the number and type of included dimensions, they all comprise two common elements. Firstly, a descriptive system for completion by patients or members of the general population comprising a set of items with multiple response categories covering the different dimensions reflecting health status. Secondly, an off the shelf scoring algorithm which reflects society's strength of preference for the health states defined by the instrument. The scoring algorithms are typically generated from large adult general population

surveys to elicit health state values for a selection of health states described by each descriptive system (Brazier et al., 2007).

A recent review by Chen and Ratcliffe (2015) identified nine generic preference based instruments available internationally that have been used in paediatric populations: the Quality of Well-Being Scale (QWB), the Health Utility Index Mark 2 (HUI2), the HUI3, the Sixteen-dimensional measure of health-related quality of life (HRQoL) (16D), the Seventeen-dimensional measure of HRQoL (17D), the Assessment of Quality of Life 6-Dimension (AQoL-6D) Adolescent, the EQ-5D Youth version (EQ-5D-Y), the Adolescent Health Utility Measure (AHUM) and the CHU9D. The majority represent an adaptation of an existing adult instrument and have been valued using adult general population samples. Notable exceptions include the 16D, the AQoL-6D Adolescent and the CHU9D, which have all been valued previously using adolescent samples. Of these, the CHU9D is unique, in that it is the only instrument that does not represent an adaptation of an existing adult instrument, having been developed from its inception with young people (Stevens, 2009). The dimensions of health-related quality of life included within the CHU9D were determined directly from qualitative interviews and analysis with young people using their own language and terminology to describe what quality of life means to them (Stevens, 2009). The original scoring algorithm for the CHU9D is based upon UK adult general population values (n=300) and was generated using the standard gamble (SG) valuation method (Stevens, 2012). A pilot Australian adolescent specific scoring algorithm (aged 11-17 years, n=590) using profile case best worst scaling (BWS) methods has also been developed (Ratcliffe et al., 2012a).

The choice of whose values to use to generate the scoring algorithms for generic preference based instruments applicable for young people may have important policy implications because there is evidence to indicate that adults' preferences for identical health states may differ from adolescents' preferences (Ratcliffe et al., 2015b; Ratcliffe et al., 2012b; Norquist et al., 2008; Saigal et al., 1999) In an empirical comparison of adult versus adolescent specific scoring algorithms for the CHU9D

and the AQoL-6D Adolescent, Ratcliffe et al. (2012b) found notable differences. For the CHU9D instrument, employment of the adolescent algorithms resulted in lower mean health state values than the adult algorithm. For the AQoL-6D, a converse relationship was found. The adolescent values were higher than the corresponding adult values. Although the differences in adult and adolescent values were not consistently found to be in the same direction for both instruments, they were are significant enough to potentially impact upon policy. Employment of the adult or adolescent algorithm for the CHU9D or the AQoL-Adolescent instruments may result in the estimation of differential incremental QALY gains; thereby potentially influencing the decision as to to whether a new healthcare technology targeted for adolescents should be funded or not (Ratcliffe et al., 2012b).

The CHU9D is a generic preference based instrument that has been specifically designed for application with children and adolescents to facilitate the estimation of QALYs for the economic evaluation of health care treatment and preventive programs targeted at young people (Stevens, 2010). The dimensions of HRQoL for inclusion in the CHU9D descriptive system were identified from in-depth qualitative interviews with young people with a variety of chronic and acute health problems (n=74) which aimed to explore how their health affects their lives (Stevens, 2009). The CHU9D has nine attributes: worried, sad, pain, tired, annoyed, schoolwork, sleep, daily routine, ability to join in activities, with five different levels representing increasing levels of severity within each attribute. Whilst it was originally developed for use with younger children aged 7 to 11 years, several recent studies have demonstrated the practicality, face and construct validity of the CHU9D in the Australian adolescent general population (Ratcliffe et al., 2012a; Stevens and Ratcliffe, 2012; Chen et al., 2015). There is increasing interest within Australia and internationally in the application of the instrument with adolescents in the 11-17 year age group and in young adults. The instrument is currently being applied in a number of research programmes internationally focused upon the adolescent age group including the economic evaluation of new innovative adolescent treatment

programs for type 1 diabetes, attention deficit hyperactivity disorder, mental health, obesity prevention and liver transplantation.

In common with traditional discrete choice experiments (DCE) and ranking exercises, profile case BWS is an ordinal approach for health state valuation which offers an attractive option for application with vulnerable population groups including adolescents and older people. It involves a potentially easier choice task to conventional approaches (including time trade off (TTO) and standard gamble (SG)) and traditional DCE. Traditional DCE involves presenting the respondent with a number of choice scenarios in which they are required to indicate their preferences between two or more health states with varying survival durations whereas profile case BWS presents the respondent with a number of choice scenarios represented by one health state only and the respondent is asked to indicate the best and worst attribute of the health state under consideration (Flynn et al. 2008). Previous research by Ratcliffe and colleagues found higher face validity and reliability for BWS methods relative to TTO and SG approaches in young people for the estimation of health state values for the CHU9D instrument (Ratcliffe et al., 2011). However it is important to note as highlighted previously that, of these approaches, only the SG includes a preference for risk and thereby satisfies the von Neumann-Morgenstern axioms of expected utility theory (Mehrez and Gafni 1993). The main objective of the study reported upon in this paper was to build upon the work previously conducted in our pilot study (Ratcliffe et al., 2012a) by utilising profile case BWS methods to generate a new Australian adolescent scoring algorithm for the CHU9D in a much larger and a more representative community based sample of adolescents originating across Australia. The availability of an updated Australian adolescent scoring algorithm will facilitate the systematic incorporation of adolescents' preferences into the health care priorities decision-making process both within Australia and internationally by allowing their values to be captured within CUA for assessing the relative benefits of competing adolescent health and public health programmes.

2. Methods

Study sample

A survey was developed for on-line administration with a community based sample of adolescents, aged 11-17 years, recruited from an Australia wide on-line panel company following parent and adolescent dyad consent for participation. Permission was sought and ethical approval was granted to conduct the study from the Social and Behavioural Research Ethics Committee of Flinders University (Approval no: 5508) . The key objective was to derive adolescent-specific health state values from a large representative sample of adolescents in the general community. A target sample size of N=2000 was considered sufficient to meet the requirements of the profile case BWS experimental design, ensuring precise estimation of model parameters for development of the adolescent specific scoring algorithm for the CHU9D whilst also protecting against any extremes of heterogeneity in preferences.

Survey Instrument

The survey included three main sections. Firstly, respondents were asked to complete the CHU9D instrument. In addition to providing an indicator of HRQoL, completion of the CHU9D helped to familiarise respondents with the wording, formatting and range of each of the 9 attributes of the CHU9D. Secondly, respondents were presented with a series of CHU9D health states for valuation via the profile case BWS task. As it was not feasible to present every possible health state to participants for valuation (the full factorial generates 5^9 = 1,953,125 health states), a fractional factorial design was generated to reduce the number of health states required for presentation. A design that permitted the estimation of main effects, whilst maintaining the properties of near level balance and near orthogonality was generated in 50 health states. Complete orthogonality in the design was not possible due to the need to eliminate a small number of implausible health states (Louviere et al., 2000; Sloane, 2007). In order to promote participant completion rates and minimise

error due to fatigue, the design was blocked into 5 versions so that each respondent was presented with 10 CHU9D health states for valuation using the blocking design principles documented by Hensher and colleagues (Hensher et al., 2005). The 10 CHU9D health states in each block were purposively chosen to include a range of mild, moderate and severe health states.

Each health state description consisted of the nine common dimensions of the CHU9D with different levels for each of the 10 health states presented. Respondents were asked to indicate the best, worst, second best of and then the second worst attributes (dimension levels) of each health state (a screen shot of an example profile case BWS question is presented in Appendix 1); this partial ranking is referred to as a semi-order, because it yields something less than a full ranking. This semi-order was collected with the intention of (1) testing the stability of partial rank orders (best, worst, second best and then the second worst) and (2) determining whether it was possible to pool the data to power the model and improve characterization of individual heterogeneity. The final section of the on-line survey comprised a series of socio-demographic questions including age, gender and additional questions relating to general health status and whether or not the respondent indicated that they were living with a disability or long standing health condition.

Data Analysis

Choice data analysis

In common with all ordinal approaches to health state valuation, profile case BWS data are used in estimated choice models for the sample assuming a random utility theory model. The analysis of choice implies that U_{iq}, the utility respondent q derives from choosing attribute level i, is additively split into an explainable component (V_{iq}) and a random component (ε_{iq}).

For K=9 (representing the 9 CHU9D attributes) attributes each with L_k representing the number of levels of attribute k (representing the 5 levels within each CHU9D attribute), there are a total of K* L_k=45 attribute levels (Flynn et al., 2007). Therefore the equation to be estimated is of the following form:

$$U_{iq}=\Sigma\beta_i X_{iq}+\varepsilon_{iq}$$

where X_{iq} represents the CHU9D attribute levels, and β_i refers to the coefficient for each attribute level to be estimated, i=1,… K* L_k, initially assumed to be constant across individuals. Assuming that the random components are distributed extreme value type 1 (EV1) enables choice data to be analysed using the conditional (multinomial) logit model (Ben-Akiva and Lerman 1985; Louviere et al., 2000):

$$P_{iq}=\exp(\lambda V_{iq})/\Sigma_j \exp(\lambda V_{iq})$$

where P_{iq} is the probability that participant q chooses alternative i, j represents all the relevant alternatives in choice set C (i.e. the descriptive dimensions of a health state), and λ represents the Extreme Value 1 scale parameter which is inversely proportional to the standard deviation of the random component ε_{iq}.

In any data-set where the estimation of one or more parameters is heavily influenced by a very small number of observations, this can lead to mis-specification of the fitted model through incorrect characterization of the underlying data generation process (Cook R, 1977). The removal of outliers had the objective of removing observations that have extreme impact on the aggregate estimates although the likelihood of them belonging to the same distribution is small. Respondents that exhibited extreme atypical trade-offs and marginal rates of substitution (swaying the conditional (MNL) logit model parameter estimates by three standard deviations) compared to the rest of the sample were identified and removed using jackknife resampling (further details are available from the authors upon request) (Babu, 2011). Conditional logit (MNL) regression models were then estimated on the remaining dataset for the prediction of CHU9D health state values for each of the choice measures: best, worst, second best and second worst, separately. To account for the sequential nature in which choices were made in the task, all models were estimated on the reduced set of options that were presented, that is, best choice among i=9 dimensions and worst choice of the remaining i-1 dimensions.

Testing for the pooling of different choice rankings

The collection of the semi-order of best and worst choice data in the profile case BWS task provided additional information on adolescent preferences and allowed for the possibility to combine or augment the data to include all choices to estimate the attribute level utilities. As is the case for exploded ranking data, models estimated from different ranks may not be pooled if both variance scale factors and parameters differ by rank level (Ben-Akiva et al., 1992). However, if adolescent preferences are similar across the different choices but vary in their error variance such that they are more or less consistent in making choices, then it is possible to pool the data sources with appropriate accounting for scale differences across context to estimate the attribute level utilities (Swait and Louviere, 1993).

As an initial informal investigation of whether data pooling was feasible, the plots of the part-worth utilities were compared. Swait and Louviere (1993) show that for the MNL model, under the hypothesis of preference homogeneity and scale differences between two data sources, plotting the preference parameters on a X-Y plot should result in proportional and positively sloped distribution points, whereby the slopes are related to the ratio of the scale factors in the two choice data sources. Thus, if the data sources can be pooled, then we would expect the beta coefficient estimates to be roughly proportional across data sources. (See Swait and Bernardino, 2000 for the introduction of this informal method across multiple segments.) The hypothesis was tested that the parameters from models estimated for different choice measures were the same, and scales between the datasets were allowed to vary for logical pairs of choice measures: best with worst; best and second best; worst and second worst. The chi-square test compared the sum of the log-likelihood from the MNL models for each choice context and the log-likelihood function from a heteroscedastic conditional (multinomial) logit model (Louviere and Swait, 1996) that adjusted for scale difference across choice measures for the combined data-set.

Latent class analysis

Latent Class models identify and cluster "types" of participants who are similar in terms of their relative preferences (Flynn et al., 2010). The behavioural choice model was assumed to be a logit model, and the preference distribution was a discrete finite mixture of logit models assumed to comprise types of participants exhibiting similar part-worth utilities and/or scale. Maximum likelihood estimates were used to classify into clusters based upon their posterior probability of class membership. The Bayesian Information Criterion (BIC) criterion was used to help guide model selection and stability of solutions was also used to select the optimal model (Hensher, 2012). The EM (Expectation-Maximization) optimization algorithm cannot guarantee that a set of parameter values globally maximizes the log-likelihood. Thus, different starting seeds for the algorithm were considered to ensure with a reasonable degree of confidence that a global maximum had been reached. The final reported model was the optimal class selection, re-estimated to include only statistically significant coefficients.

Adolescent scoring

Finally, the heterogeneity-adjusted population level scoring algorithm was estimated by producing a single set of beta coefficients corresponding to the adolescent population average preferences for the attribute levels relating to the CHU9D. The average scores across all participants were calculated, by taking the mean of the sets of preference class estimates, weighted by probability of class membership across the sample. A linear transformation was applied to the attribute level estimates to ensure that the sum of the relevant nine scores (one chosen level per attribute) were reflected on a 0-1 scale. In common with all ordinal approaches to health state valuation, the estimates obtained from the profile case BWS were not based on the 0-1 dead full health QALY scale. Since these estimates were on an interval scale, re-scaling via an external cardinal valuation task using traditional health state valuation methods, e.g., the TTO or SG, was necessary to ensure

that 0 represented the death state. Whilst it would be ideal to conduct the re-scaling using data generated from an adolescent sample, our previous research has highlighted the ethical difficulties associated with this process (Ratcliffe et al., 2011). In addition we found a poor level of understanding of TTO and SG methods in general in adolescents (Ratcliffe et al., 2011). Hence, for the purposes of this study we utilised the mean health state values derived from a conventional TTO task for a selection of CHU9D health states from a sample of young adults (aged 18 to 29 years) to re-scale the ordinal values derived from the BWS DCE task onto the 0 = death to 1 = full health QALY scale (Ratcliffe et al., 2015a). Two rescaling approaches were utilised and compared in terms of overall model fit and mean absolute errors (MAE) to determine the optimal approach. The first approach (Method 1) used the mean TTO PITS health state value only, whilst the second mapping approach (Method 2) used ordinary least squares regression with TTO derived health state values for a selection of CHU9D health states (ranging from mild impairment to the PITS state) to rescale the raw scores generated from the profile case BWS task (Rowen et al., 2015).

3. Results

Sample Characteristics

Data collection for the profile case BWS task was conducted over a two month period from October to November 2012 for a sample of adolescents aged 11-17 years in the Australian population. The completion rate for the survey was 19%, with N=2076 of the total sample of consenting respondents fully completing the survey, out of a total pool of 10,928 individuals initially approached. Respondents were randomly assigned into the five survey versions. The sampling was programmed to cease as soon as there were least 400 respondents in each version of the survey. The version with the smallest number of respondents had a sample size of N=404. We then removed the last observations that entered the survey for each of the remaining versions (using actual date and time of completion) to achieve a balanced sample (N=404) across each of the 5 versions of the survey, such that complete data from 2020 adolescents were obtained.. The model estimates were not

sensitive to the removal of these respondents. On average, respondents took a median time period of 12.2 minutes to complete the on-line survey.

The age and gender distributions from the adolescent sample were compared with the wider population of Australian adolescents using the 2011 Census ABS data (Pink, 2012). The characteristics of the study sample are presented in Table 1. It can be seen that whilst the study sample was generally representative of the wider Australian population aged 11-17 years in terms of gender, the study sample comprised a greater proportion of older adolescents with 18.8% aged 17 years compared to 14.5% of the wider population. Table 2 presents the health characteristics of the respondents. A minority (12.3%) indicated that they were living with a long standing illness or disability. As expected with a community based sample, a relatively small proportion of participants reported themselves as living with poor health (0.9%) with larger proportions of respondents reporting themselves as living in Excellent (29.5%), Very good (42.3%) or Good health (22.1%). The responses to the CHU9D are presented in Table 3. Respondents also generally reported themselves in good health according to the CHU9D descriptive system, with N=184 reporting themselves at full health (reflecting the highest or best level for all nine CHU9D dimensions). No participants reported themselves in the PITS state (reflecting the lowest or worst level of impairment for all nine CHU9D dimensions).

Choice data analysis

From the initial sample of 2020 individuals with complete responses, the jackknifing exercise identified N=38 individuals whose inclusion leads to conditional (multinomial) logit model estimates that are more than +/- 3 standard deviations from the aggregate model counterpart. These 38 individuals were then removed from the dataset, reducing the final useable sample to N=1982. Figure 1 plots the MNL coefficient estimates across pairs of choice measures for each CHU9D dimension. For comparison, reverse coding (-1) was applied on the worst and second worst results

presented in Figure 1. Whilst all of the plots were positively sloped, they were not linearly proportional. The greatest difference was highlighted in the plot for best against worst where there appeared to be less differentiation amongst the lower levels on the best data compared to the worst data, particularly for the mental health dimensions, '*Worried*', '*Sad*' and '*Annoyed*'. The lowest level of the '*Activities*' dimension was differentiated as being worst compared to all other levels, and there was no discernible distinction between the remaining levels (levels 2-5) for this particular dimension. Overall, the results are consistent with the majority of respondents choosing the best level of a CHU9D dimension when it was presented within a given health state.

Pooling of choice types

Table 4 presents the Swait and Louviere (1993) test for pooling various pairs of choice measures: best, worst, second best and second worst (Cases A-E). Scale was specified as $\lambda=\exp(Z_q\theta)$ where Z_q is dummy indicator for the data source, and θ was the parameter to be estimated. The pooling test of best and worst data produced a chi-squared =2020.23 with 45 degrees of freedom, therefore rejecting the hypothesis that the parameters across datasets were the same whilst permitting the scale factor to vary. Similarly the hypothesis was rejected across the other pairs of choice measures at the 95% confidence level. The results suggest the differences are not only attributable to variance scale but also differences in preferences between the choice measures among adolescents, providing evidence against pooling of the different data sources. The pooling test was further relaxed to allow for partial preference heterogeneity across data sources or attributes of the CHU9D, thus allowing more noise among some attributes by data source (Swait and Bernardino, 2000). Parameters chosen to be freed included specific attribute level parameters and entire attributes with particular focus on worst and second worst data in which the test statistic was smallest. Whilst allowing for partial preference heterogeneity reduced the chi-squared statistic, the reduction was not significant enough to allow for the (partial) pooling of data. It was therefore concluded that the final scoring of the adolescent values of the CHU9D should be based on one single choice measure assessed to be most appropriate for the task. Consistent with the traditional choice literature and traditional DCEs in

which we typically associate choices as reflecting underlying values (Ryan et al., 2008), the best choices have therefore been utilised in the development of an updated Australian adolescent specific scoring algorithm for the CHU9D.

Latent class analysis of best data

Latent class analysis on the best choice data led to the selection of the two class solution; the coefficients for each class are provided in Columns (1) and (2) of Table 5. This model was characterized by strongest preferences for the following:

- Class 1 (size=63.2%) – most importance placed on mental health dimensions including *Worried* and *Annoyed*, and least importance placed on daily activities such as *Activities, Daily routine*, *Sleep*.
- Class 2 (size=36.8%) – equal weights on all attributes: valued the top level of every CHU9D attribute most highly and appear largely insensitive to the remaining levels.

Socio-demographic variables were included in the class membership model to characterise the classes. Wald and likelihood ratio test statistics indicated that none of the included covariates (age, gender, disability, number of cars, level of difficulty, health slider value) were statistically significant at the 5% level in predicting the class membership probabilities. This supports an interpretation that individual heterogeneity arises through the behaviour surrounding evaluation of the health state dimensions rather than arising from systematic individual differences due to age, gender, etc.

The scores based upon the latent class model for the best data are presented in Column (3) of Table 5. The scores are weighted averages of the class segments (Flynn et al., 2015). The scores presented in Column (3) of Table 5 are anchored to the least valued attribute level. The scores indicate that all nine attributes make a contribution to an individual's HRQoL as classified by the CHU9D, with the *Worried* and *Activities* dimensions having the largest effect on the overall scores.

Collapsing of levels and scores

Consistent with the health state valuation literature, it was expected *a priori* that lower health state values would be associated with greater impairments amongst levels of the same CHU9D attribute or dimension. Upon calculation of the scores for the CHU9D using the latent class analysis, a number of inconsistencies in coefficient values were noted across dimension levels. The first example of such an inconsistency was observed with the *Worried* attribute in which the 4th level was valued more highly than the 3rd level. Similar inconsistencies are identified and bolded in Column (3) of Table 5. Inconsistencies may represent CHU9D dimension levels that are not statistically different from each other and therefore signify a need to collapse specific levels for particular dimensions and present them as a single (combined) level. In such cases it is more accurate and reliable to collapse levels of attributes. Previous large scale valuation studies for other generic preference based instruments with relatively large descriptive systems, e.g., the UK valuation of the SF-6D, have identified similar levels of attribute level inconsistencies to those found in this study and have also adopted this approach (Brazier et al., 2002).

The collapsing of levels was determined by imposing restrictions to the original model presented in Table 5. Parameters were restricted to be equal for chosen levels of attributes at the class level in the model. As previously indicated, the scores represent weighted averages of the class segments. The selection of which levels to equate was defined so as to satisfy the following four criteria:

1. Attribute level coefficients for the same dimension that were not statistically different were restricted to be equal. All restriction and tests were performed by class. Statistical significance was based on t-tests performed on the associated levels from the original latent class model.
2. Monotonicity was achieved for each attribute at the aggregate level. That is, lower health state values would be associated with greater impairments amongst levels of the same

CHU9D attribute or dimension.

3. The new model was not statistically different from the final non-restricted model in Table 5. (A log-likelihood ratio test was used to test for statistical differences between the models).
4. An overall good model fit in terms of the Rho-squared and the BIC.

Criteria 1 and 2 were directly imposed, while criteria 3 and 4 were utilised post hoc to validate the specification of the final model.

Scoring with collapsed levels

It was not possible to identify a model which satisfied all four criteria simultaneously. Equating the non-significantly different attribute levels (criterion 1) produced a model that also satisfied criteria 3 and 4 but failed to satisfy monotonicity for the *Tired* attribute (results are available from the authors upon request). Ensuring monotonicity for all attributes resulted in the model presented in Table 6. This model satisfied criteria 1, 2 and 4 but failed to satisfy criterion 3 because it was found to be statistically different to the original latent class model. However, this model produced a better model fit in terms of the BIC to both the first and main (non-restricted) model. The final column (3) of Table 6 presents the raw scores for the CHU9D based upon the new latent class model where monotonicity is ensured.

Rescaling onto the QALY scale

The raw scores for the CHU9D were rescaled onto the QALY scale by utilising the mean TTO values derived for a selection of CHU9D health states from a sample of young adults. Details of the methods and findings from the TTO study are provided elsewhere (Ratcliffe et al., 2015a). It is notable that the mean PITS health state value from the TTO study (-0.21) was significantly lower than the mean PITS health state score generated from application of the original adult scoring algorithm which utilised the SG approach (0.34). Table 7 presents two groups of rescaled BWS

DCE estimates corresponding to the mean TTO health state values for the four selected CHU9D health states, as well as the goodness-of-fit MAE values. It is evident that rescaling the profile case BWS estimates using the mapping approach (Method 2) exhibited the best performance (i.e. lowest MAE) in this dataset. The scatter plot between the preferred rescaled BWS estimates (Method 2) and TTO health state values is presented in Figure 2.

4. Discussion

This study provides important insights into the use of profile case best worst scaling (BWS) methods to generate health state values with adolescent populations. The findings indicate that the cognitive decision processes adolescents use to make 'best' and 'worst' choices respectively may be quite different. In this study it was not possible to combine the choice data (best, worst, second best and second worst) to provide more information about preferences and thereby improve the estimation of the coefficients attached to attribute levels. Consistent with conventional discrete choice experiments (DCEs) in which we typically associate choices as reflecting underlying values, the best choices were therefore utilised to develop the updated community based adolescent scoring algorithm for the CHU9D.

The latent class modelling identified two key groups of respondents characterised according to their underlying preferences. The first key group represented the majority of respondents and tended to place more weight on the CHU9D attributes relating to mental health impairments relative to those reflecting daily activities and/or physical health impairments. These findings are consistent with those from our previous study utilising the pilot adolescent scoring algorithm for the CHU9D which also highlighted that adolescents tend to place more importance upon mental health impairments than adults (Ratcliffe et al., 2012a; Ratcliffe et al., 2015b). The second key group of respondents tended to value the top (no impairment) level of each of the CHU9D attributes most highly and

appeared largely insensitive to the remaining levels (2 to 5) reflecting increasing degrees of impairment. The reasons for this type of choice behaviour are unclear but may be reflective of the use of decision heuristics (or short cuts) to simplify the choice task (Lloyd, 2003). More substantively, however, this may indicate that non-compensatory decision making is occurring; this gives rise to potentially very interesting alternative models of decision making to evaluate health states (Lancsar and Swait 2013). This type of choice behaviour also has important implications for health care policy as it implies a reduced likelihood of finding significant QALY differences between groups where incremental changes are observed between degrees of impairment (specifically between levels 2 to 5 of the CHU9D instrument) over time. Relative to other popular generic preference based instruments, in particular the EQ5D, the CHU9D has a relative large descriptive system and it may be that the presentation of nine attributes simultaneously within a single health state was cognitively challenging to the extent that these respondents opted to focus only upon a limited number of attributes to make their choice decisions. It is also possible that the use of an on-line mode of administration may have reduced the level of engagement for some respondents and therefore increased the likelihood of the application of decision heuristics relative to an interviewer administered mode of administration.

A further possible explanation for the lack of differentiation between CHU9D attribute levels beyond the best level may be a reflection of the utilisation of a community based sample of largely healthy adolescents. It is likely that the majority of these individuals had little or no previous experience of, and therefore found it difficult to imagine living with, moderate or severe health impairments. As such, their preferences were largely insensitive to increasing degrees of impairment. Further research including qualitative ‘think aloud’ approaches would be helpful in this regard in determining a detailed examination of adolescent respondents understanding and level of engagement with the profile case BWS task (Whitty et al., 2014).

In general, the mean health state values generated from application of the updated adolescent specific scoring algorithm for the CHU9D are lower than the two previous scoring algorithms generated for this instrument, i.e., [1] the original scoring algorithm based upon application of the SG method in the UK with adults of all ages (age range: 16 to 87 years) (Stevens, 2012) and [2] the pilot Australian adolescent specific scoring algorithm (age range: 11 to 17 years) (Ratcliffe et al., 2012a). The main reason for these differences is likely largely due to the differences in the mean cardinal health state values utilised for rescaling. The mean PITS health state value generated from the TTO study with young adults (-0.21) was significantly lower than the mean PITS health state score generated from application of the original adult scoring algorithm utilising the SG approach (0.34). The pilot Australian adolescent specific scoring algorithm was also generated using profile case BWS methods to generate raw scores. These scores were then rescaled using the mean PITS health state value from the original adult scoring algorithm based upon the SG approach (Ratcliffe et al., 2012a). In contrast the updated adolescent specific scoring algorithm reported in this paper was developed using a mapping approach to rescaling, involving TTO values from a series of four CHU9D health states reflecting increasing health impairments and including the PITS State. The TTO derived health state values were noticeably lower than the corresponding values for identical CHU9D health states generated using SG. In addition, a significant proportion of young adult participants considered the PITS state to be worse than death when directly valuing it using the TTO method. Hence, overall the mean health state value for the PITS state indicated that this state was considered worse than death (Ratcliffe et al., 2015a). Overall, these findings are consistent with evidence from the literature to indicate that the SG method tends to bias health state values upwards relative to the TTO method due to probability weighting (the tendency for individuals to overweight small probabilities and underweight large probabilities) and loss aversion (a tendency to be more sensitive to losses than to gains) (Bleichrodt, 2002).

This study raises important questions and adds to the debate in the literature about whose values should be used in valuing health states for economic evaluation. Whilst the study sample was large and generally representative of the wider Australian population aged 11-17 years in terms of gender, it also contained a greater proportion of older adolescents compared to the wider population. The sample may not, therefore, be considered as entirely representative of the adolescent population of Australia. In addition, it is possible that the utilisation of a predominantly healthy sample of adolescents to value CHU9D impairment states contributed to the apparent insensitivities and lack of differentiation at the lower levels and the apparent adoption of different decision rules for the identification of best and worst attribute levels. Further research including similar valuation studies conducted in adolescent patient samples with more direct experience of health impairments would be helpful in indicating the effects of experience and whether or not this facilitates more differentiation at the lower levels and the adoption of more consistent decision rules. A variant of this suggestion is to use stratified sampling on the basis of health states, and weight appropriately to the population level. This would ensure adequate representation of the full spectrum of health states to enable population predictions, while permitting more accurate inferences within health state.

An argument often propagated in favour of using adult general population preferences for health states for incorporation into economic evaluation is that adults in the general population are eligible to pay general taxation which provides financial support for the health systems of many countries (Gunning, 2003). This argument appears to be at the root of the guidance to health technology appraisal provided by the National Institute for Health and Care Excellence (NICE) in the UK (NICE, 2013) and that of other regulatory authorities (Brazier et al., 2007). However, a converse, and potentially more compelling argument which forms the underlying premise of the work presented in this paper, is that the incorporation of the preferences of adolescents into cost-effectiveness analyses of health and public health programmes designed for this age group has the potential to facilitate the development of programmes that are more relevant to their needs,

ultimately leading to improvements in adolescent health as a consequence of improved treatment compliance and service utilisation.

5. Conclusion

This study has provided important insights into the use of profile case best worst scaling methods to generate health state values with adolescent populations. Post hoc it is evident that different decision processes may underlie the observed best and worst choices, so it was decided that the optimal adolescent specific scoring algorithm should be derived using the best choices. The availability of an updated adolescent specific scoring algorithm for the CHU9D will enable the health state values of a large community based sample of Australian adolescents to be incorporated directly into economic evaluation studies through calculation of the incremental QALY gains associated with new treatment and preventive programs targeted for this age group. The new updated adolescent specific scoring algorithm will facilitate the systematic incorporation of the views of young people into the health care priorities decision-making process both within Australia and internationally, with the ultimate aim of improving the health of the adolescent population through the development of cost effective treatment and preventive programs whose effectiveness is defined to incorporate the needs and preferences of adolescents.

References

Babu, G., 2011. Resampling methods for model fitting and model selection. Journal of Biopharmaceutical Statistics, 21, 1177-86.

Ben-Akiva, M., and Lerman, S. (1985) Discrete Choice Analysis: Theory and Application to Predict Travel Demand, Cambridge: MIT Press.

Ben-Akiva, M., Morikawa, T., Shiroishi, M., 2002. Analysis of the Reliability of Preference Ranking Data Journal of Business Research, 24, 149-64.

Bleichrodt, H., 2002. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Economics, 11, 447-56.

Brazier, J., Ratcliffe, J., Salomon, J., Tsuchiya, A., 2007. Measuring and valuing health benefits for economic evaluation, Oxford University Press.

Brazier, J., Roberts, J., Deveril, M., 2002. The estimation of a preference based measure of health from the SF-36. Journal of Health Economics, 21, 271-292.

Chen, G., Flynn, T., Stevens, K., Brazier, J., Huynh, E., Sawyer, M., Roberts, R., Ratcliffe, J., 2015.. Assessing the health related quality of life of Australian adolescents: an empirical comparison of the EQ-5D-Y and CHU9D instruments. Value in Health, 18, 432-438.

Chen, G., Ratcliffe, J., 2015. A review of the development and application of generic multi-attribute utility instruments for paediatric populations. Pharmacoeconomics, 33: 1013-28.

Flynn, T., Huynh, E., Peters, T., Al-Janabi, H., Clemens, S., Moody, A., Coast, J., 2015. Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. Health Economics, 24, 258-269.

Flynn, T., Louviere, J., Marley, A., Peters, T., Coast, J., 2008. Rescaling quality of life tariffs from discrete choice experiments for use as QALYs: a cautionary tale. Population Health Metrics, 6, 6.

Flynn, T., Louviere, J., Peters, T., Coast, J., 2010. Using discrete choice experiments to understand preferences for quality of life. Variance-scale heterogeneity matters. Social Science and Medicine, 70, 1957–196.

Gunning, J., 2003. Understanding Democracy: An Introduction to Public Choice. Nomad press, Taiwan.

Hensher, D., Rose, J., Greene, W., 2007. Applied Choice Analysis, A Primer. Cambridge University Press, Cambridge.

Hensher, D.A., 2012. Accounting for scale heterogeneity within and between pooled data sources. Transportation Research Part A Policy and Practice, 46, 480-6.

Kleinet, S., 2007. Adolescent health: an opportunity not to be missed. The Lancet, 369, 1057-1058.

Lancsar, E., Swait, J., 2014. Reconceptualising the External Validity of Discrete Choice Experiments, PharmacoEconomics, 32(10), 951-965.

Lloyd, A., 2003. Threats to the estimation of benefit: are preference elicitation methods accurate? Health Economics, 12, 393-402.

Louviere, J., Hensher, D., Swait, J., 2000. Stated Choice Methods: Analysis and Application. Cambridge University Press, Cambridge.

Louviere, J., Swait J., 1996. Searching for regularities in choice processes, or the little constant that could. Working Paper, Dept. of Marketing, Faculty of Economics, University of Sydney.

Mehrez, A., Gafni, A., 1993. HYEs versus QALYs: in pursuit of progress. Medical Decision Making, 13, 287–92.

National Health and Hospitals Reform Commission., 2009. A Healthier Future for all Australians-Final Report of the National Health and Hospitals Reform Commission. Commonwealth of Australia, Canberra.

National Institute for Health and Care Excellence., 2013. Guide to the Methods of Technology Appraisal, NICE, London..

Norquist, G., McGuire, T., Essock, S., 2008. Cost effectiveness of depression treatment for adolescents. American Journal of Psychiatry, 165, 549-552.

Pink, B., 2012. Census of Population and Housing, - Products and Services 2011. Australian Bureau of Statistics, Canberra.

Ratcliffe, J., Chen, G., Stevens, K., Bradley, S., Couzner, L., Brazier, J., Sawyer, M., Roberts, R., Huynh, E., Flynn, T., 2015a. Valuing Child Health Utility 9D Health States with Young Adults: Insights from A Time Trade Off Study. Applied Health Economics and Health Policy, 13, 485-92.

Ratcliffe, J., Flynn, T., Huynh, E., Stevens, K., Brazier, J., Sawyer, M., 2015b. Nothing about us without us? A comparison of adolescent and adult health state values for the Child Health Utility-9D using profile case best worst scaling. Health Economics Feb 16. doi: 10.1002/hec.3165, [Epub ahead of print].

Ratcliffe, J., Flynn, T., Terlich, F., Brazier, J., Stevens, K., Sawyer, M., 2012a. Developing adolescent specific health state values for economic evaluation: an application of profile case best worst scaling to the Child Health Utility-9D. Pharmacoeconomics, 30(8), 713-27.

Ratcliffe, J., Stevens, K., Flynn, T., Brazier, J., Sawyer, M., 2012b. Whose values in health? An empirical comparison of the application of adolescent and adult values for the CHU9D and AQOL-6D in the Australian adolescent general population. Value in Health, 15(5), 730-736.

Ratcliffe, J., Couzner, L., Flynn, T., Sawyer, M., Stevens, K., Brazier, J., Burgess, L., 2011. Valuing Child Health Utility 9D health states with a young adolescent sample: a feasibility study to compare best-worst scaling discrete-choice experiment, standard gamble and time trade-off methods. Applied Health Economics and Health Policy, 9(1),15-27.

Richardson, J., McKie, J., 2014. Multi attribute utility instruments and their use. In: Culyer AJ (ed.) Encyclopedia of Health Economics. San Diego: Elsevier Science, 341–357.

Rowen, D., Brazier, J., Van Hout, B., 2015. A comparison of methods for converting DCE Values onto the Full Health-Dead QALY scale. Medical Decision Making, 35, 328-40.

Ryan, M., Gerard, K., Amaya-Amaya, M., 2008. Using discrete choice experiments to value health and health care (Vol. 11). Dordrecht: Springer.

Saigal, S., Stoskopf, B., Feeny, D., Furlong, W., Burrows, E., Rosenbaum, P., Hoult, L., 1999. Differences in preferences for neonatal outcomes among health care professionals, parents and adolescents. Journal of the American Medical Association. 281, 1991-1997.

Sloane, N., 2007. A library of orthogonal arrays. Available at : http://neilsloane.com/oadir/.2007.

Stevens, K., Ratcliffe, J., 2012. Measuring and valuing health benefits for economic evaluation in adolescence: an assessment of the practicality and validity of the Child Health Utility 9D in the Australian adolescent population. Value in Health, 15, 1092-9.

Stevens, K., 2009. Developing a descriptive system for a new preference-based measure of health related quality of life for children. Quality of Life Research, 18, 1105-1113.

Stevens, K., 2012. Valuation of the Child Health Utility 9D Index. Pharmacoeconomics, 30, 729-747.

Stevens, K., 2010. Working with children to develop dimensions for a preference-based generic paediatric health related quality of life measure. Qualitative Health Research, 20, 340-351.

Swait, J., Bernardino, A., 2000. Distinguishing taste variation from error structure in discrete *choice* data. Transportation Research Part B, 34, 1-15.

Swait, J., Louviere, J., 1993. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. Journal of Marketing Research, 30, 305–314.

Torrance,G., Feeny, D., Furlong, W., Barr, R., Zhang, Y., Wang, Q., 1996. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Medical Care, 34, 702-22.

Whitty, J., Walker, R., Golenko, X., Ratcliffe, J., 2014. A think aloud study comparing the validity and acceptability of discrete choice and best worst scaling methods. PLoS One, 23, 9(4):e90635.

Williams, P., Holmbeck, G., Greenley, R., 2002. Adolescent Health Psychology. Journal of Consulting and Clinical Psychology, 70, 828-842.

Table 1: Characteristics of participants

	Frequency (%)	Percent*(ABS)
Gender:		
Male	996 (49.3)	48.63
Female	1024 (50.7)	51.37
Age at survey completion:		
11 years	262 (13)	14.02
12 years	239 (11.8)	14.13
13 years	234 (11.6)	14.12
14 years	266 (13.2)	14.26
15 years	329 (16.3)	14.32
16 years	310 (15.3)	14.63
17 years	380 (18.8)	14.52
Total	2020 (100)	100

*Percent of adolescents between ages 11 and 17 years (inclusive)

Table 2: Health characteristics of participants

	Frequency (%)
Do you have a long-term disability, illness or medical condition?	
Yes	249 (12.3)
No	1771 (87.7)
In general would you say your health is:	
Excellent	596 (29.5)
Very good	854 (42.3)
Good	447 (22.1)
Fair	104 (5.1)
Poor	19 (0.9)
Total	2020 (100)

Table 3: CHU9D responses	Frequency (%)
Worried	
I don't feel worried today	991 (49.1)
I feel a little bit worried today	669 (33.1)
I feel a bit worried today	257 (12.7)
I feel quite worried today	78 (3.9)
I feel very worried today	25 (1.2)
Sad	
I don't feel sad today	1315 (65.1)
I feel a little bit sad today	464 (23)
I feel a bit sad today	172 (8.5)
I feel quite sad today	52 (2.6)
I feel very sad today	17 (0.8)
Pain	
I don't have any pain today	1170 (57.9)
I have a little bit of pain today	596 (29.5)
I have a bit of pain today	195 (9.7)
I have quite a lot of pain today	45 (2.2)
I have a lot of pain today	14 (0.7)
Tired	
I don't feel tired today	418 (20.7)
I feel a little bit tired today	899 (44.5)
I feel a bit tired today	380 (18.8)
I feel quite tired today	226 (11.2)
I feel very tired today	97 (4.8)
Annoyed	
I don't feel annoyed today	1169 (57.9)
I feel a little bit annoyed today	536 (26.5)
I feel a bit annoyed today	199 (9.9)
I feel quite annoyed today	90 (4.5)
I feel very annoyed today	26 (1.3)
Schoolwork/Homework	
I have no problems with my schoolwork/homework today	1046 (51.8)
I have a few problems with my schoolwork/homework today	644 (31.9)
I have some problems with my schoolwork/homework today	223 (11)
I have many problems with my schoolwork/homework today	80 (4)
I can't do my schoolwork/homework today	27 (1.3)
Sleep	
Last night I had no problems sleeping	1101 (54.5)
Last night I had a few problems sleeping	629 (31.1)
Last night I had some problems sleeping	196 (9.7)
Last night I had many problems sleeping	76 (3.8)
Last night I couldn't sleep at all	18 (0.9)
Daily Routine	
I have no problems with my daily routine today	1506 (74.6)
I have a few problems with my daily routine today	355 (17.6)
I have some problems with my daily routine today	121 (6)
I have many problems with my daily routine today	28 (1.4)
I can't do my daily routine today	10 (0.5)
Able to join in activities	
I can join in with any activities today	1336 (66.1)
I can join in with most activities today	400 (19.8)
I can join in with some activities today	136 (6.7)
I can join in with a few activities today	96 (4.8)
I can join in with no activities today	52 (2.6)
Total	2020 (100)

Table 4 - Pooling test of choice measures

Case	**A**	**B**	**C**	**D**	**E**
Hypothesis (H1A)*	$B_{best}=B_{worst}$	$B_{sec\ best}=B_{sec\ worst}$	$B_{worst}=B_{sec\ worst}$	$B_{worst}=B_{sec\ worst}$	$B_{worst}=B_{sec\ best}$
LL (best)	-34339.62		-34339.62		
LL (worst)	-37919.74			-37919.74	-37919.74
LL (second best)		-33830.2	-33830.2		-33830.2
LL (second worst)		-34294.64		-34294.64	
LLu (pooled)**	-73269.47	-68476.79	-69025.16	-72360.45	-72636.36
θ***	0.43	0.67	-0.33	0.53	0.18
Chi2 (45 df)	2020.23	703.9	1710.68	292.14	1772.84
Reject H1A?	YES	YES	YES	YES	YES

* While permitting scale to vary.

**Log-likelihood from the heteroscedastic conditional logit model of the pooled data. The scale parameter is specified as $\lambda=\exp(Z_q\theta)$ where Z_q represents an indicator for choice task, and θ is the parameter to be estimated. Z_q is indicator for data source such that the first data source is given +1 and -1 for the alternate data source, e.g.) for Case A Zq is a best-worst indicator: 1 for best data and -1 for worst data.

*** The parameter in the scale function is significant at the 1% level of significance.

Table 5: Latent class model with adolescent scoring – main model with no restrictions.

.		(1) Class 1		(2) Class2		(3) SCORE
Worried	Level 1	5.67	***	1.5	***	0.2361
	Level 2	1.362	***	1.337	***	0.1503
	Level 3	1.463	***	1.155	***	**0.1314**
	Level 4	1.587	***	1.164	***	**0.1343**
	Level 5	0		1.299	***	0.1247
Sad	Level 1	5.357	***	0.844	***	0.1573
	Level 2	1.343	***	0.527	***	0.0587
	Level 3	1.56	***	0.448	***	0.0532
	Level 4	0		0.477	***	0.0321
	Level 5	0		0.38	***	0.0212
Pain	Level 1	5.15	***	0.432	***	0.1076
	Level 2	1.27	***	0.305	***	0.0326
	Level 3	1.252	***	0.307	***	0.0325
	Level 4	0		0		-0.0217
	Level 5	0		0		-0.0217
Tired	Level 1	4.89	***	0.227	**	**0.0805**
	Level 2	3.063	***	0.655	***	**0.1001**
	Level 3	2.363	***	0.597	***	0.0826
	Level 4	1.821	***	0.522	***	0.0657
	Level 5	1.328	***	0.243	**	0.0265
Annoyed	Level 1	4.312	***	0.294	**	0.079
	Level 2	1.463	***	0		0.0012
	Level 3	1.366	***	0		-0.0003
	Level 4	0.65	**	0		-0.0115
	Level 5	0		-0.202	*	-0.0444
Schoolwork	Level 1	5.197	***	0.305	***	0.0941
	Level 2	2.037	***	0		0.0102
	Level 3	1.584	***	0		0.0031
	Level 4	0.688	**	0		-0.0109
	Level 5	1.58	***	-0.247	**	-0.0248
Sleep	Level 1	5.383	***	0		0.0626
	Level 2	1.635	***	0		0.0039
	Level 3	1.795	***	0		0.0064
	Level 4	0.856	***	0		-0.0083
	Level 5	0		-0.298	***	-0.0552
Daily Routines	Level 1	5.497	***	0		0.0644
	Level 2	1.995	***	-0.479	***	-0.0444
	Level 3	1.791	***	-0.543	***	-0.0548
	Level 4	0.779	***	-0.6	***	**-0.0771**
	Level 5	1.319	***	-0.418	***	**-0.0481**
Activities	Level 1	5.902	***	0.424	***	**0.1185**
	Level 2	5.273	***	0.618	***	**0.1305**
	Level 3	4.837	***	0.308	***	0.0888
	Level 4	4.581	***	0.265	***	0.0799
	Level 5	2.776	***	0		0.0218
Class membership		0.5521	***	0		
Posterior class probabilities		0.635		0.365		
Log-likelihood		-30638.91				
BIC(LL)		61809.24				
Npar		70				

***,**,* significant at the 1,5,10% level of significance

Table 6 – Model with collapsed levels and monotonicity

		(1) Class 1		(2) Class2		(3) SCORE
Worried	Level 1	5.687	***	1.504	***	0.2163
	Level 2	1.364	***	1.334	***	0.1326
	Level 3	1.55	***	1.155	***	0.118
	Level 4	1.55	***	1.155	***	0.118
	Level 5	0		1.296	***	0.1077
Sad	Level 1	5.376	***	0.849	***	0.1475
	Level 2	1.35	***	0.524	***	0.0532
	Level 3	1.578	***	0.447	***	0.0492
	Level 4	0		0.474	***	0.0274
	Level 5	0		0.376	***	0.0178
Pain	Level 1	5.168	***	0.436	***	0.1039
	Level 2	1.282	***	0.305	***	0.0308
	Level 3	1.265	***	0.305	***	0.0305
	Level 4	0		0		-0.0189
	Level 5	0		0		-0.0189
Tired	Level 1	4.903	***	0.51	**	0.107
	Level 2	3.085	***	0.51	***	0.0788
	Level 3	2.376	***	0.51	***	0.0678
	Level 4	1.819	***	0.522	***	0.0603
	Level 5	1.33	***	0.242	**	0.0254
Annoyed	Level 1	4.329	***	0.297	**	0.0773
	Level 2	1.469	***	0		0.0039
	Level 3	1.377	***	0		0.0024
	Level 4	0.664	**	0		-0.0086
	Level 5	0		-0.201	*	-0.0386
Schoolwork	Level 1	5.216	***	0.309	***	0.0922
	Level 2	2.044	***	0		0.0128
	Level 3	1.601	***	0		0.0059
	Level 4	0.688	**	0		-0.0083
	Level 5	1.588	***	-0.248	**	-0.0185
Sleep	Level 1	5.402	***	0		0.0649
	Level 2	1.747	***	0		0.0082
	Level 3	1.747	***	0		0.0082
	Level 4	0.869	***	0		-0.0054
	Level 5	0		-0.31	***	-0.0492
Daily Routines	Level 1	5.516	***	0		0.0667
	Level 2	2.012	***	-0.482	***	-0.0348
	Level 3	1.804	***	-0.543	***	-0.044
	Level 4	1.116	***	-0.496	***	-0.0501
	Level 5	1.116	***	-0.496	***	-0.0501
Activities	Level 1	5.922	***	0.524	***	0.1242
	Level 2	5.291	***	0.524	***	0.1144
	Level 3	4.854	***	0.314	***	0.0871
	Level 4	4.591	***	0.271	***	0.0788
	Level 5	2.794	***	0		0.0244
Class membership		0.5468	***	0		
Posterior class probabilities		0.633		0.367		
Log-likelihood		-30650.442				
BIC(LL)		61763.9872				
Npar		61				
LRT Chi-squared(9)		23.070***				

***,**,* significant at the 1,5,10% level of significance

LRT: Likelihood ratio test of current model against main model with no restrictions

Table 7 - Comparison of two rescaling approaches

Health states	CHU9D classification	TTO scores	BWS DCE estimates (0 - 1 scale)	Rescaled scores based on PITS value only (Method 1)	Rescaled scores based on mapping approach (Method 2)
1	414355432	0.3421	0.3223	0.1788	0.2505
2	231345314	0.4592	0.4801	0.3700	0.4250
3	423141114	0.6263	0.6027	0.5186	0.5606
4	555555555	-0.2118	0.0000	-0.2118	-0.1059
MAE (Range)	-	-	-	0.0901 (0-0.1633)	0.0743 (0.0342-0.1059)

Note: MAE - mean absolute error.

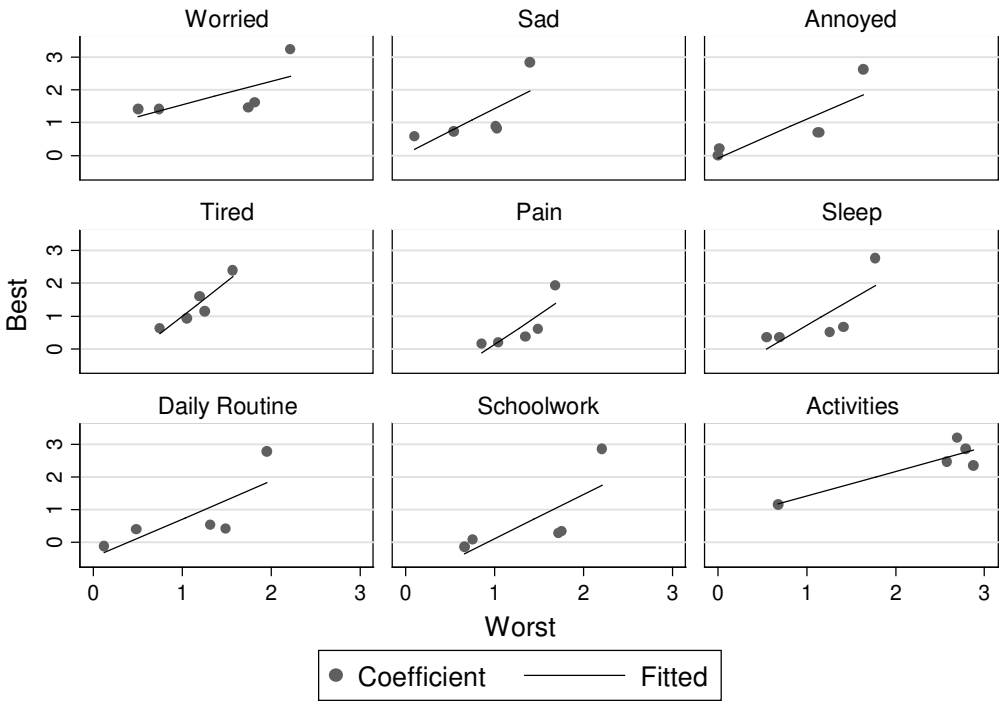

Figure 1 – Plot of MNL coefficients for best versus worst choices

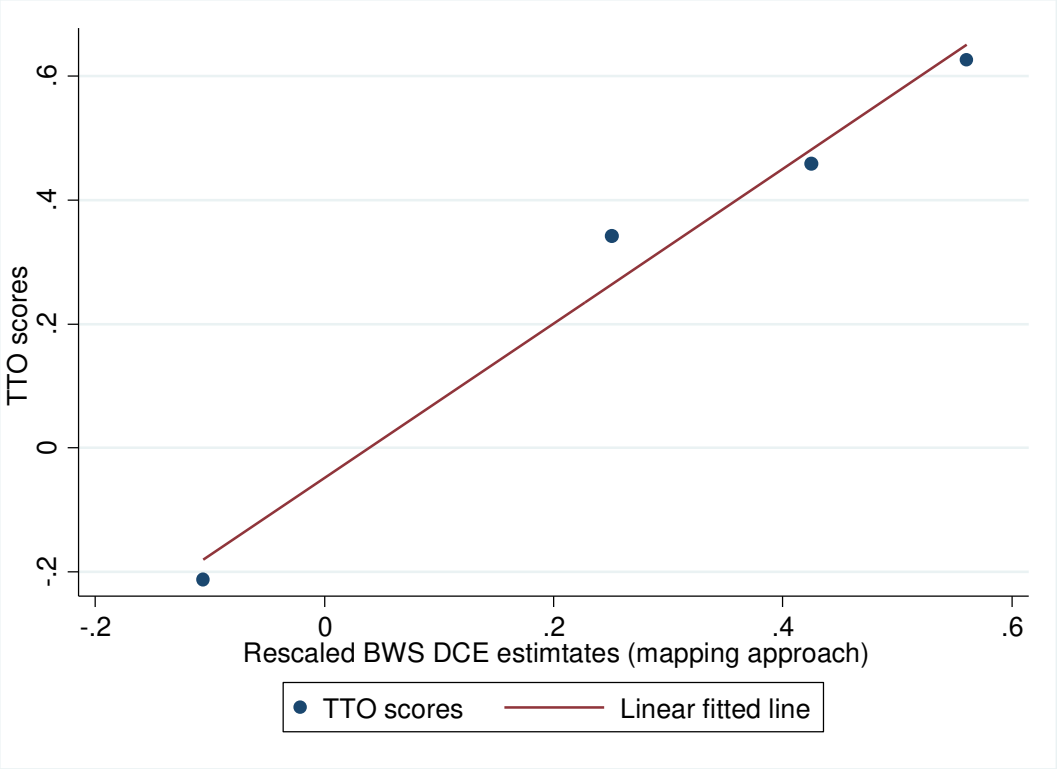

Figure 2 – Scatter plot between TTO scores and rescaled BWS DCE estimates

Appendix 1: Screen shot example of profile case BWS question

Health State X	Best	Worst	Second Best	Second Worst
I feel a little bit worried today	○	○	○	○
I feel a little bit sad today	○	○	○	◉
I have a little bit of pain today	○	◉	○	○
I feel a little bit tired today	○	○	○	○
I feel a little bit annoyed today	○	○	○	○
I have a few problems with my school work today	○	○	○	○
Last night I had a few problems sleeping	○	○	◉	○
I have a few problems with my daily routine today	○	○	○	○
I can join in with most activities today	◉	○	○	○