UNIVERSITY
OF YORK

CENTRE FOR HEALTH ECONOMICS
HEALTH ECONOMICS CONSORTIUM

Comparing Scaling Methods for Health State Valuations - Rosser Revisited

by

Claire Gudex, Paul Kind, Harmanna van Dalen
Mary-Alison Durand, Jenny Morris
and Alan Williams

DISCUSSION PAPER 107

COMPARING SCALING METHODS FOR HEALTH STATE VALUATIONS - ROSSER REVISITED

by

Claire Gudex
Paul Kind
Harmanna van Dalen
Mary-Alison Durand
Jenny Morris
Alan Williams

July 1993

The Authors

Claire Gudex is a Research Fellow and Paul Kind is a Senior Research Fellow in the Centre for Health Economics. Alan Williams is Professor of Economics in the Economics Department. Harmanna van Dalen, Mary-Alison Durand and Jenny Morris are former Research Fellows in the Centre for Health Economics.

Acknowledgements

The authors wouild like to acknowledge the contribution of David Lewis, formerly a member of the MVH Project Group. We would also like to thank Paul Dolan for his comments on an earlier draft of the paper. The Department of Health and the Economic and Social Research Council have provided financial support for this work.

Further Copies

Further copies of this document are available (at price £7.00 to cover the cost of publication, postage and packing) from:

The Publications Secretary
Centre for Health Economics
University of York
York Y01 5DD

Please make cheques payable to the University of York. Details of other papers can be obtained from the same address, or telephone York (0904) 433648 or 433666.

CONTENTS

ABSTRACT

The methodology of valuing health states remains a key issue in the construction of health-related quality of life measures. Different scaling methods appear to yield different sets of valuations, and as yet there is no consensus as to which method is the preferred technique. Many of the cost-per-QALY estimates produced in the United Kingdom have been based on the Rosser Classification and its associated Valuation Matrix, which used the method of Magnitude Estimation.

This paper reports on a study comparing three main scaling methods (Category Rating, Magnitude Estimation, and Time Trade-off) using the Rosser Classification to describe states of health. The objectives of the study were two-fold. First, to assess whether the Magnitude Estimation valuations obtained from a random sample of the general population correspond to the original values obtained from a convenience sample by Rosser, and second, to compare health state valuations produced by three different scaling methods.

The values in the original Rosser matrix were not exactly reproduced in this study. Possible contributing factors are the differing demographic characteristics of the respondents, as well as differences in the design of the study and in the detail of the Magnitude Estimation technique used.

There is a high degree of consensus between Category Rating, Magnitude Estimation, and Time Trade-off methods in the ranking of states, however there appear

to be important differences between the actual valuations produced by the different methods. In addition valuations can be sensitive to order of presentation of method and interviewer bias.

It appears from a reworking of cost-per-QALY estimates using these 'new' matrices that the precise "quality-adjustments" used in QALY estimates do matter. Given that no-one is yet in a position to claim to have established a "definitive" set of valuations, empirical cost-per-QALY measures thus need to be interpreted with some caution. In the meantime however, those wishing to calculate QALYs on the basis of Rosser's descriptive system will have to choose from one of three alternatives - the original matrix, a new matrix based on Magnitude Estimation, and a 'synthesised' matrix based on a commonsense interpretation of scores from all three scaling methods. This may not be an easy choice, and for the time being it might be advisable to conduct analyses of QALY data using all three matrices so as to test the robustness of any conclusions to be drawn.

1.0 INTRODUCTION

Despite increased interest and activity in the measurement of health-related quality of life in recent years, some key methodological issues in the construction of health status measures remain open. One of these is how valuations for health states should be obtained. There is no consensus as to which of the various scaling methods should be the preferred technique, and, as a further complication, different methods appear to yield different sets of valuations. However, since the choice of the valuation method is often restricted by practical constraints, relatively few studies have obtained within-subject comparisons of the different valuation methods[1-4].

A variety of methods has been used for valuing health states[5], the four most common being category rating (CR), magnitude estimation (ME), standard gamble (SG), and time trade-off (TTO). These methods are based on different theoretical frameworks and require respondents to answer quite different types of questions. SG has been considered by economists to be the classical or 'gold standard' method as it is based directly on the fundamental axioms of the von Neumann-Morgernstern 'expected utility theory'. This has been called into question, however, on the basis that SG values can be strongly influenced by the characteristics of the outcomes used in the gamble and the way in which the information is presented[6].

In an early paper Patrick et al[1] compared valuations from CR, ME, and the equivalence method. They concluded that the equivalence technique was too complex

for use in general population surveys, that there was a linear relationship between the results obtained from CR and ME, and that the CR method was the simplest and most reliable of the methods. The form of ME procedure in their study used a scale for which both extremes were fixed, and it was later argued that this was in fact a form of category rating.

Torrance[2] on the other hand, when comparing CR, TTO and SG, found that CR was the most difficult method and TTO the easiest. Differences in test–retest reliability were not statistically significant (though these tests were only undertaken on a high education sample). For population means, Torrance found the SG and TTO data to be equivalent and the relationship between TTO and CR to be curvilinear. However, these relationhips did not hold at the individual level. In conclusion, Torrance suggested that TTO was the best method since it gave comparable results to SG but was simpler and more reliable. In a later paper[3], Torrance found no significant difference in the reliability of TTO and CR and recommended the use of TTO if financial resources permitted, and CR if not.

Read et al[4] found moderately high correlations between SG, TTO and CR but noted that high correlations do not guarantee equivalent ratings. As noted elsewhere also[7], SG appears to generate values higher than CR.

Froberg and Kane [8] have more recently reviewed a number of valuation methods for reliability, validity and feasibility. While they highlighted the need for further research to establish validity and reliability, they concluded that "... the most

promising scaling methods are the category rating, magnitude estimation, and the time trade-off methods" (p. 470).

One of the most widely quoted studies of health state valuations in the United Kingdom is that undertaken by Rosser and Kind.[9] Combining 8 levels of disability with 4 levels of distress (Figure 1), they defined a total of 29 health states and obtained valuations of these states from a convenience sample of 70 respondents (10 psychiatric patients, 10 psychiatric nurses, 10 experienced doctors, 10 state registered nurses working on medical wards, 10 medical patients and 20 healthy volunteers) using ME. 50 respondents in a second group also used equivalence techniques.

The matrix of valuations produced from their work (Figure 2) has been used to compute quality-adjusted life years (QALYs) and to construct cost-per-QALY league tables.[10,11,12] However, the results of Rosser's study have attracted some criticisms on the grounds that the sample size was too small to support any analysis of individual differences, and that the study respondents were not representative of a wider population.

These criticisms, plus the lack of consensus regarding the most appropriate valuation method, and the potential widespread use of Rosser valuations and of QALY estimates indicate a need to undertake further comparison of the methods used to obtain health state valuations.

Such a study has been conducted by the Research Group on the Measurement and Valuation of Health (MVH) at York, with two objectives in mind:

(i) to assess whether the valuations derived by Rosser from a convenience sample correspond to values obtained from a random sample of the general population using similar methods

and

(ii) to compare valuations produced from different methods of valuing health states.

This paper describes the method and results of the MVH study. Valuations obtained from three different methods are compared, and implications for QALY estimates are discussed.

2.0 METHOD

2.1 Study Design

On the basis of age and social class characteristics, three electoral wards in the York Parliamentary Constituency were selected as being (collectively) representative of the population of England and Wales[13,14], see Table 1. Assuming a response rate of about 40%, it was estimated that about 800 individuals would have to be approached

in order to yield the 320 respondents which the study design required (including pilot work). An initial 859 respondents were selected by taking every 55th name on the 1988 electoral registers for the three targeted wards. Letters were sent out inviting participation in the study and also giving an opportunity to opt-out. A final total of 327 people were interviewed in their own homes by eight specially trained interviewers. Forty of these interviews were conducted as part of a pilot study (see Table 2).

All respondents carried out a category rating (CR) exercise and either a magnitude estimation (ME) task or a time trade-off (TTO) task. They were also asked to provide background socio-demographic information.

Rosser's classification contains descriptions of 29 living states described in terms of disability and distress, plus the state 'dead'. Six of these states were selected in her original study as 'marker' states, being broadly representative of the full range of severity. These same states were given a similar role in the study described here. It would have been impractical to expect respondents to value 28[1] states on each of two main valuation methods (ME, TTO) and a pre-pilot had suggested that the valuation of 19 states on TTO was too tiring. Hence a factorial block design was adopted (see Figure 3) in which the states were divided into two subsets (A and B) for the ME and CR exercises, and into four subsets (C, D, E and F) for the TTO exercise. States were assigned to the ME and CR subsets by including each of the six original Rosser marker states (IC, IID, VC, VIB, VIIB and VIID) and then by allocating the remaining

[1] State 1A is a reference state with the assigned value of 1.0 and thus does not need to be valued.

22 states to either A or B. Each of the TTO subsets included the six marker states and six other states.

The order of presentation of task was varied so that half of the respondents completed the CR task first and half completed either the ME or TTO task first. The resulting 36 packs of interview material were distributed across the three wards, with every pack being distributed at least once to each ward, and each of the eight interviewers being assigned contact addresses in each of the wards.

2.2 Valuation Methods

The three methods used in the study were ME, TTO and CR.

(i) Category Rating:

Three variants of CR were used.

CRT: a visual analogue rating scale was used in the form of a thermometer where 100 represented "best imaginable health state" and 0 represented "worst imaginable health state". Respondents rated each state by drawing a line to whichever point on the scale best reflected their perception of how good or bad it was to be in the health state for 20 years.

CRN: respondents were asked to put a cross in a box labelled between 1 to 9 to reflect their perception of how good or bad it was to be in the health state for 20 years.

CRL: respondents were asked to put a cross in a labelled box to reflect their perception of how good or bad it was to be in the health state for 20 years. The boxes were labelled "best imaginable health state"; "very good"; "good"; "fairly good"; "neither good nor bad"; "fairly bad"; "bad"; "very bad"; and "worst imaginable health state".

(ii) Time Trade-Off:

Respondents were first asked to rank the states in order from best to worst and then to identify those states which were considered to be worse than death (assuming they lasted 20 years). Taking first the states considered to be <u>better</u> than death, an iterative process was used to find out how many years in state IA would be equivalent to spending 20 years in each dysfunctional health state. The time in state IA was varied until the point at which the respondent was indifferent between the two states was reached. For states considered <u>worse</u> than death, a similar iterative process was used to find out how many years in each dysfunctional state would be equivalent to immediate death. The time in the dysfunctional state was varied until the point at which the respondent was indifferent between the state and immediate death was reached.

(iii) Magnitude Estimation:

Respondents were first presented with the six marker states printed on cards and were asked to judge each health state in terms of its perceived severity compared with the reference state of "no disability, no distress" (IA) which was assigned a value of 1.0. They were asked whether each state was better or worse than the reference state, and then how many times better or worse. The number given was written on the card. This procedure was repeated with the other 11 states. When all 17 states had been valued, the interviewer placed them in rank order with state IA at the top, and the respondent was given the opportunity to make any changes he or she wished.

Comparison with Original Rosser Procedure

The ME procedure used here differs from that originally used by Rosser in a number of ways. Firstly Rosser used descriptions of distress which included a reference to pain and its treatment with aspirin or heroin. Although these formed part of the original study the distress classification was subsequently published as a four-point categorical scale -none/mild/moderate/severe - and these labels were adopted in the MVH study. The reference to being in a wheelchair was also omitted, since this had proved ambiguous in meaning.

Secondly, Rosser asked respondents to rank and score six 'marker' states. The least severe of these states was 1C (no disability, moderate distress). The valuation process

continued by presenting successive pairs of the remaining marker states: the state ranked second was valued with respect to IC; the state ranked third was valued with respect to that ranked second, and so on. Hence there was no fixed reference point. Respondents were offered the chance of assigning a zero score to that state which they regarded as describing optimal function. This resulted in IA being valued as 0 in the Rosser method (and was subsequently re-scaled to 1.0)[15]. In the MVH study, a more orthodox ME procedure was adopted, in which respondents were required to estimate valuations for health states always with respect to the same reference state of IA (no disability, no distress), and thus state IA is by definition 1.0.

Thirdly, Rosser engaged respondents in a discussion of the interpretation of their scores so that any difficulties were resolved during the interview. This was felt to be unduly directive and was excluded in this study.

2.3 Analysis of Data

In order to compare results from different methods, valuations from individual respondents were transformed into a 0 to 1 scale where 0 = 'death' and 1.0 = 'full health'.

For both the **ME** and **CRT** tasks, raw valuations were transformed into a 0 to 1 scale using the following formula:

$$V_{ij}' = \frac{(V_{ij} - V_{DEATH})}{(V_{1A} - V_{DEATH})}$$

where V_{ij}' = transformed valuation (on 0–1 scale) for state with disability level i and distress level j;

V_{ij} = observed valuation for state with disability level i and distress level j;

V_{DEATH} = observed valuation for the state of 'being dead';

V_{1A} = observed valuation for state IA (no disability, no distress)

For **TTO** valuations, subjects' observed valuations for states initially rated better than being dead were transformed onto the 0–1 scale as follows:

$$V_{ij}' = \frac{V_{ij}}{20}$$

where V_{ij}' = transformed valuation (on 0–1 scale) for disability level i and distress level j;

V_{ij} = observed valuation (i.e., equivalent number of years in state IA) for 20 years in state with disability level i and distress level j.

Due to a problem with the TTO scoring protocol, it was not possible to transform the observed valuations of those states initially rated worse than being dead onto the 0–1 scale. These states were thus described simply as having a "negative" value.

Although individual respondents' ratings are made from a limited range of discrete integers in the **CRN** method, these data can be treated as continuous for the purposes of sub-group analysis. The value given to each state was based on the median category rating score. Once the group valuation matrix has been completed the entire set (rather than that of the individual respondents) was transformed according to the same formula applied to the ME and CRT data.

The **CRL** data consists strictly of ordinal information on individual preferences, but here they have been treated as equivalent to the ratings generated by the CRN method. The CRL data were coded by giving each labelled category an equivalent numeric value 1 through 9. They were then aggregated and transformed in the same way as the CRN data.

Subjects' ME and CRT valuations were included in subsequent analyses if they were complete (i.e. all states were given valuations) and transformable (i.e. valuations could be transformed onto the 0-1 scale which essentially requires that 'being dead' be considered worse than 'no disability, no distress').

The distributions of the transformed valuations for all methods were found to be skewed with mean values lying consistently below median values. Since the mean as a measure of central tendency is particularly sensitive to values at the extremities of a distribution, the median has been considered more appropriate. For this reason non-parametric statistical methods have been used to analyze the data.

3.0 RESULTS

3.1 The Study Population

As Table 3 shows there were more respondents in the 25–44 age group than in the England and Wales population (χ^2=4.3,p<0.01). The study was also over–represented on social classes II and III(NM) (χ^2=17.0, χ^2=35.6 respectively, both p<0.01), and under–represented on social classes III(M), IV and the economically inactive (χ^2=14.6, χ^2=6.5, χ^2=100.0 respectively, all P<0.01).

There were no statistically significant differences in age, sex or social class between the ME and TTO subgroups, nor between respondents grouped by interviewer.

The study population differs in several respects from that of Rosser (Table 4). 20% of subjects in the MVH study reported past experience as an inpatient, a quarter of the rate in the Rosser study, which included two subgroups of patients (20 of the total 70 respondents). The MVH study contained no current inpatients at all. The biases in the Rosser study towards young female respondents is absent from the MVH study, and nearly four times as many MVH respondents were classified as being of the Protestant faith.

3.2 Health State Valuations

Summaries of the transformed valuations obtained from each of the methods are presented in Tables 5 to 9.

Within the Rosser classification the disability and distress descriptions constitute two ordinal scales. Thus the relationship between some of the combinations of disability and distress are logically defined e.g. state VC should be more severe than state IVC and than state VB. For other comparisons, such as VC and IID, there are no such inherent logical orderings, since the trade-off between disability and distress is not known.

The transformed ME values (Table 5) generally conform with the logical ordering inherent in the disability and distress scales, with more severe states having lower values. However there are three reversals – between states VB and VC, between states VB and VIB, and between states VIIA and VIIB. In all three cases it is probable that the arithmetic differences are insignificant, given the overlap in interquartile ranges (significance tests could not be performed as the values were generated from different subgroups). All three reversals of logical ordering are between states in which one of the pair is a marker state (therefore valued by all respondents) and the other is a non-marker state (therefore valued by only one of the two ME subgroups). When the valuations provided by the two subgroups were examined separately however, precisely the same reversals involving the same pairs of states were

found. Therefore, the factorial block design appears not to be responsible for the reversals found in the ME matrix.

The matrix of transformed TTO values show six reversals of logical ordering (Table 6). All are between pairs of states which were valued by different subgroups of respondents. When the valuations provided by the four TTO subgroups were examined separately, only one reversal was found – yet this occurred between a pair of states (VIIB and VIID) where no reversal exists in the aggregated matrix! Therefore, while the combination of data sets helps to eliminate this reversal, it is responsible for the six reversals that are present in the aggregated TTO matrix. Unlike the other methods, the TTO matrix contains negative values for two states.

There are 11 violations of logical ordering within the transformed CRT matrix (Table 7). Nearly half of these violations involve states VA, VB and VC. Nine of the reversals are between pairs of states valued by different subgroups. When the valuations of the two subgroups are examined separately, seven reversals remain. The factorial block design appears to be responsible therefore for 4 of the 11 reversals in the CRT matrix.

Only two violations of logical ordering are evident in the CRN valuations matrix (Table 8). Values for states ID and IID, and for IIID and IVD are reversed. Both reversals are accounted for by the combination of data from different subgroups.

There are no violations of logical ordering within the CRL valuations matrix (Table 9), presumably since there are relatively few distinct scores and many states share common values. Paradoxically, one reversal appears when valuations from different subgroups are examined separately.

There are in total only eight scores generated for the entire CRN matrix, and only six for the CRL matrix (excluding 0.0). These two matrices have thus not been used further in the following analysis.

3.3 Differences in Valuations

Influence of Health State Subsets

Within both the ME method and the CRT method, Mann–Whitney U tests indicated that there were no statistically significant differences in valuations for the marker states between those who valued the states in subset A and those who valued the states in subset B. Similar results relating to the TTO subsets (C, D, E and F) were found using Kruskal–Wallis one–way analyses of variance.

Influence of Order of Presentation

A summary of significant differences in valuations relating to order of task presentation is presented in Table 10. For each valuation method, analyses were carried out on:

(a) the valuations for the marker states when the valuations of all the health state subsets are pooled; and

(b) the valuations of all the states within each of the health state subsets.

The data were analyzed in two ways:

(a) Mann–Whitney U tests were performed to test for differences in valuations for specific states between those who carried out a given task first and those who carried it out second; and

(b) Sign tests were carried out to test for systematic differences in median valuations between the two groups across all states.

It can be seen that there were some associations between valuations and the order of task presentation. The patterns of association, however, were not consistent across the valuation methods.

Influence of Interviewer

Similar analyses were carried out to test for the influence of interviewers on valuations, except that:

(a) Kruskal–Wallis analyses of variance were performed to test for differences in valuations between interviewers for specific states;

(b) Friedman's rank tests were performed to test for systematic differences in median valuations between interviewers across all states; and

(c) The statistical significance of differences in median valuations of the states within the TTO blocks were not computed because of the very small numbers of subjects who fell into some cells.

A summary of significant differences in valuations relating to interviewer is presented in Table 11. From this, it appears that there were some associations between valuations and interviewers within each of the valuation methods, but that the associations appear to be strongest within ME. Again, the patterns of association were not consistent across methods.

Table 12 presents the data for ME in more detail, in order to better illustrate this 'interviewer effect'. Median valuations by interviewer are given for each state. These valuations have been ranked within each state, and the rankings then summed for each interviewer. According to the Friedman 2-way analysis of variance technique, if there is negligible interviewer influence, then there should be little difference between interviewers in the sum of ranks. However as can be seen from the table, there are large differences with interviewers #1 and #5 producing higher scores than the other interviewers.

Similar analysis of the TTO data suggest that interviewer #6 produced scores which were significantly higher than those collected by the other interviewers. On CRT, respondents seen by interviewers #5 and #6 produced the highest scores.

A review of the interviewer characteristics did not reveal any significant differences that might explain these results.

Influence of Respondent Characteristics

Although no single characteristic was linked to all methods, both personal experience of illness and experience of working in health or social services appeared to influence both ME and TTO scores.

Respondents who rated their own current health below the median value of 85 on the visual analogue scale (which has scores ranging from 0 to 100) produced ME scores for the marker states which were significantly lower than those produced by respondents with own health greater than 85 ($z=-2.2014$, $p<0.05$). Respondents who had recent personal experience as in-patients produced significantly lower TTO scores for the marker states than those who had no such experience ($z=2.02$, $p<0.05$).

Both ME and TTO scores for marker states were significantly lower for those respondents who had experience of working for health or social services ($z=-2.0226$, $p<0.05$ for both ME and TTO).

On the CRT there were significant differences with respect to age (respondents aged 61 years or over produced higher scores than younger respondents, $p<0.05$); sex (males produced higher scores than females, $p<0.05$); educational attainment (respondents

with qualifications gained after leaving school gave lower scores than those without, $p<0.05$); and personality (respondents with low Neurotic scores produced higher CRT scores than those with high Neurotic scores, $p<0.05$, while respondents with high Psychotic scores produced higher CRT scores than those with low Psychotic scores, $p<0.05$). There were no significant difference in ME or TTO scores on the basis of these characteristics.

3.4 Comparison of Valuations by Method

There is a high degree of consensus between methods in the ranking of states and there are highly significant rank correlation coefficients between methods (Table 13). However there are some large differences in the ranking for some states e.g. state VA is ranked 8th on the TTO scale but is ranked 16th on ME and 20th on CRT, and there are similar differences for states IID and VIC.

Despite the strong correlations between rankings, the Friedman test statistic for the valuations given in Table 13 (excluding the original Rosser matrix valuations) is highly significant ($Q=73.32, p<0.001$), indicating important differences between valuations produced by different methods.

In order to investigate these differences further, data were analyzed from the following groups:

(a) those who completed both ME and CRT (n=39); and

(b) those who completed both TTO and CRT (n=44)

To test for significant differences in valuations according to valuation method, sign tests were carried out on the valuations attached to specific states by individuals within each of the groups. For most states, ME and TTO valuations tended to be higher than CRT valuations. A summary of significant differences is presented in Table 14.

3.5 Comparison with Original Rosser Matrix

Scores from Rosser's original ME data are also shown in Table 13. In terms of the ranking of states, the ME matrix from the MVH study is the closest to the original Rosser matrix (Spearman's rho = 0.97, $p<0.001$). Most states have the same or similar rank to that in the original matrix, though larger discrepancies appear on states 11D, VA, VB and VD, and VIB amd VIIB.

Spearman's correlation coefficients between the TTO and CRT matrices and the original ME matrix are 0.91 and 0.88 respectively, with $p<0.001$ in both cases).

Despite the similarity of the MVH ME matrix to the original Rosser ME matrix with respect to ranking, there are clear differences in the transformed valuations. The original matrix assigns relatively high scores to nearly all 29 states, with 18/29 scoring 0.9 or more. In the MVH study, only 3/29 states in each of the ME and CRT matrices and 2/29 in the TTO matrix have scores as high as this. Only 4 states in

the original matrix score below 0.5, compared to over half of the states on the MVH ME, TTO and CRT matrices. No score in either of CRN or CRL matrices are higher than 0.75 and in fact the majority are below 0.5.

Differences in scores within the MVH matrices are substantially greater for states at the less severe end of the dysfunction range. State ID on the Rosser matrix scores 3.3% below IA, whereas the difference on the MVH ME matrix is 10–times greater, and on the other matrices even greater, with ID on CRN scoring 69% below IA.

Perhaps most striking of all, no states in the MVH ME or CR matrices have negative values. Although individual respondents scored some states below zero, the aggregate results show no states worse than death. The Rosser matrix, as does the TTO matrix, identified two such states – VIID and unconscious.

3.6 A Synthesised Matrix?

Although there is much agreement in the ranking of states according to the various methods adopted in this study, considerable differences remain in the values for any given health state. In addition to this variation across matrices, there is the underlying phenomenon of internal inconsistency within each valuation matrix. If called upon to designate a revised set of values for the Rosser Classification based on the data from this study, it would therefore be difficult to choose between the various matrices presented here.

Is there a way therefore, of combining the data from the various matrices, to produce a 'synthesised' matrix? Table 15 presents the median values for each state from the ME, TTO and CRT matrices, and a suggested 'synthesised' value is then also given for each state. It should be emphasised that these values are based essentially on a "common sense" interpretation of the other scores, and do not arise from an underlying mathematical or statistical technique. It will be seen that these new values decline more steadily than those in Rosser's original matrix, and that no states are rated as worse than being dead.

3.7 Consequences for QALY Computations

The original Rosser valuation matrix has been used in previous studies as the quality adjustment in the calculation of QALYs for a variety of health care interventions. The production of a reworked ME matrix, and of 'new' TTO and CR matrices raises obvious questions about their effect on both the absolute value of the various QALYs estimates in current circulation, and about the impact on the rank ordering of those interventions. Some reworked cost-per-QALY estimates using the 'new' matrices produced from the MVH valuation study are presented here (Note that cost data have NOT been reworked).

"Economics of Coronary Artery Bypass Grafting (CABG)"

Table 16 shows the reworked cost-per-QALY estimates and rankings for CABG and percutaneous transluminal coronary angioplasty (PTCA) as evaluated by Williams[10].

It is clear that no two matrices produce precisely the same rank ordering of treatments. The relevant rank correlation coefficients are displayed in Table 17, with the highest being between the cost-per-QALY estimates based on the original and the reworked ME valuations.

All the matrices from the MVH study identify the same three treatments as being the most cost-effective: PTCA for severe or moderate angina with 1 vessel disease, and CABG for severe angina with left main vessel disease. Two of these differ from the top three identified by the original Rosser matrix (CABG for severe angina with left main vessel disease or with 3 vessel disease, and CABG for moderate angina with left main vessel disease).

"QALYs and their Use by the Health Service"

The reworked cost-per-QALY estimates and rankings for the treatments evaluated by Gudex[16] are shown in Table 18. The relevant rank correlations are given in Table 19.

The main change here is that whichever of the MVH matrices is used, shoulder joint replacement emerges as the most cost-effective treatment (instead of second most cost-effective). At the other extreme when using any but the TTO matrix, surgery for idiopathic adolescent scoliosis emerges as the least cost-effective of the seven treatments, and even appears to generate a negative cost-per-QALY estimate suggesting that surgery is actually harmful rather than beneficial.

"Prioritising Waiting Lists"

The reworked QALYs per hour of operating time and the respective rankings of treatments evaluated by Gudex et al[17] are shown in Table 20. The associated rank correlation coefficients shown in Table 21 are higher than in the preceding cases, indicating that the priority rankings remain broadly the same whichever matrix is used. The best correlations are again between the estimates based on the original Rosser matrix and those based on the MVH ME matrix.

4.0 DISCUSSION

There are several issues that arise from this work. Firstly, why is the Rosser ME matrix produced here different from the original one produced by Rosser in 1978? Secondly, what more has been learnt about the methodology of health state valuation? Thirdly, what are the implications of these different valuation matrices for QALY computations?

Comparison with Original Rosser Matrix

As noted earlier there are some differences in the design of the MVH and Rosser studies, and in the detail of the scaling procedures used.

The descriptions of health states are essentially the same for disability. Rosser incorporated more detail in her references to distress, but the use of "pain" or

"anxiety" or "depression" as additional material in the health states descriptions had reportedly made no significant difference to the results she obtained. In addition it is unlikely that these differences in description account for all the observed differences in scores. In the present study State ID (no disability, severe distress) attracted a score of 0.967 in Rosser's study, compared to 0.667 in the MVH study.

Both the Rosser and MVH studies were conducted by interviewers specially trained in the psychometric techniques required. The MVH interviewers were given a standard protocol to follow and so far as is known did not discuss respondents' results with them. This was not so in the Rosser study where the more active interaction between interviewer and respondents may have played a key role in eliminating inconsistency in responses.

The form of ME used by Rosser differs from that described in the psychometric literature in two respects. Firstly, respondents made judgements about the relative severity of successive pairs of marker states. While the first of these pairs always included IC, the following comparisons were made with different states as determined by their ranked order. Hence not only did individual subjects encounter a changing reference state for each pairwise judgement, but these reference states differed across the 70 subjects also. The MVH study adopted a more conventional form of ME in which all comparisons are made with respect to a single fixed reference state. Contextual and range effects within ME have been noted by previous authors[18,19], but it is doubtful that these effects could now be investigated within the Rosser data.

Secondly, in the Rosser study the nature of the valuations task changed over the course of the interview. Respondents were initially asked to make judgements about the six marker states. This produced a numeric framework in which the remaining 23 states were placed. These states, however, were assigned values without specific reference to any other one state, and the prompt of "How many times more severe is state B compared to state A?" which had been used for the marker states was not used for these 23 other states. In the MVH study, all states were compared to state IA.

There were also clear differences between the two studies with respect to the respondents who took part. While the MVH respondents were drawn from a general population, Rosser's study involved doctors, nurses, inpatients, and only 20/70 respondents from a general population. The Rosser group was also younger and included more women. Previous authors have noted that demographic characteristics may be linked with differences in valuations[20].

Thus it is not easy to isolate any one factor which accounts for the differences between the Rosser and MVH ME scores. Since the MVH study has varied the precise detail of the original protocol it could be argued that like is not being compared with like. It is probably the impact that these different matrices have on QALY estimates that determines whether these differences matter or not.

Implications for QALY Computations

As might be expected, the precise "quality adjustments" used in QALY estimates do matter, and they may be crucial in certain circumstances. The differences observed here only serve to reinforce the notion that such estimates need to be interpreted with some caution. It may also be that uncertainty regarding the epidemiological evidence, and the status of cost data, are more important sources of variability in cost-per-QALY estimates than differences arising from changes in the valuation matrix. A more broadly-based sensitivity analysis could address such issues, but these did not form part of the substantive MVH study.

In the meantime however, should one of these revised matrices replace the original Rosser matrix, and if so, which one? There is an argument that because this study was not an exact replication of Rosser's original work, then not even the revised ME matrix is appropriate to use. In addition the study population was not representative of the general British population despite efforts to achieve this. On the other hand, it can be argued that these revised matrices are better than Rosser's original one because they are based on valuations from a general population and not from medical staff and patients, and the marker states at least in all methods were valued by 143 and 144 respondents respectively compared to the original 70.

It is difficult to identify one method as being better than the others: the ME method was more vulnerable to interviewer effects, and also produced the most

violations of logical ordering when the factorial block design was taken into account. The TTO performed better on consistency but there were difficulties in determining valuations for states considered 'worse than death'. CRT also performed better on consistency. There were fewest violations of logical ordering when respondents were restricted in their choice of rating e.g. in CR with numbered or labelled boxes, but the scores produced were clearly categorical.

Thus one option is to use the 'synthesised' matrix based essentially on a "common sense" interpretation of the other scores. Although this matrix does not arise from an underlying mathematical or statistical technique, it has the advantage that its values decline more steadily than those in Rosser's original matrix and appear to be more realistic intuitively.

The Methodology of Health State Valuation

This study has shown that the valuations from ME, TTO and CRT are sometimes sensitive to order of presentation of method. Although the direction of influence was inconsistent across methods, it does suggest that in all such comparative studies order of presentation should be varied in a systematic manner, since it appears to be a potential source of variation on valuations.

The results also indicate that all methods are vulnerable to interviewer bias, although ME was the most affected. Not surprisingly, the CRT method displayed less interviewer bias than the others, since the interviewer simply gave the questionnaire to

the respondents who completed it on his/her own. In those circumstances it is perhaps surprising to find any interviewer effect at all though Sutherland et al[21] have reported an interviewer effect where "the interviewer's role was only to describe the study, obtain informed consent for participation in the study, and sit with the patient while s/he completed the tasks". It is clear that great care needs to be taken in the training and monitoring of interviewers.

All of the methods yielded results which contain violations of the logical ordering of states. No other studies have been found that present 'inconsistencies' in valuation data but it is clearly an important issue in the interpretation of valuation matrices. Inconsistencies in this study appeared to be related to the factorial block design used, and to disability level V, which is described by a large amount of text. Evidence from other parts of the MVH study suggested that there might also be some difficulty in making a distinction between 'mild' and 'moderate' levels of distress. However there is almost certainly a residual amount of inconsistency that has not been explained by the experimental design or ability to interpret the descriptive information. Where such inconsistency arises– and it seems likely that it is an intrinsic feature of all scaling tasks– should it be used to distinguish different levels of respondent performance, or as an index of performance of the scaling method itself? This issue of inconsistency has been further investigated elsewhere [22].

5.0 CONCLUSIONS

This paper describes the results obtained using different scaling methods to elicit valuations for health states described in terms of the Rosser Classification. Since one of the methods used was closely related to that originally employed by Rosser, the study can be described as a 'partial' replication of that original work. Two principal findings emerge from the present work.

Firstly, a reinforcement of the finding that different scaling methods, when applied to the same health states by the same people, yield different valuations. Since there are no a priori grounds for recognising one or other of these methods as having absolute or comparative advantage over rival methods, the study findings further emphasise the absence of a 'standard' method.

Secondly, the values in the original Rosser matrix were not reproduced in this study. The differences may be accounted for in several ways, but essentially this means that those wishing to calculate QALYs on the basis of Rosser's descriptive system will have to choose from one of three alternatives – the original matrix, a new matrix based on ME (despite some logical inconsistencies), and a 'synthesised' matrix based on a commonsense interpretation of scores from all three scaling methods. This may not be an easy choice, and for the time being it might be advisable to conduct analyses of QALY data using all three matrices so as to test the robustness of any conclusions to be drawn.

Note

The data generated in this study is available to other researchers through the ESRC Data Archive.

REFERENCES

1. Patrick, D., Bush, J. and Chen, M. Methods for Measuring Levels of Well-Being for a Health Status Index. Health Services Research, Fall, pp.228-245, 1973.

2. Torrance, G. Social Preferences for Health States: An Empirical Evaluation of Three Measurement Techniques. Socio-Econ. Plan. Sci. 10:129-136, 1976.

3. Torrance, G. Utility Approach to Measuring Health-Related Quality of Life. Journal of Chronic Disease, 40:593-600, 1987.

4. Read, J.L., Quinn R.J., Berwick, D.M., Fineberg, H.V. and Weinstein, M.C. Preferences for Health Outcomes: Comparison of Assessment Methods. Medical Decision Making 4:315-329, 1984.

5. Torrance, G. Measurement of Health State Utilities for Economic Appraisal. Journal of Health Economics, 5:1-30, 1986.

6. Llewellyn-Thomas, H., Sutherland, H., Tibshirani, R., Ciampi, A., Till, J. and Boyd, N. The Measurement of Patients' Values in Medicine, Medical Decision Making, 2:449-462, 1982.

7. Llewellyn-Thomas, H., Sutherland, H., Tibshirani, R., Ciampi, A., Till, J. and Boyd, N. Describing Health States: Methodologic Issues in Obtaining Values for Health States. Medical Care, 22:543-552, 1984.

8. Froberg, D. and Kane, R. Methodology for Measuring Health-State Preferences - II: Scaling Methods. J. Clin. Epidemiol., 42:459-471, 1989.

9. Rosser, R. and Kind, P. A Scale of Valuations of States of Illness: Is There a Social Consensus?, Int. J. Epidemiology, 7:347-358, 1978.

10. Williams, A. Economics of Coronary Artery Bypass Grafting. British Medical Journal, 291:326-329, 1985.

11. Forrest Report. Breast Cancer Screening. (Chairman of Working Party: Professor P. Forrest), London, HMSO, 1986.

12. Parker, M.J., Myles, J.W., Anand, J.K. and Drewett, R. Cost-Benefit Analysis of Hip Fracture Treatment. Journal of Bone and Joint Surgery 74B:261-264, 1992.

13. 1981 Census for England and Wales.

14. 1981 Small Area Statistics for York.

15. Kind, P., Rosser, R. and Williams, A. Valuation of Quality of Life: Some Psychometric Evidence, in Jones-Lee, M.W. (editor), The Value of Life and Safety, North Holland, 1982.

16. Gudex, C. QALYs and their Use by the Health Service. CHE Discussion Paper No. 20, University of York, 1986.

17. Gudex, C. Williams, A., Jourdan M. et al. Prioritising Waiting Lists. Health Trends 22(3):103-108, 1990.

18. Poulton, E.C. The New Psychophysics: Six Models for Magnitude Estimation. Psychological Bulletin 69(1):1-19, 1968.

19. Teghtsoonian, R. Range Effects in Psychophysical Scaling and a Revision of Stevens' Law. American Journal of Psychology 86(1):3-27, 1973.

20. Froberg, D. and Kane, R. Methodology for Measuring Health-State Preferences - III: population and Context Effects. J. Clin. Epidemiol., 42:585-592, 1989.

21. Sutherland, H.J., Lockwood, G.A., Tritchler, D.L., Sem, F., Brooks, L. and Till, J.E. Communicating Probabilistic Information to Cancer Patients: Is there "noise" on the line?, Soc. Sci. Med. In press.

22. Kind P et al. Inconsistency and the Judgement of Health State Valuation: Results from 3 Scaling Methods, Mimeo, Centre for Health Economics, University of York, 1993.

Table 1: Sociodemographic Characteristics of Sampling Frame

	3 York Wards	England and Wales
Age Structure		
under 5	6.0	6.0
5 – 15	16.5	16.2
16 – 24	13.6	14.1
25 – 44	26.4	26.4
45 – 60/65*	20.0	19.6
over 60/65*	17.5	17.7
Social Class		
I	4.3	4.4
II	17.5	17.9
III (non manual)	9.4	9.5
III (manual)	26.1	26.7
IV	12.2	12.2
V	3.9	4.1
Owner Occupier	61.4	57.3
Car Owner	57.0	59.5

* 65 for males, 60 for females

Table 2: Response to Request for Participation in Study

	%	n
Completed interviews	33.4	287
Not approached	4.8	41
Refusals by post	41.1	353
Doorstep refusals	16.1	138
Used in pilot	4.7	40
TOTAL	100.1*	859

* Total greater than 100 due to rounding

Table 3: Characteristics of Study Population
(figures expressed as percentages)

	Study Sample (n=287)	England and Wales (1981)
Age		
18–24	13.9	14.6
25–44	43.6	35.4*
45–60/65	22.3	26.3
Pensionable age and over	20.2	23.7
Sex		
Male	48.1	48.6
Female	51.9	51.4
Social Class		
I	5.7	4.4
II	35.0	17.9*
III Non Manual	18.8	9.5*
III Manual	19.5	26.7*
IV	8.7	12.2*
V	4.0	4.1
Armed Forces	0.4	2.4
Head of Household economically inactive	6.9	22.8*
Missing	1.0	–

* $p<0.01$

Table 4: Characteristics of MVH and Rosser Populations

	Rosser		MVH	
Age				
under 30	46	(66%)	37	(26%)
31 - 45	17	(24%)	43	(30%)
over 45	7	(10%)	63	(44%)
Female	42	(60%)	75	(52%)
Nationality - British	58	(83%)	140	(98%)
Inpatient experience	56	(80%)	28	(20%)
Hospitalisation of family	39	(56%)	69	(48%)
Religion - Protestant	14	(20%)	108	(76%)

Table 5: Magnitude Estimation – Median Values Based on Transformed Data

	A	B	C	D
I	[1.00]	.89 (.80–.98)	.89 (.72–.95)	.67 (.36–.87)
II	.89 (.78–.98)	.81 (.65–.90)	.78 (.59–.89)	.56 (.33–.76)
III	.70 (.53–.89)	.63 (.44–.80)	.57 (.33–.83)	.44 (.22–.67)
IV	.63 (.45–.86)	.56 (.34–.79)	.51 (.33–.74)	.40 (.11–.67)
V	.44 (.33–.79)	.43 (.11–.61)	.44 (.22–.78)	.22 (.08–.50)
VI	.44 (.22–.75)	.44 (.22–.78)	.34 (.17–.64)	.22 (.03–.51)
VII	.38 (.13–.60)	.40 (.22–.64)	.33 (.11–.62)	.20 (.00–.45)
VIII	.01 (.00–.19)			

[Dead = 0]

Table 6: Time Trade-Off - Median Values based on Transformed Data

	A	B	C	D
I	[1.00]	.90 (.76–.95)	.80 (.55–.95)	.45 (.00–.70)
II	.90 (.71–.95)	.70 (.55–.90)	.60 (.45–.84)	.35 (.10–.56)
III	.55 (.30–.80)	.53 (.34–.70)	.45 (.20–.70)	.20 (.00–.45)
IV	.70 (.38–.85)	.45 (.30–.95)	.55 (.20–.83)	.33 (.15–.55)
V	.55 (.30–.83)	.45 (.20–.78)	.43 (.20–.55)	.20 (.00–.45)
VI	.43 (.21–.75)	.35 (.15–.60)	.45 (.20–.70)	.15 (neg.–.30)
VII	.20 (neg–.45)	.10 (.00–.45)	.03 (–0.14–.29)	[negative] (neg.–.05)
VIII	[negative] (neg.–neg.)			
	[Dead = 0]			

Values are based on all TTO data – including states worse than death

[negative] = rated worse than death, but no numerical scores available

Table 7: Category Rating – Thermometer Version based on Median Values of Transformed VAS Scores

	A	B	C	D
I	[1.00]	.85 (.80–.95)	.85 (.66–.90)	.35 (.05–.57)
II	.85 (.70–.90)	.45 (.30–.65)	.60 (.32–.75)	.50 (.26–.66)
III	.50 (.20–.70)	.44 (.29–.59)	.30 (.08–.53)	.30 (.18–.47)
IV	.55 (.38–.70)	.40 (.25–.58)	.40 (.26–.55)	.22 (.10–.44)
V	.25 (.05–.42)	.35 (.15–.55)	.26 (.07–.50)	.17 (.02–.35)
VI	.41 (.17–.60)	.39 (.19–.51)	.30 (.05–.45)	.14 (.05–.30)
VII	.20 (.00–.45)	.16 (.05–.30)	.20 (.10–.35)	.05 (.00–.10)
VIII	.00 (.00–.10)			
	[Dead = 0]			

Table 8: Category Rating (Numbered Boxes) – Transformed Scale Based on Median Category Ratings

	A	B	C	D
I	[1.00]	.75	.75	.31
II	.75	.56	.50	.38
III	.63	.50	.38	.25
IV	.50	.50	.38	.31
V	.50	.38	.38	.13
VI	.38	.25	.25	.13
VII	.25	.25	.25	.00
VIII	.00			
	[Dead = 0]			

Table 9: Category Rating (Labelled Boxes) – Transformed Scale Based on Median Category Ratings

	A	B	C	D
I	[1.00]	.63	.63	.38
II	.63	.50	.38	.25
III	.50	.38	.38	.25
IV	.38	.38	.38	.25
V	.38	.25	.25	.19
VI	.38	.25	.25	.13
VII	.25	.25	.13	.13
VIII	.13			
	[Dead = 0]			

Table 10: Summary of Significant Differences in Valuations by Order of Presentation

[* = p<.01; ** = p<.001; *** = p<.0001]

Data Set Used	Differences Tested	Valuation Method		
		ME[1]	CRT[2]	TTO[3]
All subsets pooled	State by State	–	–	–
	Medians	–	–	–
Subset A only	State by State	–	–	n.a.
	Medians	1st>2nd**	–	n.a.
Subset B only	State by State	–	2nd>1st*[VIB]	n.a.
	Medians	–	2nd>1st***	n.a.
Subset C only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	–
Subset D only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	–
Subset E only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	–
Subset F only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	1st>2nd**

n.a. not applicable

1 Magnitude Estimation
2 Category Rating using the thermometer
3 Time Trade–Off

Note: (a) The states to be rated were divided into 2 subsets (A and B) for ME and CRT, and into 4 subsets (C,D,E and F) for TTO. The 6 "marker" states appeared in all subsets.

(b) The state for which a significant difference was detected is indicated in square brackets, thus [VIB], meaning disability state VI and distress state B (see Figure 1 for full description of states)

Table 11: Summary of Significant Differences in Valuations by Interviewer

[* = p<.01; ** = p<.001; *** = p<.0001]

Data Set Used	Differences Tested	Valuation Method		
		ME[1]	CRT[2]	TTO[3]
All subsets pooled	State by State	1>N>6*[VC]	–	5,6>N>1*[IC]
	Medians	5>N>6***	–	6>N>3*
Subset A only	State by State	–	–	n.a.
	Medians	1>N>6***	4>N>7***	n.a.
Subset B only	State by State	5>N>4*[IVA]	–	n.a.
	Medians	5>N>3***	–	n.a.
Subset C only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	–
Subset D only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	–
Subset E only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	–
Subset F only	State by State	n.a.	n.a.	–
	Medians	n.a.	n.a.	–

n.a. not applicable

1 Magnitude Estimation
2 Category Rating using the thermometer
3 Time Trade-Off

Note: (a) The states to be rated were divided into 2 subsets (A and B) for ME and CRT, and into 4 subsets (C,D,E and F) for TTO. The 6 "marker" states appeared in all subsets.

(b) Interviewers are numbered 1, 2, ... 8

Table 12: ME Median Valuations for Each State by Interviewer

Marker State	Interviewer							
	1	2	3	4	5	6	7	8
IC	.90 (1)	.80 (3.5)	.73 (6)	.70 (7.5)	.81 (2)	.70 (7.5)	.80 (3.5)	.75 (5)
IID	.50 (3.5)	.50 (3.5)	.29 (8)	.44 (6)	.60 (1)	.38 (7)	.50 (3.5)	.50 (3.5)
VC	.63 (1)	.40 (4)	.29 (6)	.27 (7)	.50 (2)	.18 (8)	.42 (3)	.30 (5)
VIB	.50 (3)	.40 (4.5)	.20 (8)	.28 (6.5)	.75 (1)	.28 (6.5)	.56 (2)	.40 (4.5)
VIIB	.50 (2.5)	.30 (4)	.10 (8)	.23 (7)	.50 (2.5)	.25 (6)	.55 (1)	.26 (5)
VIID	.20 (3)	.05 (5)	.00 (6.5)	-0.01 (8)	.40 (1)	.00 (6.5)	.25 (2)	.07 (4)
Sum of Ranks	14	24.5	42.5	42	9.5	37.5	19	27

Table 13: Rank Order of States by Principal Scaling Methods

	ME[1]	TTO[2]	CRT	ORIGINAL ROSSER VALUES
IB	.89 (2)	.90 (1.5)	.85 (2)	0.995 (1)
IC	.89 (2)	.80 (3)	.85 (2)	0.990 (2.5)
ID	.67 (7)	.45 (13)	.35 (14.5)	0.967 (8)
IIA	.89 (2)	.90 (1.5)	.85 (2)	0.990 (2.5)
IIB	.81 (4)	.70 (4.5)	.45 (8)	0.986 (4)
IIC	.78 (5)	.60 (6)	.60 (4)	0.973 (6)
IID	.56 (11.5)	.35 (18.5)	.50 (6.5)	0.932 (15)
IIIA	.70 (6)	.55 (8)	.50 (6.5)	0.980 (5)
IIIB	.63 (8.5)	.53 (10)	.44 (9)	0.972 (7)
IIIC	.57 (10)	.45 (13)	.30 (17)	0.956 (10.5)
IIID	.44 (16)	.20 (22)	.30 (17)	0.912 (16)
IVA	.63 (8.5)	.70 (4.5)	.55 (5)	0.964 (9)
IVB	.56 (11.5)	.45 (13)	.40 (11.5)	0.956 (10.5)
IVC	.51 (13)	.55 (8)	.40 (11.5)	0.942 (13)
IVD	.40 (20.5)	.33 (20)	.22 (21)	0.870 (19)
VA	.44 (16)	.55 (8)	.25 (20)	0.946 (12)
VB	.43 (19)	.45 (13)	.35 (14.5)	0.935 (14)
VC	.44 (16)	.43 (16.5)	.26 (19)	0.900 (17)
VD	.22 (25.5)	.20 (22)	.17 (24)	0.700 (21)
VIA	.44 (16)	.43 (16.5)	.41 (10)	0.875 (18)
VIB	.44 (16)	.35 (18.5)	.39 (13)	0.845 (20)
VIC	.34 (23)	.45 (13)	.30 (17)	0.680 (22)
VID	.22 (25.5)	.15 (24)	.14 (26)	0.000 (25.5)
VIIA	.38 (22)	.20 (22)	.20 (22.5)	0.677 (23)
VIIB	.40 (20.5)	.10 (25)	.16 (25)	0.564 (24)
VIIC	.33 (24)	.03 (26)	.20 (22.5)	0.000 (25.5)
VIID	.20 (27)	[NEGATIVE]	.05 (27)	−1.486 (28)
VIIIA	.01 (28)	[NEGATIVE]	.00 (28)	−1.028 (27)

(1) Correlation coefficient between ME and TTO = 0.85
Correlation coefficient between ME and CRT = 0.91

(2) Correlation coefficient between TTO amd CRT = 0.83

Table 14: Summary of Significant Differences in Valuations by Method of Eliciting Valuations (within Subjects)

[* = p<0.01; ** = p<0.001; *** = p<0.001]

Data Set Used	ME vs CRT		TTO vs CRT	
	State	Difference	State	Difference
Subset A	2C 3B 4B 7B	ME>CRT* ME>CRT* ME>CRT* ME>CRT*	n.a.	n.a.
Subset B	–	–	n.a.	n.a.
Subset C	n.a.	n.a.	–	–
Subset D	n.a.	n.a.	–	–
Subset E	n.a.	n.a.	–	TTO>CRT***
Subset F	n.a.	n.a.	–	–

n.a. = not applicable

Table 15: Synthesised Valuation Matrix

Synthesised values (derived by personal judgement) from the medians of individually transformed data elicitied by ME, TTO and CRT valuation methods

DISABILITY STATES	DISTRESS STATES			
	A	B	C	D
I	[1.00]	.89 .90 .85	.89 .80 .85	.67 .45(.35)
		.90	.85	.55
II	.89 .90 .85	.81 .70(.45)	.78(.60)(.60)	.56 .35(.50)
	.90	.70	.60	.45
III	.70(.55)(.50)	.63 .53 .44	.57(.55)(.30)	.44 .20 .30
	.65	.55	.50	.35
IV	.63(.70)(.55)	.56(.45).42	.51(.55)(.40)	.40 .33 .22
	.60	.50	.45	.30
V	.44 .55(.25)	(.43).45(.35)	(.44(.43)(.26)	.22 .20 .17
	.50	.40	.35	.20
VI	.44 .43(.41)	(.44)(.45)(.39)	.34(.45)(.30)	.22 .15 .14
	.40	.35	.30	.15
VII	(.38).20 .20	(.40).10(.16)	.33 .03(.20)	.20[neg].0
	.30	.25	.20	.10
VIII	.01	[neg]		.00
	00			

The data in each cell is as follows:

ME	TTO	CRT
SYNTHESISED		

(--) indicates that the state was one of a pair manifesting inconsistency

Table 16: Cost per QALY Data for Williams (1985)

	Rosser's Original Matrix		Medians of Individually Transformed Data							
			Magnitude Estimation		Time Trade–Off		CR (Thermometer)		Synthetic	
	a	b	a	b	a	b	a	b	a	b
CABG										
Severe Angina										
LMD	1.04	1	1.27	3	1.27	3	1.14	3	1.27	3
3VD	1.27	2	1.63	4	1.63	4	1.90	9	1.63	5
2VD	2.28	4	1.90	7.5	1.90	7.5	1.63	4.5	1.63	5
1VD	11.40	12	1.90	7.5	1.90	7.5	1.63	4.5	1.63	5
Moderate Angina										
LMD	1.33	3	1.71	5	1.71	5	1.71	6.5	1.71	7
3VD	2.40	5.5	3.00	9	3.00	10.5	2.40	11	3.00	10
2VD	4.00	9	4.00	11.5	3.00	10.5	2.00	10	3.00	10
1VD	12.00	13	4.00	11.5	2.40	9	1.71	6.5	3.00	10
Mild Angina										
LMD	2.52	7	3.15	10	3.15	12	3.15	12.5	3.15	12
3VD	6.30	10	6.30	13.5	4.20	13.5	3.15	12.5	4.20	13
2VD	12.60	14	6.30	13.5	4.20	13.5	4.20	14	6.30	14
1VD	63.00	15	12.60	15	12.60	15	63.00	15	12.60	15

a = Cost per QALY b = Rank Order

Table 16 (continued)

	Rosser's Original Matrix		Medians of Individually Transformed Data							
			Magnitude Estimation		Time Trade–Off		CR (Thermometer)		Synthetic	
	a	b	a	b	a	b	a	b	a	b
PTCA										
Severe – 1VD	2.40	5.5	0.64	1	0.64	1	0.56	1	0.60	1
Moderate – 1VD	3.40	8	1.02	2	0.85	2	0.57	2	0.85	2
Mild – 1VD	10.72	11	1.79	6	1.79	6	1.79	8	1.79	8

a = Cost per QALY b = Rank Order

Table 17: Spearman's Rank Correlation Coefficients between the Various Sets of Cost-per-QALY Measures in Table 16

	ME	TTO	CRT	Synthetic
Rosser Original	0.70	0.63	0.47	0.63
ME		0.98	0.84	0.94
TTO			0.90	0.96
CRT				0.95

Table 18: Cost per QALY Data for Gudex (1986)

	Rosser's Original Matrix		Medians of Individually Transformed Data							
			Magnitude Estimation		Time Trade–Off		CR (Thermometer)		Synthetic	
	a	b	a	b	a	b	a	b	a	b
Scoliosis Surgery – Neuromuscular Illness	0.19	1	3.14	3	1.43	2	2.10	2	2.86	3
Shoulder Joint Replacement	0.59	2	0.27	1	0.22	1	0.30	1	0.25	1
Kidney Transplant	1.41	3	2.09	2	2.43	3	3.48	4	2.43	2
Scoliosis Surgery – Idiopathic Adolescent	2.62	4	*	7	15.72	5	*	7	31.43	7
Treatment of Cystic Fibrosis with Ceftazidime	8.23	5	4.70	4	2.99	4	2.53	3	3.29	4
Haemodialysis	9.08	6	13.50	5	15.82	6	22.14	5	15.38	5
Continuous Ambulatory Peritoneal Dialysis	13.43	7	19.03	6	21.75	7	25.38	6	20.76	6

* indicates that procedure generates a negative number of QALYs (i.e. the procedure is harmful) when calculated using the specified matrix

a = Cost per QALY

b = Rank Order

Table 19: Spearman's Rank Correlation Coefficients for the Various Sets of Cost-per-QALY Measures in Table 18

	ME	TTO	CRT	Synthetic
Rosser Original	0.68	0.93	0.68	0.68
ME		0.86	0.89	1.00
TTO			0.86	0.86
CRT				0.89

Table 20: QALYs per Hour of Operating Time: Treatment Now vs Treatment One Year Later
(Based on Gudex et al. 1990)

	Rosser's Original Matrix		Medians of Individually Transformed Data							
			Magnitude Estimation		Time Trade-Off		CR (Thermometer)		Synthetic	
	a	b	a	b	a	b	a	b	a	b
Anal Fissure	0.139	1	1.001	1	1.202	1	1.462	1	1.217	1
Bilateral Inguinal Hernia – Male	0.095	2	0.518	4	0.637	6	0.781	9.5	0.613	6
Ingrowing Toe Nail	0.068	3	0.696	3	0.872	2	1.049	3	0.874	2
Recurrent Inguinal Hernia – Male	0.059	4	0.483	5	0.587	8	0.824	8	0.609	7
Anal Fistula	0.055	5	0.721	2	0.866	3	1.092	2	0.855	3
Piles	0.050	6	0.470	6	0.698	4	0.860	7	0.637	5
Hyperhidrosis	0.049	7	0.385	10	0.510	9	0.596	14	0.473	11
Ganglion	0.037	8	0.463	7	0.618	7	0.906	5	0.594	8
Circumcision	0.035	9	0.426	8	0.651	5	0.959	4	0.643	4
Unilateral Inguinal Hernia – Female	0.032	10	0.397	9	0.506	10	0.781	9.5	0.528	9

Table 20: Continued

	Rosser's Original Matrix		Medians of Individually Transformed Data							
			Magnitude Estimation		Time Trade–Off		CR (Thermometer)		Synthetic	
	a	b	a	b	a	b	a	b	a	b
Incisional Hernia	0.027	11	0.376	11	0.423	15	0.534	16	0.444	13
Anal Tags	0.024	12	0.313	12	0.456	13	0.691	12	0.456	12
Skin Lesions	0.023	13	0.299	15	0.331	16	0.544	15	0.331	16
Gynaecomastia	0.022	14	0.298	16	0.461	12	0.731	11	0.431	14
Unilateral Inguinal Hernia – Male	0.018	15	0.302	13	0.440	14	0.665	13	0.430	15
Subcutaneous Lumps	0.017	16.5	0.225	18	0.250	21	0.410	22	0.250	21
Unilateral Varicose Veins – Male	0.017	16.5	0.301	14	0.478	11	0.881	6	0.478	10
Excision of Mole	0.016	18	0.257	17	0.286	18	0.469	19	0.286	18
Epigastric Hernia	0.014	19	0.206	19	0.301	17	0.427	21	0.297	17
Unilateral Varicose Veins – Female	0.010	20	0.176	20	0.263	20	0.477	18	0.263	20

Table 20: Continued

	Rosser's Original Matrix		Medians of Individually Transformed Data							
			Magnitude Estimation		Time Trade–Off		CR (Thermometer)		Synthetic	
	a	b	a	b	a	b	a	b	a	b
Bilateral Varicose Veins – Female	0.009	21	0.149	22	0.236	22	0.434	20	0.236	22
Bilateral Varicose Veins – Male	0.008	22	0.175	21	0.279	19	0.515	17	0.279	19

a = Cost per QALY
b = Rank Order

Table 21: Spearman's Rank Correlation Coefficients for the Various Sets of Cost-per-QALY Measures in Table 20

	ME	TTO	CRT	Synthetic
Rosser Original	0.97	0.91	0.78	0.92
ME		0.95	0.87	0.96
TTO			0.94	0.99
CRT				0.94

Figure 1: Rosser Classification of Illness States

Disability		Distress	
I	No disability	A	No Distress
II	Slight social disability	B	Mild
III	Severe social disability and/or slight impairment of performance at work Able to do all housework except very heavy tasks	C	Moderate
		D	Severe
IV	Choice of work or performance at work very severely limited Housewives and old people able to do light housework only but able to go out shopping		
V	Unable to undertake any paid employment Unable to continue any eudcation Old people confined to home except for escorted outings and short walks and unable to do shopping Housewives able only to perform a few simple tasks		
VI	Confined to chair or able to move around in the house only with support from an assistant		
VII	Confined to bed		
VIII	Unconscious		

Figure 2: Rosser's Matrix of Health State Valuations

DISABILITY RATING	DISTRESS RATING			
	A	B	C	D
I	1.000	0.995	0.990	0.967
II	0.990	0.986	0.973	0.932
III	0.980	0.972	0.956	0.912
IV	0.964	0.956	0.942	0.870
V	0.946	0.935	0.900	0.700
VI	0.875	0.845	0.680	0.000
VII	0.677	0.564	0.000	-1.486
VIII	-1.028			

Source: Kind, Rosser and Williams (1982)

Figure 3: Factorial Design used in Study