THE UNIVERSITY of York

CENTRE FOR HEALTH ECONOMICS
YORK HEALTH ECONOMICS CONSORTIUM
NHS CENTRE FOR REVIEWS & DISSEMINATION

The Measurement and Valuation of Health : A Chronicle

Alan Williams

DISCUSSION PAPER 136

THE MEASUREMENT AND VALUATION OF HEALTH

A CHRONICLE

Alan Williams
Centre for Health Economics
University of York, England

June 1995

Research Group on the Measurement and Valuation of Health ("the MVH Group") membership, 1987 to 1995:

Dalen, Harmanna van (1988 to 1991)
Dolan, Paul (1991 to 1994)
Durand, Mary-Alison (1988-1990)
Gudex, Claire (1987 to 1990 and 1991 to 1995)
Kind, Paul (1987 to date)
Lewis, David (1990 to 1991)
Morris, Jenny (1988 to 1991)
Williams, Alan (1987 to date)

Collaborators in Social and Community Planning and Research:

Erens, Bob
Thomas, Roger
Thomson, Katarina
and the key field workers and their managers

The Author

Alan Williams is Professor of Economics at the University of York, and founder of the EuroQol Group.

Further Copies

Further copies of this document are available (at price £6.50 to cover the cost of publication, postage and packing) from:

The Publications Secretary
Centre for Health Economics
University of York
York YO1 5DD

Please make cheques payable to the University of York. Details of other papers can be obtained from the same address, or telephone York (01904) 433648 or 433666.

ABSTRACT

Objective: to measure health-related quality-of-life in a way that reflects the salient features of health as perceived by a representative sample of the adult population of the UK.

Choosing a Descriptive System: with data from a survey of 600 people in the West Midlands we appraised 6 existing ways of measuring health-related quality-of-life and chose EuroQol.

Choosing a valuation method: Phase I: two direct methods (Time Trade-Off [TTO] and Magnitude Estimation [ME]) and two indirect ones (Pairwise Comparison and Category Rating [CR]) were studied. A survey of almost 300 subjects in the City of York led us to discard ME in favour of TTO.

Choosing a valuation method: Phase II: in the early stages of Phase II we tested TTO against SG on a within-subject basis. Our findings were that there was little to choose between them, but TTO had more complete data and more consistent valuations at individual level .

The Main Survey: 1993 Pilots: the first pilot indicated that respondents could not handle more than 15 states. The second was a "full dress rehearsal" for the main survey.

The Main Survey: Design & Execution: each interview consisted of: self-reported health; ranking of states; VAS rating of states; TTO rating of states; personal data. The main fieldwork was conducted in late 1993.

The Main Survey: Results: 3395 interviews were achieved, a response rate of 64%. The data on self-reported health showed that problems generally increase with age, and, within every age group, by social class too. With the VAS, median scores were all positive (ie every state was rated as better than being dead by a majority of respondents). Higher median scores were given by the lower social classes and by the less educated, meaning that they do not think that the poorer health states are as bad as the others do. With the TTO far more states were rated worse than being dead. There were some differences between men and women, and also according to marital status and employment status, but the most marked effect was age. At retest all 3 methods proved very reliable at both group and individual levels.

The Main Survey: Modelling the "Tariff": to interpolate values for the remaining 200 EuroQol states from the 45 on which we had direct valuations, the preferred model (known as "Dolan-N3") predicts the value of a health state from its components by attaching a (negative) value to each separate deviation from good health. Our basic tariff, for use when a weighting system is required for use in an economic evaluation, is the one of mean values based on the individual TTO scores. Not everyone may wish to use this basic tariff, though we recommend for comparative purposes that even if another one is preferred, the basic one is used too.

The Next Phase: The main future activity of the MVH Group is going to be the implementation of the benefit measures we have already generated. This is our next challenge.

THE MEASUREMENT AND VALUATION OF HEALTH

A CHRONICLE

THE TASK

No country can afford to do all the things that might improve somebody's health and thus some systematic method of establishing priorities is required. Commonsense suggests that resources should be concentrated where they will do the most good. In health care, "doing the most good" means maximising the improvements in people's health. Improvements in health have two broad dimensions: improvements in life expectancy, and improvements in quality of life. Various classification schemes have been developed to provide descriptions (or simple scores) for particular dimensions of health-related quality-of-life that are of particular relevance for patients with particular conditions (e.g. arthritis, cancer, heart disease, kidney disease). Because these "specific" schemes have much content in common, more general health "profiles" have been developed which encompass a broader range of dimensions than any one specific scheme, and which enable comparisons to be made in a more standard way across different conditions.

But because of the multi-dimensional nature of profiles they do not generate an overall "index" number to indicate the extent of any health benefit. Indeed, in cases in which a patient is better on some dimensions but worse on others, they will even fail to indicate whether on balance there is any benefit or not. For this purpose a "global" index is required, the descriptive content of which is neither condition-specific nor treatment-specific, and the valuation content of which is such that it generates a single number which summarises the relative value attached to each of the multi-dimensional health states it encompasses. The valuation process must also include the relative value attached to improved life expectancy on the one hand and improved quality of life on the other.

From its inception, the basic objective of the Research Group on the Measurement and Valuation of Health (henceforth "the MVH Group") at the University of York has been to find practical ways of measuring health-related quality-of-life (HRQOL) that reflect the salient

features of health as perceived by ordinary people. With this in mind, an important task for the MVH Group has been to elicit the valuations that ordinary people attach to different (multi-dimensional) health states.

This extremely ambitious task needs to be broken up into segments if it is to constitute a workable research programme. The key strategic decision made by the MVH Group was to investigate separately the choice of the descriptive system for health states, and the method of valuing them. The first task involved the elicitation of lay concepts of health. The second task (which turned out to be the much larger one) involved the testing of various valuation methods (and different practical means of administering each of them). The valuation task faced two further requirements: firstly, that, in order to fulfil the basic objective, the valuations had to be capable of being represented on a scale in which 0 = being dead, and 1 = being healthy: and, secondly, that, in order to be useful for general policy purposes, such valuations were needed from a representative sample of the general public.

BACKGROUND

The foundations for the work of the Group were laid by Rosser and her collaborators some years ago [Rosser and Watts (1972), Rosser and Watts (1978), Rosser and Kind (1978), Rosser (1983)]. The central feature of this approach is a simple descriptive classification defining 28 states in terms of disability and distress, plus a 29th state "unconscious", with "dead" as the (implicit) 30th state (Annexe A). A valuation matrix across these states was elicited from 70 respondents (Annexe B), who were not a random sample of the population at large, but a selection of fairly accessible doctors, nurses, patients and healthy volunteers. The original objective was to use this system to measure the "sanative output" of a hospital, i.e. the extent to which patients benefitted from hospital treatment.

Meanwhile a separate stream of methodological work from within health economics was developing the concept of the quality-adjusted-life-year as an outcome measure for use in health care. [Culyer, Lavers and Williams, (1972), Torrance, Sackett and Thomas (1973), Patrick, Bush and Chen (1973), Sackett and Torrance (1978)]. Williams noted that the Rosser valuation matrix (suitably transformed to a scale in which dead = 0 and healthy = 1) had the

appropriate properties for it to become the "quality adjustment" in the QALY concept, and this idea was then published jointly by Kind, Rosser and Williams (1982) and applied for the first time by a small group of economists in the now classic study of the Economics of Coronary Artery Bypass Grafting (Williams, 1985a).

The enormous interest generated in the potential of this approach led the York Centre for Health Economics to devote more resources to its development, and in particular to pursue the following activities:

(a) Conducting surveys on people's attitudes to health (see, for instance, Wright, 1986);

(b) Developing a simple self-assessment questionnaire to describe patients' current quality-of-life (see Kind & Gudex 1994; Gater et al 1995);

(c) Helping the various clinicians and medical researchers who have approached us to incorporate quality-of-life measurement into their studies;

(d) Incorporating such measures, where appropriate, in studies with which we were associated (e.g. studies of CT and MRI) (Kind & Sims 1987; Hutton & Williams 1988);

(e) Setting up such studies de novo wherever the opportunity arose (e.g. on the general surgical waiting list at Guy's Hospital, and a nationwide study of the QoL of patients in end-stage renal failure, in collaboration with EDTA) (Gudex et al 1990, Gudex 1995 forthcoming);

(f) Assisting in the use of the QALY concept in priority-setting at management level in the NHS (e.g. for the North West Regional Health Authority) (Gudex 1986);

(g) Liaising more closely with other British researchers working on global indexes (e.g. Buxton, Rosser);

(h) Seeking out other European groups engaged on parallel exercises (Dutch, Finnish, Swedish)(later to become the EuroQol Group);

(i) Establishing direct face-to-face contact with the US and Canadian groups working in this same territory, to ensure prompt exchange of ideas;

(j) Devising ways of strengthening and developing the measurement of health benefits generally (Williams 1985b, Williams 1987, Williams 1988a);

(k) Conducting a pilot study of people's attitudes concerning the extent to which the NHS should discriminate between different sorts of people (e.g. old versus young) in the distribution of the benefits of health care (Williams 1988b)

Out of these diverse activities grew the MVH Group at York.

PHASE I - 1987 to 1990

Against the background described above, in 1987 we proposed a major programme of development work to the then DHSS, which would extend over 4½ years and the core of which would be "a major survey designed to elicit the relative valuations of approximately 2000 respondents, to replace the current Rosser classification and its associated valuation matrix". However, in order to prepare the ground as thoroughly as possible for this, the first 2½ years of the project (Phase I) would be devoted to three preliminary tasks:

(a) to establish whether a Rosser valuation matrix based on the views of a sample of the general population would be similar to or different from the existing matrix based on a convenience sample of 70 respondents;

(b) to establish whether the descriptors used in the Rosser Classification were the most suitable ones to use in future work, and, if not, what descriptive system should be used in its place;

(c) to establish which of the available valuation/scaling methods would be the best one to use in the major study.

This preliminary work would proceed on as large a pilot sample as could be afforded, and to do this we were supported with core staff financed by the ESRC, and fieldwork financed by the Nuffield Provincial Hospitals Trust (NPHT), as well as DHSS support. It was envisaged that there would be a review of progress about 18 to 24 months into the study, which, if favourable, would lead to the release of the funds earmarked for Phase II.

(a) Replicating Rosser

For this task we used Rosser's descriptive system (see Annexe A) and the "magnitude estimation" valuation method. In its conventional form this involves asking how many times worse is each state than some reference state. Rosser had posed the question in a different way, first asking how many times worse was State X than the reference state, and then how much worse was state Y than state X, and then Z compared with Y, and so on. We used the conventional method, in which the same state is the reference state throughout, so it was not an exact replication. Moreover, it was not possible for each subject to value all 28 states, so they were divided into 2 subsets, with six states common to both. We got very different valuations from our sample of 140 members of the general public compared with those that Rosser had elicited, as can be seen from Annexe B.

(b) Finding the best descriptive system

This was a much more elaborate exercise. In order to establish our own baseline data on what the general population regard as the salient features of health, so that we could appraise different approaches to the construction and content of generic (health related) quality-of-life measures, we conducted a survey (financed by the NPHT) in the West Midlands. The aim was to recruit the following:

(a) A random sample of 200 members of the general public aged 18 and over.

(b) 100 physically disabled young people, aged between 18 and 25 years and living at home; and 100 able-bodied young people, aged 18 to 25 and living at home, to act as controls.

(c) 100 individuals who have been caring (at home) for physically disabled children who are now young adults; and 100 individuals who have brought up able-bodied children who are now young adults, to act as controls.

The interview schedule was carefully designed so that it was identical for all respondents. The main body of the interview had three phases:

(i) an unprompted section in which we elicited what individuals thought were the distinguishing features of good or bad health in themselves or in others.

(ii) a prompted section in which they were presented with 37 statements about health, which they were asked to endorse using a series of categories from "very important" to "not at all important".

(iii) a section in which 6 groups of statements, each representing a particular concept of health, were presented to subjects, who were asked to indicate which of successive pairs of such concepts better represented their own notions of health.

As well as this core data, information was collected from each subject on sociodemographic characteristics, Health Locus of Control, Eysenck Personality Questionnaire, and the Euroqol health questionnaire.

The most important general finding from our survey data was that although initially, when encouraged to offer unprompted ideas about health or illhealth in self or others, the presence or absence of diseases or symptoms played a significant role, later, when offered items to rate in order of importance, and later still, when asked to choose between broad conceptualisations of health, the notion of health as simply not being ill faded into insignificance, and notions of functional capacity, feelings and general fitness came to predominate.

For our immediate purpose, the data was used to appraise existing (and proposed) instruments for measuring health-related quality-of-life. The unprompted health items were used to see what proportion of the items expressed by the public are actually covered by the different instruments.

With one exception, the instruments included in this comparison were all generic measures designed to yield a single index number, e.g. the Rosser Index, EuroQol, Sickness Impact Profile (SIP) and the Quality of Well-Being Scale (QWB). The exception is the Nottingham Health Profile (NHP), which was designed, but some users have nevertheless converted it from a profile to an index simply by adding together the different items across dimensions. The SIP and QWB are based on US weights (though there is a UK adaptation of the SIP called the Functional Limitations Scale). Rosser has English weights, and at the time the Euroqol had English, Dutch and Swedish weights. The greater complexity of the NHP, QWB and SIP measures is largely due to their aim to pick-up relatively specific variations in health status which may be of particular significance in particular circumstances. The Euroqol measure on the other hand was not intended as a "stand alone" instrument, but as a comparative tool to be used alongside more specific instruments which were not directly comparable with each other. It therefore had a less complex descriptive system.

Concentrating only on the general population sub-sample (of 196 respondents), and pooling all of their unprompted responses (whether relating to health or illhealth, or to self or others), we arrive at the data set out in the left-hand columns of Annexe C. In the subsequent columns are indicated those items which are included in each of the five descriptive systems that are being compared. From the bottom line it will be seen that coverage generally increases as the number of items of information collected increases. It seems that the simple systems provide about 26-36% coverage of the items mentioned spontaneously by our respondents, and the more complex systems around 50-60% coverage. But about a third of the items mentioned are not relevant for a general health state index designed to appraise clinical interventions or variations in health-related life-style. Of the remaining omitted items the most significant is that relating to energy and tiredness (10.5% of all mentions). The omission of this item accounts for most of the difference in coverage between the Rosser and Euroqol instruments on the one hand, and the NHP and SIP on the other. Judging by our

data, it was a strong candidate for inclusion. [An experimental version of the Euroqol Questionnaire, with energy/tiredness included as a sixth dimension, was subsequently tested in a pilot study. The additional dimension was found to have such a small impact on the valuations of the health states concerned that, in the interests of parsimony, it was not included in the standard Euroqol Instrument.]

We also tested the importance of items, as rated by respondents in the prompted section of our interview material. This was summarised as the percentage of the general population rating each item as "very important", averaged over the items covered by each instrument. There was little to choose between the various instruments on these grounds.

A further consideration to be borne in mind was the sheer size of the classification system offered by each instrument. If it were very small, then fine discrimination between health states would not be feasible. If it were very large, direct valuation of each health state would become impossible, and short-cut methods would have to be adopted to fill the gaps. Of those considered here, Rosser's is the most parsimonious, with only 29 states (excluding dead). The Euroqol system has 244 states (excluding dead), which means that valuations can only be conducted on a subset of them, the rest (apart from "unconscious") being estimated by a formula working in the 5-dimensional space. The NHP generates a very large number of possible states (more than 10,000!). Since any of the SIP's items may appear singly or in combination with one or more of the others, again the number of logical possibilities is enormous (more than 100,000!), and the valuation problem is "solved" by simply attaching a score to each item and using it additively whenever it appears. A similar consideration applies to the rather more complex two-stage system adopted by the QWB scale, which also generates over 100,000 different possible states, and deals with the valuation problems by using simple additive weights for the 40 adjustment factors in the "symptom-problem complexes".

Against this background, the task before us was to balance five considerations against each other with respect to each instrument, namely information demands, coverage, importance of items, complexity, and scope for full valuation of states. In summary, the situation was as follows:

	Rosser	Euroqol	NHP	QWB	SIP
Information Required for classification	2 items	5 items	45 items	43 items	136 items
Coverage	37%	39%	58%	59%	49%
Average importance	54	57	54	55	57
Complexity	29 states	244 states	>10,000 States	>100,000 States	>100,000 States
Valuation strategy	Inter-active & complete	Inter-active but selective	No over-all val-uations	Additive & complete	Additive & complete

Both the SIP and the QWB seemed to be too complex for our purposes, the NHP was inappropriate (being a profile measure), and the Euroqol seemed slightly better than Rosser if interpolated valuations derived from a subsample of directly valued states were acceptable.

(c) Choosing a valuation method

The following criteria were employed in deciding which valuation methods would be studied.

1. use in other relevant studies
2. methodological importance
3. efficiency
4. ease of use in large scale studies
5. type of method (direct or indirect valuations)
6. orientation (individual vs aggregate scales)

The Equivalence Technique was excluded because it introduced an additional element, interpersonal comparisons, which were to be taken up explicitly at a later stage in the research programme. Standard Gamble, Magnitude Estimation, Time Trade-Off, Pairwise Comparison and Category Rating had all been used in a number of relevant studies and all were considered to be methodologically important. It was decided to test two direct methods and

two indirect ones: Standard Gamble, Time Trade-Off and Magnitude Estimation fell into the former category, and Pairwise Comparison and Category Rating into the latter. Standard Gamble was rejected because it was considered to be too time-consuming, and because it was concurrently being studied by Rosser's group at the Middlesex Hospital. Magnitude Estimation, Time Trade-Off, Category Rating and Pairwise Comparisons were therefore the methods chosen for use in the pilot study. Furthermore, it was decided that three variants of Category Rating would be employed: visual analogue (thermometer)(CRT), labelled boxes (Likert version)(CRL), and numbered boxes (CRN). The various methods are described in Annexe D. We envisaged that in the main survey we would eventually use one direct method as our "main" method (for possible use in economic evaluations), and one indirect method (for possible use in other types of evaluation).

Once more we proceeded by conducting a survey, this time of almost 300 subjects in the City of York. We did not manage to achieve a close match to the population of England and Wales, but we did achieve a fairly wide spread of characteristics amongst our respondents (apart from ethnicity). The core of each interview was made up either of an ME task, or of a TTO task, accompanied by one of the 3 variants of the CR method, which sometimes preceded the other task and sometimes succeeded it. In addition the same supplementary data was collected as with the lay concepts of health study. By a complex factorial block design the different combinations of tasks were randomised between subjects, between interviewers and between geographical areas. To avoid subjects being overloaded, each valuation task could cover only a subset of all 29 Rosser Health States, but the factorial block design also incorporated a careful distribution of these subsets so that each state was valued by at least 35 people.

We devoted a great deal of attention to identifying "inconsistent" responses (based on the assumption that in Rosser's Classification the disability states get successively worse, as do the distress states). Rather than rejecting inconsistent data as unreliable, we decided to analyse it carefully to see what information it yielded, and to retain it if at all possible. We found that the inconsistency rates of individual subjects were particularly high for older people from manual occupations. We also found that they varied by interviewer. Order of presentation of task made no difference. Although it is clear that some subjects (the older

people from manual occupations) experience greater difficulties than most people, no clear explanation has been found for the inconsistencies produced in all methods. Judging by the experiences of others who have reported such inconsistencies (but then discarded the inconsistent data) some "random noise" is to be expected, and the problem is how to minimise it, rather than how to eliminate it. One limited explanation for some of our inconsistencies is that some of the Rosser descriptors seem difficult to digest or conceptualise (for instance not everyone may regard "moderate" distress as being worse than "mild" distress).

Of the two direct scaling methods considered here (ME and TTO), there seemed little to choose between them on grounds of ease of completion, but TTO generated less inconsistencies. Both were more difficult for respondents than category rating methods. Within the three category rating methods, the Thermometer method generated the most inconsistencies, yet is rated the easiest to complete.

In the middle of this valuation study we obtained copies of the Manual and Standard Gamble Board, devised by the McMaster Group in Canada, which was intended to facilitate the use of that valuation method, which we had initially excluded as being too complex (see Annexe D for a brief description of this method). We decided to extend our study by inserting the SG method into our design combined with a category rating (CR) task. Since the resources available to finance extra interviews were severely constrained, we were limited to 72 interviews (i.e. only half the number who did TTO and ME) so we decided to present CR before SG and forgo any test of the effect of the order of presentation. As recommended by the McMaster Group, we limited coverage of the Rosser States to eight, which were always presented in the same order. Unfortunately, when we came to process the data we found an error in the Manual that had been supplied to us, and which we had followed carefully, and this error rendered this supplementary study abortive. We were thus no further forward in judging the relative merits of SG.

To convert the raw data from each of our methods into a comparable valuation matrix requires transformation of the data to a scale in which dead = 0 and healthy ("No disability and No distress") = 1. For the ME, TTO and CRT methods it is possible to make this transformation for each individual separately, and then construct a group matrix from the

medians of these transformed valuations for each state. For the CRN and CRL methods this transformation can only be done on the medians (or means) of the raw scores. The data was also processed as if ordinal rather than cardinal. The resulting valuation matrices contained a fair number of "reversals" of logical orderings, some of which were due to the partitioning of the states into small subsets which were valued by different subsamples of the population. Again TTO showed up quite well and CRT badly by this test. Reversals tended to concentrate around certain Rosser states, suggesting that the descriptive system itself may have been partly responsible. But the rank ordering of states was similar for all valuation methods.

Our conclusion at this stage was therefore that ME should be discarded in favour of TTO, but since we had been unable to test TTO against SG, this task remained. CRT proved easier to do than TTO or ME but was much less reliable in the data it generated. It is, however, the preferred method for postal questionnaires in the Euroqol Group, so needed to be kept in play for that reason. We therefore proposed to test SG against TTO in the early stages of Phase II, and to use the "winner", plus CRT, in the main valuation study.

PHASE II - THE PILOT STUDIES - 1991 to 1993

At this point a major shift occurred in the way the MVH Group worked, in that instead of organising and conducting our survey work ourselves, we joined forces with Social and Community Planning and Research (SCPR), a London-based research-orientated survey organisation with a strong interest in attitudinal research. All future survey work was designed jointly between SCPR and the MVH Group, but it was carried out by SCPR field staff. Amongst the joint responsibilities were the training (and de-briefing) of interviewers and quality control of data from the fieldwork, which became a central feature of the next stage of the work as we sought to find the best way of getting good quality data from the various valuation methods that were still in play.

The 1992 Pilots

As planned, the early stages of Phase II were concerned with testing TTO against SG as the main valuation method to be used in the main survey. From the literature it was evident that in principle neither method could be regarded as a "gold standard", each having its advocates and detractors. We therefore decided to base our selection upon more practical considerations, namely:

1. LOGICAL CONSISTENCY: the extent to which the health states used are given a logical ordering within each method.

2. VALIDITY – Concurrent: the extent to which people's valuations correspond between methods within each subject. Discriminant: the extent to which valuations differ (in accordance with prior expectations) by respondent characteristics.

3. TEST–RETEST RELIABILITY: the extent to which respondents' responses are stable within each method over a relatively short time interval.

4. COMPLETENESS: the extent to which each method produces a complete data set.

There were some further criteria which, although not sufficiently important to be used in choosing between methods, nevertheless offered additional evidence on the performance of the methods. They mostly concern the burden placed upon subjects and interviewers, and they needed to be taken into account when designing the main survey. They were:

(a) The time taken to complete each task
(b) The difficulty of each task as reported by both respondents and interviewers
(c) Respondents' willingness to be re–interviewed

The principal research objective was to achieve within–subject comparisons for the two main methods under review. Each method was tested in two variants, one of which used specially designed boards and cards as an aid to decision–making by respondents (Props), and the other

used a self-completed booklet (No Props). Since we felt that the average length of an interview should be about one hour, we abandoned any attempt to generate enough valuations in this pilot survey to make it possible to estimate the valuation space generally, so we used only 6 health states (apart from healthy and dead).

The target population were adults aged 18 and over in the general population, with no upper age limit. A random sample of 700 addresses was drawn from 11 regional areas in the U.K. using the Postcode Address File. The fieldwork was carried out between March and May 1992.

Of the 525 "in scope" addresses, 190 (36%) yielded refusals and 335 (64%) yielded an interview. A sub-sample of those who had said they would be willing to be re-interviewed were approached again 4 to 12 weeks after the original interview. Respondents were asked to do exactly the same tasks as before, with the additional question of whether anything important had happened to them since the last interview.

Each interview followed the same pattern, namely: description and rating of own health state (using the EuroQol Classification as in Annexe E); ranking of health states; category rating (using the thermometer as in Annexe F); SG followed by TTO (or vice-versa); sociodemographic background data. The questionnaire had an additional section requiring interviewer feedback.

Compared to the general population, in the survey population there were more people with no children at home, and more with a degree or professional qualification. There were also slightly fewer people aged under 20 and slightly more aged over 60, and fewer in paid work (20% were retired). This seems to indicate that there may be some response bias in favour of the more educated, and that people with children at home are less willing to undertake a rather time-consuming interview. But there seemed to be no difficulty in eliciting the co-operation of older people (though they had greater difficulty with some of the tasks). On average the time taken for an interview was just over an hour, but the time taken at retest was, on average, shorter. Of the 14 respondents with incomplete interviews, 71% were aged 61 or over, and none was in paid work. These were important pointers to problems that we

might face in the Main Survey.

Turning to the criteria for choice set out above, our findings were as follows:

Logical consistency: there were no significant differences between the methods, but TTO props performed slightly better than the others, and seemed more robust to characteristics such as age and education level.

Validity: TTO props avoided most of the more extreme valuations and it also identified some of the significant differences related to respondents' own health states, but these observations are rather tentative in view of the difficulty in establishing "correct" values.

Test–retest reliability: the median values for all states were extremely reliable for all methods, but at individual level valuations were most consistent with TTO props.

Completeness: TTO props was the best of the four main methods.

Thus, although there was not a lot in it, as regards the quality of the data, there was an accumulation of evidence in favour of TTO props, so this was chosen as the best valuation method to use in the main survey.

Whichever method had been chosen, there would remain some consequential problems to be considered, mainly relating to the conduct of the interview itself. The "props" variants took 4 or 5 minutes longer on average than their corresponding "no props" variants. This may be because a larger percentage of states were considered to be worse than death in the "props" variants. In addition, more states were considered to be worse than death on TTO than SG. From this it will be noted that the method we chose is the one that takes the longest. This indicated a need to explore further ways of streamlining the presentation of the TTO Props method in the interview situation. TTO Props was not the easiest of tasks to understand from the respondents' point of view, although nor was it the hardest. Interviewers considered TTO Props to be the most easily understood of the major tasks.

The 1993 Pilots

The number of states that could be valued within a single interview was a key variable in determining the required sample size for the Main Survey. So although we had now chosen our preferred valuation method, there remained two outstanding problems that needed to be resolved before embarking upon the Main Survey. The first of these was to work out with the interviewers ways of making the TTO Props method as user–friendly as possible, from both the interviewer's and from the respondent's perspective. The second was to discover, in the light of this, the maximum number of states that could be valued by a respondent, within the context of a one–hour interview in which the TTO method would be preceded by a ranking and a rating exercise. These tasks occupied us for most of 1993.

To pursue the first task a brainstorming session was held between the MVH Group, 2 senior members of the SCPR staff, and six of the interviewers who had used the TTO Props version in the 1992 Pilot Study. This led to various suggestions for simplifying the choice process in the TTO method. These modifications were then tested in the field in the first pilot. This pilot also tested the bisection method for use with the CRT scaling exercise (see Annexe D). The purpose of this "bisection" process is to ensure that the resulting valuations have interval scale properties (Stevens 1971). We were here simply testing its feasibility however. Finally, the number of states used by each respondent was raised to 15 to see whether this number could be handled under the new procedure.

The outcome was that the revised procedures worked well and appeared to be understood by the interviewers; the bisection approach also worked well in practical terms; and most respondents were able to evaluate the full set of states, but interviewers felt that 15 states should be regarded as an absolute upper limit. Interviews took an hour on average, which was the target. One new problem arose with the TTO method, however, which was that at the mild end of the spectrum quite a few respondents were unwilling to give up any time to improve a health state, so it was decided to allow some extra "fine tuning" at this end of the scale.

The second pilot in 1993 was designed as a "full dress rehearsal" for the main survey. One

of the purposes here was to test the procedures as stringently as possible by training a completely new batch of interviewers and exercising rigorous quality control over the data they generated. It transpired that the key aspect of the interviewer briefing was the conduct of practice interviews under the guidance of experienced staff who acted as dummy respondents. This was the system eventually adopted for the Main Survey.

Sample size

The final matter that had to be determined before embarking on the main survey was the sample size. In the TTO method the smallest difference that it would generally be possible to express would be .025 (3 months out of 10 years). To detect such a difference at the 5% significance level with 80% power would require 3235 valuations for each state. But although some states would be valued by all respondents, 36 states would be valued only by 25% of respondents, so nearly 13,000 interviews would be required. In the end we settled for a sample size of 3235, which meant that we would have only about 800 valuations for most of the 45 states to be included in the survey. On that basis we expected to be able to detect a 0.1 difference in valuations between subgroups at the 5% level of significance.

THE MAIN SURVEY - DESIGN & EXECUTION - 1993

The objective of the Main Survey was to elicit the views of a representative sample of the non-institutionalised adult population of England, Scotland and Wales, by interviewing them in their own homes. Broad geographical coverage was required in case it emerged that there were marked regional differences in valuations.

The Euroqol Classification generates 245 theoretically possible health states, some of which are unlikely to occur in practice. Respondents cannot handle more than 15 each, and about 40 are required for modelling purposes (ie to estimate valuations for the states that are not directly valued). Valuations for two of the states ("unconscious" and "dead") cannot be estimated from the valuations given to any other state, so must be directly valued. The state 11111 ("healthy") is essential to the re-scaling of the VAS (thermometer) data, so must also be directly valued by all respondents. For all other states we had discretion. In exercising

that discretion we had several considerations in mind. First of all, we wanted the states to be widely spread over the valuation space in terms of mildness or severity (as indicated from earlier valuation data). Secondly, we wanted the set of states to include all plausible combinations of "levels" across each of the 5 dimensions, so as to be able to test for significant interaction effects (for example, to test whether the weight given to "moderate pain or discomfort" is different if it is combined with "some difficulty in walking" from what it is when combined with "moderately anxious or depressed"). Thirdly, we wanted to stay as close as possible to the selection of states that had been used by Finnish EuroQol colleagues in a major postal survey which they had just conducted. Fourthly, we wanted to exclude states which seemed prima facie implausible to respondents, so as to sustain motivation and credibility. The result of applying these criteria was the selection of states shown in Annexe G. The reason why the states in that Annexe are stratified in the way they are is that, apart from unconscious and dead (for which we had to have valuations from everyone), we wanted the two "reference" states (11111 and 33333) to act as a common frame of reference for all respondents. We also wanted each individual to have in their valuation set 2 of the 5 mildest states (11112, 11121, 11211, 12111, and 21111). Amongst the remaining 36 states we needed to ensure balance at individual level between the relatively "mild", "moderate" and "severe" states. Thus 3 out of each group of 12 states were randomly selected within this stratification system for each individual respondent.

The core of each interview contained five elements:

Self-reported health
Ranking of states
VAS (Thermometer) rating of states
TTO rating of states
Personal background data

Respondents were also asked whether they would be willing to be re-interviewed at a later date, because in order to test the reliability of the three valuation methods, a representative sub-sample of approximately 200 respondents was to be interviewed approximately 3 months after the original interview.

A great deal of emphasis was placed on interviewer training. All interviewers attended personal briefings (held in Birmingham, London, Manchester and Newcastle) which involved intensive training in the three valuation methods. Any interviewers who appeared to be having problems with their first few interviews were asked to attend a half-day re-briefing session before they were permitted to carry on with their assignment (13 were so recalled).

The main fieldwork was conducted between August and November 1993, and the reinterviews during December 1993.

THE MAIN SURVEY - RESULTS - 1994

Of the 6080 addresses selected for sampling 706 (12%) were found to be `out of scope' of the survey, being non-residential, empty/derelict, untraceable, or even not yet built. Of the remaining 5324 addresses, 3395 interviews were achieved, giving a response rate of 64% on in-scope addresses. After the survey data had been weighted to correct for the effect of varying household size on selection probabilities, the sample was found to have nearly identical characteristics as the general population.

There were few *missing data* from 3395 respondents,and logical consistency within method was also surprisingly high, with an average of 97.5% on the VAS and 93.8% on the TTO. Four separate data sets were assembled: a ranking data set, a VAS data set, a TTO data set and a combined VAS and TTO data set. Some respondents have been excluded from each set on the grounds of missing data and logical inconsistency but, despite stringent criteria, the numbers are extraordinarily small: 107 (3.2%) from the VAS data set, 58 (1.7%) from the TTO data set and 398 (11.7%) from the combined data set. Although the excluded respondents have tended to be those older than 60 years and with no educational qualifications, the respondents remaining in each data set are still representative samples of the general population. The entire data set has been deposited with the ESRC Survey Research Archive.

As an incidental byproduct of our survey we assembled data on the self-reported health of a representative sample of the non-institutionalised adult population of the UK. Some

illustrative examples of our findings are given in Annexe H. For example, we found 33% of respondents reporting pain or discomfort, and 21% reporting anxiety or depression. Health problems generally increase with age, and, within every age group, by social class too. These data provide evidence for the validity of the EuroQol instrument as a measure of health-related quality-of-life.

Each respondent rated 15 health states on the Visual Analogue Scale shown in Annexe F. In order to compare scores from different respondents, these `raw' scores have been adjusted relative to two states that all respondents rated: the state 11111 (full health) (set equal to 1), and death (set equal to 0). Both median and mean scores were logically consistent, and median scores were all positive (ie every state was rated as better than being dead by a majority of respondents). Significantly higher median scores were given by the lower social classes and by the less educated, meaning that they do not think poor health states are as bad as the others do.

The set of valuations emerging from the TTO task contained no logical inconsistencies, but far more states were rated worse than being dead than was the case with the VAS valuations. There were some significant differences in valuations between men and women, and also according to marital status and employment status. But the background factor which had the most marked effect was age (which had no effect in the VAS data). Looking at the age effect more closely it appears that respondents over the age of 60 give significantly lower values to the more severe states than do the rest of the population. One possible explanation is that as people's life expectancy shortens, they see less reason to tolerate suffering during their remaining years. An alternative explanation might be that it is an artefact of the TTO method. If respondents do not believe that they have 10 years life expectancy, they might willingly give up these 'excess' years, thereby depressing the apparent value attached to the more severe states. This puzzle we came back to later.

The relationship between the VAS valuations and the TTO valuations did not appear to be the power relationship found in earlier studies (see for instance Torrance 1976), but a "spreading" relationship, in which the TTO valuations are more extreme than the VAS ones at both ends of the valuation spectrum, but especially with respect to the more severe states.

This means that people are relatively unwilling to sacrifice life expectancy to improve mild states, but relatively more willing to sacrifice it to avoid severe states, where "mild" and "severe" relate to their VAS ratings.

At retest all 3 methods proved very reliable at both group and individual levels, and the already low inconsistency rates declined to still lower levels.

THE MAIN SURVEY - MODELLING THE TARIFF - 1994

At this stage in the proceedings we had valuation data on 45 states, from which we needed to interpolate values for the remaining 200 states in the EuroQol Classification. One of these states, "unconscious", lies outside the 5 dimensional scheme, and has been directly valued by all respondents. The states 11111 ("healthy") and "dead" act as calibration points in the valuation scale, so it is the valuations given to the other 42 that constitute the core data set. The data set used for all the modelling activity was that which contained the valuations of the 2997 respondents for whom we had complete data over both the VAS and TTO valuations (ie the "combined VAS and TTO data set" referred to earlier).

In estimating valuations for those EuroQol states on which we did not have direct valuations, we enlisted the collaboration of those members of the EuroQol Group with a special interest in modelling, and also obtained the services of Ian Russell and Mona Abdalla to act as external statistical consultants during this phase of the work. Each of the four participating groups was given access to our data set, and invited to enter a sort of competition to come up with the "best" model. The following criteria were used to help us choose the "best" model:

1) Goodness-of-fit i.e. how well the model explains the differences in the valuations given to those states on which there is direct data.

2) Parsimony i.e. the simplicity of the model.

3) Consistency i.e. states that are logically worse must have lower predicted values.

4) Transparency i.e the ease with which non-experts can understand the manipulations made.

Another member of the EuroQol Group, who was not in the "modelling competition", provided an independent critique of the various approaches when the results of all these different analyses were presented at the Plenary Meeting of the EuroQol Group, held in London in October 1994. At this meeting, it was decided that the model presented by Paul Dolan (a member of the MVH Group) satisfied the above criteria most fully. Before discussing this model in more detail, it is encouraging to note that the results presented by Abdalla and Russell using a different technique corresponded closely to the Dolan model.

Essentially the preferred model (known as "Dolan-N3") predicts the value of a health state from its components by attaching a value to each separate deviation from good health. In the EuroQol system there are 10 such "decrements" in health, made up of a moderate and a severe level of dysfunction for each of the five dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression). The model contains two other terms, one of which "N3" is active whenever any of the 5 dimensions of health is "severe" (ie at level 3), and the other is simply a constant term (which might be interpreted as the loss of value involved with being in any kind of dysfunctional state whatever). This approach has been used by the MVH Group for the estimation of all its "tariffs" of social values, so that they are all based on a common analytical approach. But the data used for each tariff is different, and there are some important matters of interpretation and use which need careful consideration. It is to these that we now turn.

The structure of the analytical work surrounding this modelling work is shown schematically in Annexe I. At top centre is the modelling data set, containing both VAS and TTO valuations from each respondent for the 15 states in that person's set. These data can be used in two different ways: either treating each individual as a separate observation, or taking the median values for each state, thus simplifying the situation but losing information. We have used the first approach when generating tariffs of mean values, and the second approach when generating tariffs of median values. The main modelling activity is represented straight down the middle of the chart, in the generation of "tariffs" of social values for all the EuroQol

states, representing the mean or median values of the general population, as elicited either by the VAS method or by the TTO method. The other items on the chart will be described later.

Our basic tariff, when a weighting system is required for use in an economic evaluation, is the one of mean values based on the individual TTO scores. The reason for taking this as the base case is that the use of individual scores retains the maximum amount of data, and the use of TTO scores means using a valuation method which involves tradeoffs, which in an economic context is more appropriate than a simple rating scale. There remains the choice between means and medians. The median value for a state is the value given by the person in the middle of the distribution, so it is insensitive to the particular valuations provided by people at the extremes of the distribution. Many people prefer to use medians as the measure of central tendency when a distribution is strongly skewed (as these distributions are). By contrast, means give every respondent some weight, but are sensitive to "outliers". In this case the "outliers" fall into two groups. There are those who rated the "mild" states as exceptionally good, resulting in means which exceed medians at the upper end of the scale. And there are those who rate the severe states as exceptionally bad, hence the mean values for the severe states tend to be much lower than the medians. The basic tariff is given in Annexe J, which also contains an explanation of how the tariff is calculated from the coefficients of the Dolan–N3 model.

Not everyone may wish to use this basic tariff, though we recommend for comparative purposes that even if another one is preferred for a specific purpose, the basic one is used too. The possible reasons for preferring a different tariff are many and varied, and we have tried to cater for as may of them as our data permits. Thus, although not shown here, we have tariffs for medians as well as means, for VAS as well as TTO, VAS tariffs for different educational levels, and TTO tariffs for different age–groups (and for each sex within age groups). All this is indicated at the bottom of the chart in Annexe I.

OTHER MATTERS

The valuations of the elderly

The chart in Annexe I also indicates some other matters that have required our attention. As was mentioned earlier, when using the TTO method the older members of the population rated the more severe health states as very much worse than did the younger people, and we puzzled over this, wondering whether it was genuine or an artefact of the method (since we did not observe this from the same people when they were using the VAS method). Fortunately we had the services of Angela Robinson (of Newcastle University) during the summer of 1994 to go out and reinterview a sample of our respondents in the North East of England, and get them to talk their way through the valuation process as they were doing it, to see whether this yielded any clues as to why their valuations were as they were. As far as possible her interviews followed the protocol originally used in the Main Study, but since the nature of her study was qualitative, rather than quantitative, respondents were asked to value only 7 states (as opposed to 15 in the Main Study). Respondents were asked to `think aloud' as they completed the interview, and to explain why they made certain decisions during the TTO exercise. The findings with respect to the TTO valuations from the elderly can be summarised as follows :

(a) no evidence was found to support the view that variation in values was primarily an artefact of the TTO method

(b) evidence was found of a `threshold of tolerability', below which states would have to fall before some respondents were prepared to give up even a few days, let alone months or years of life, to get out of them

(c) no convincing evidence was found that the elderly are more concerned than younger respondents about becoming a burden to their families

(d) older respondents appear genuinely more likely than younger respondents to consider severe states as worse than death

It appears that the key artefactual element is that older people do not believe that after being in a very severe state for any length of time they will in fact recover full health, whereas younger people do believe this. So the time they are going to gain is thought of as being time in poor health, not, as required by the method, time in good health.

Since it appears that the low valuations from the elderly are partly an artefact and partly genuine, we explored the possibility of modifying the TTO tariffs for the elderly (and the TTO tariffs for the general population) to eliminate the artefactual element. This was possible because there is nothing in the TTO method which would lead older people not to indicate accurately whether they considered a state to be better or worse than being dead. The problem arises when they come to attach a value to the states that are worse than dead. If at that stage we assumed that their valuations are the same as anyone else who rated that state worse than dead, we would have an estimate of the maximum effect that could be attributable to the artefactual element. This would still leave older people valuing the more severe states lower than younger people would, because they are more likely to rate such states worse than dead. We have calculated such modified tariffs should anyone wish to use them.

The duration of health states

Another complicating factor is that the valuations were derived for states lasting 10 years, which means that the tariffs are most appropriate for chronic conditions, or for "before" and "after" comparisons when patients' health states have stabilised once more. It is reasonable to suppose that a severe condition which lasted only a short time would be more tolerable than if it lasted a long time, and to test this we conducted a supplementary survey of 312 subjects from the main survey, with a 76% response rate. Only the VAS method was used, and the supposed duration of each state was taken to be 10 years (to maintain comparability with the main survey), 1 year, and 1 month. For 38 of the 43 states valued, the value attached to it when it lasted one month was significantly higher than when it lasted for ten years, but there is much less difference between durations of 1 year and 1 month. When modelled, the key element causing these changes seems to be the "N3" element in the formula (see Annexe J), indicating that it is the presence of an extreme level of any dimension which makes a state particularly intolerable if it persists for any length of time.

Converting VAS scores into TTO scores

Scaling methods which are utility–based, such as TTO and SG, tend to be resource–intensive in that they are often interviewer–based, and may require special aids to present descriptions of health states, and to enable respondents to record their valuations of them. Such methods, however, are favoured by many researchers who consider them to be well–grounded in theory, or who demand that expressions of preference must involve an element of choice. Simpler, less demanding methods have been utilised, and foremost amongst these has been category rating, of which the visual analogue scale is a graphical form. This technique has been adopted by the EuroQoL Group as the standard form by which valuations are obtained in postal surveys, and for that reason it was included as a candidate in the earlier MVH study which compared the performance of different scaling methods, and was included in the Main Survey.

Given that different scaling methods applied to the same health states, tend to yield different valuations, a question arises as to the form of any relationship between these values. Quite apart from this methodological interest, there is a strong practical reason for considering this question. If the results obtained using a `simple', technically accessible method could be systematically related to those obtained using a more `complex' method, then the former could be deployed when resources precluded the use of interviewer–based methods.

For the purposes of this study it was considered appropriate to investigate only the form of any relationship that linked the estimated values for any health state. Hence two general sets of data exist – estimated scores for 243 health states produced using the standard models applied to the individual and aggregate [median] TTO data, and an equivalent set of scores based on the VAS models. The general problem to be investigated amounted to seeking an arithmetic process by which TTO values (based on models of either individual responses or median values) for all health states, could be estimated, given knowledge of the corresponding VAS rating. The resulting equation is shown in Annexe K, together with the coefficients for an estimate based on medians and for an estimate based on means.

This conversion method could be used to convert our duration–specific VAS valuations into

duration–specific TTO values, though the chain ot reasoning to support such a process of recalculation is somewhat tenuous. It requires two key assumptions: firstly that the same relationship exists between TTO valuations for states of different duration as exists between VAS valuations, and secondly that the relationship that exists between VAS 10 year valuations and TTO 10 year valuations is similar to that between the two methods for each of the other durations. Unfortunately we have no data that sheds any light on the reasonableness of those assumptions, because we were unable to devise a feasible method for using the TTO method for very short durations. We have nevertheless taken that important step, in order to fill an important void until such time as someone can generate the extra data required to do the job more directly. Meanwhile we have estimated TTO tariffs for states of one year duration and one month duration.

UNFINISHED BUSINESS

There is plenty of work still to be done on the measurement and valuation of health, even in the particular domain of generic indexes of health–related quality–of–life, which is only one part of this burgeoning field. For instance, one difficult technical problem that needs investigation is the extent to which the value given to a health state is influenced by the health state that precedes it, or the health state that is expected to succeed it.

A very important policy issue concerns variations in values between different sub–groups in the population. We have explored these with respect to the sociodemographic characteristics of our study population, but a more important issue is whether doctors and nurses have significantly different valuations from the general public. If they do, it would raise serious doubts about the appropriateness of professionally defined measures of benefit from health care, and about the basis of clinical priority–setting. In conjunction with SCPR, the MVH Group has worked up a detailed protocol which involves the replication of our Main Survey on 1000 doctors and nurses chosen to be representative of those currently working in the NHS. This protocol has been alpha–rated on scientific grounds by the relevant MRC Board, but its funding is nevertheless in the balance, and due to be decided in July 1995.

One issue that was on our original research agenda, but which has been set aside to enable

the main flow of work to proceed, is whether the value attached to a health benefit depends on who is to receive it. The general assumption underlying all measurement of health–related quality–of–life is that it is the nature of the change in the individual's situation that is the focus of interest, not who the individual is. This position has a strong ethical justification, as well as being a convenient simplifying assumption for research purposes. It enables simple aggregation of results to proceed untrammelled, a feature of all such measures in practice, and indeed a feature of much cruder measures such as survival or mortality rates. But many people think that, in some circumstances at least, priority should be given to the young over the old, or the parents of young children over their childless contemporaries. And there are even more contentious issues often raised here, concerning those who have cared for their own health and those who have not (eg by smoking, heavy drinking, or drug abuse). We need to know a lot more about the attitudes of the general public towards these matters, so that policy–making can be better informed. Methodological work to improve on existing survey work remains on our research agenda.

But the main future activity of the MVH Group is going to be the implementation of the benefit measures we have already generated. This has already been going on at a low level during the development phase, since many people have been keen to experiment in the use of the earlier HMQ and the later EuroQol in a practical setting. To expand this realm of activity requires the production of user–friendly documentation and instruction manuals which are different from those designed for methodological work by the research community. It also requires a support facility to ensure that the instruments are used appropriately, and their full potential exploited. This is our next challenge.

MVH Group
Centre for Health Economics
University of York

June 1995

Membership:

Dalen, Harmanna van (1988 to 1991)
Dolan, Paul (1991 to 1994)
Durand, Mary-Alison (1988-1990)
Gudex, Claire (1987 to 1990 and 1991 to 1995)
Kind, Paul (1987 to date)
Lewis, David (1990 to 1991)
Morris, Jenny (1988 to 1991)
Williams, Alan (1987 to date)

SCPR Collaborators:

Erens, Bob
Thomas, Roger
Thomson, Katarina
and the key field workers and their managers

Official Reports:

Jan 1991	Report on Phase I of Research
Dec 1992	Valuing Health States: A Comparison of Methods. Final Report - Part I.
Dec 1992	Valuing Health States: A Comparison of Methods. Final Report - Part II.
Dec 1992	Valuing Health States: A Comparison of Methods. Technical Appendix
May 1993	Supplementary report.
Jun 1993	Second Developmental Pilot: Dress rehearsal
May 1994	First Report on the Main Survey
Oct 1994	Generating a UK EuroQol Tariff (Interim report on modelling)
Jan 1995	Final Report on the Modelling of Valuation Tariffs

SCPR Reports:

Aug 1992	Health Related Quality of Life: Technical Report (Thomas R and Thomson K)
Aug 1992	Health Related Quality of Life: Comments on the Pre-Pilots and Pilot (Thomas R and Thomson K)
Sep 1993	Health Related Quality of Life: the 1993 Pilots (Thomson K)
Apr 1994	Health Related Quality of Life: General Population Survey Technical Report (Erens, R)

Other Publications Arising from this work: (as at May 1995)

Gudex, C & Kind, P
The QALY toolkit. York: Centre for Health Economics, (Discussion Paper 38), 1988.

Kind, P
The design and construction of quality of life measures. York: Centre for Health Economics, (Discussion paper 43), 1988.

Kind, P
Issues in the design and construction of a quality of life measure. Baldwin, S, Godfrey, C & Propper, C (eds.), The Quality of Life: perspectives and policies. London: Routledge, 1990: 63–71.

Kind, P
Measuring valuations for health states: a survey of patients in general practice. York: Centre for Health Economics, (Discussion Paper 76), 1990.

Williams, A H, Durand, M–A, Gudex, C, Kind, P, Morris, J & van Dalen, H
Lay concepts of health with special reference to severely physically handicapped young adults and their carers. Report to the Nuffield Provincial Hospitals' Trust. York: Centre for Health Economics, 1990.

Kind, P & Gudex, C
The HMQ: measuring health status in the community. York: Centre for Health Economics, (Discussion paper 93), 1991.

Williams, A H & Kind, P
The present state of play about QALYs. Hopkins, A (ed.), Measures of the quality of life and the uses to which such measures may be put. London: Royal College of Physicians of London, 1992: 21–34.

Kind, P
Quality of life and the calculation of disability–free life years. Robine, J–M, Blanchet, M & Dowd, J E (eds.), Health expectancy. First workshop of the International Healthy Life Expectancy Network (REVES). London: HMSO, (OPCS Studies on Medical and Population Subjects, 54), 1992: 99–104.

Gudex, C
Are we lacking a dimension of energy in the Euroqol instrument?. Bjork, S (ed.), Euroqol conference proceedings, Lund, October 1991. Lund: Swedish Institute for Health Economics, (Discussion paper 1 Working paper 1992:2), 1992: 97.

Gudex, C, Drummond, M F & Williams, A H
Quality–Adjusted Life Years: a review. Committee on Core Health Services Report, Wellington, New Zealand. York: Centre for Health Economics, 1992.

Brazier, J, Jones, N & Kind, P
Testing the validity of the EuroQol and comparing it with the SF–36 health survey questionnaire. Quality of Life Research, 1993, 2: 169–180.

Kind, P & Gudex, C
The role of QALYs in assessing priorities between health care interventions. Drummond, M & Maynard, A (eds.), Purchasing and providing cost-effective health care. Edinburgh: Churchill Livingstone, 1993: 94-108.

Gudex, C, Kind, P & Dolan, P
The valuation of death. Sintonen, H. (ed.), EuroQol Conference Proceedings, Helsinki, Oct. 1992. Kuopio: Kuopio University Publications, (Discussion paper 2 Kuopio University Publications E.Social Sciences 8), 1993: 23-39.

Gudex, C, Kind, P, van Dalen, H, Durand, M-A, Morris, J & Williams, A
Comparing scaling methods for health state valuations: Rosser revisited. York: Centre for Health Economics, (Discussion paper 107), 1993.

van Dalen, H, Williams, A & Gudex, C
Lay people's evaluations of health: are there variations between different sub-groups?. Journal of Epidemiology and Community Health, 1994, 48(3): 248-253.

Kind, P & Gudex, C
Measuring health status in the community: a comparison of methods. Journal of Epidemiology and Community Health, 1994, 48(1): 86-91.

Kind, P, Dolan, P, Gudex, C & Williams, A H
Practical and methodological issues in the development of the EuroQol: the York experience. Advances in Medical Sociology, 1994, 5: 219-253.

Gudex, C (ed.)
Time Trade-off user manual: props and self-completion methods. York: Centre for Health Economics, 1994.

Gudex, C (ed.)
Standard gamble user manual: props and self-completion methods. York: Centre for Health Economics, 1994.

Williams A
The role of the EuroQol instrument in QALY calculations. York: Centre for Health Economics, 1995.

REFERENCES

Culyer, A.J., Lavers, R. and Williams, A. (1972), "Health Indicators", in Social Indicators and Social Policy, (editors) Shonfield, A. and Shaw, S., Heinemann, London.

Gater R, Kind P, & Gudex C (1995), "Quality of life in liaison psychiatry: a comparison of patient and clinician assessment" Brit J of Psychiatry 166:515–520

Gudex C (1986) "QALYs and their use by the health service" Discussion Paper No 20, Centre for Health Economics, University of York.

Gudex C (1995, forthcoming) "Quality of life in end-stage renal disease" Quality of Life Research

Gudex C, Williams A, Jourdan M, Mason R, Maynard J, O'Flynn R, & Rendall M (1990) "Prioritising Waiting Lists" Health Trends 22:103–108.

Hutton, J & Williams, A (1988) "Cost-effectiveness analysis of diagnostic technology: a decision framework applied to the case of MRI". in Duru, G, Engelbrecht, R, Flagle, C D & Van Eimeren, W (eds.), System science in health care. Vol. 2. Health care system and actors. Fourth International Society for System Science in Health Care Conference, Lyon, 4–8 July 1988. Paris: Masson, (Collection de medicine legale et de toxicologie medicale No. 139), 493–496.

Kind, P. and Gudex, C. (1994), "Measuring health status in the community: a comparison of methods", J. of Epid and Community Health 48:86–91.

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

Kind, P. and Sims, S. (1987), "CT Scanning in a District General Hospital", Research Report Centre for Health Economics, University of York

Patrick, D.L., Bush, J.W. and Chen, M.M. (1973), "Methods for Measuring Levels of Well-Being for a Health Status Index", Health Services Research, 8, 229–44.

Rosser, R. and Kind, P. (1978), "A Scale of Valuations of States of Illness: Is there a social consensus? Int. J. of Epidemiology, 7, 347–57.

Rosser, R. and Watts, V.C. (1972), "The Measurement of Hospital Output", Int. J. of Epidemiology, 1, 361–7.

Rosser, R. and Watts, V.C. (1978), "The Measurement of Illness", J. of Operational Research, 29, 529–40.

Rosser, R. (1983), "Issues of Measurement in the Design of Health Indicators: A Review", in Culyer, A.J. (editor), Health Indicators, Martin Robertson, Oxford (for the European Science Foundation).

Sackett, D.L. and Torrance, G.W. (1978), "The Utility of Different Health States as Perceived by the General Public", J. of Chronic Disease, 31, 697–704.

Stevens, S.S., (1971) "Issues in Psychophysical Measurement" Psychological Review 78(5) 426–450.

Torrance, G.W., (1976) "Social Preferences for Health States: An Empirical Evaluation of 3 Measurement Techniques", Socio–Economic Planning Science 10, 129–136.

Torrance, G.W., Sackett, D.L. and Thomas, W.H. (1973), "A Utility Maximisation Model for Evaluation of Health Care Programs", Health Service Research, 7, 118–33.

Williams, A. (1985a), "The Economics of Coronary Artery Bypass Grafting", BMJ, 291, 326–9.

Williams, A (1985b) "The nature, meaning and measurement of health and illness: an economic viewpoint". Social Science and Medicine, 20(10): 1023–1027.

Williams, A (1987) "The importance of quality of life in policy decisions" in Walker, S R & Rosser, R M (eds.), Quality of life: assessment and application. Proceedings of the Centre for Medicines Research Workshop held at the CIBA Foundation, London, March 3rd 1987. Lancaster: MTP Press, 279–290.

Williams, A (1988a) "Applications in management" Teeling Smith, G (ed.), in Measuring health: a practical approach. Chichester: Wiley, 225–243.

Williams, A (1988b) "Ethics and efficiency in the provision of health care" in Bell, J M & Mendus, S (eds.), Philosophy and Medical Welfare. Cambridge:Cambridge University Press, (Royal Institute of Philosophy Lecture Series 23, Supplement to `Philosophy'), 111–126.

Wright, S.J. (1986), Age, Sex and Health: A Summary of Findings from the York Health Evaluation Survey, University of York, CHE Discussion Paper, No. 15.

ANNEXE A

Rosser's Classification of Illness States

Disability		Distress	
1	No disability	A.	No Distress
2	Slight social disability	B.	Mild
3	Severe social disability and/or slight impairment of performance at work Able to do all housework except very heavy tasks	C.	Moderate
		D.	Severe
4	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		
5	Unable to undertake any paid employment Unable to continue any education 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		
6	Confined to chair or able to move around in the house only with support from an assistant		
7	Confined to bed		
8	Unconscious		

ANNEXE B Rosser Revised Matrix (original values in parentheses)

	A	B	C	D
I	[1.00]	.89 (.995)	.89 (.990)	.67 (.967)
II	.89 (.990)	.81 (.986)	.78 (.973)	.56 (.932)
III	.70 (.980)	.63 (.972)	.57 (.956)	.44 (.912)
IV	.63 (.964)	.56 (.956)	.51 (.942)	.40 (.870)
V	.44 (.946)	.43 (.935)	.44 (.900)	.22 (.700)
VI	.44 (.875)	.44 (.845)	.34 (.680)	.22 (.000)
VII	.38 (.677)	.40 (.564)	.33 (.000)	.20 (−1.486)
VIII	.01 (−1.028)			

[Dead = 0]

ANNEXE C **Coverage of Various Descriptive Instruments**

Item	Code No.	All Unprompted Responses (Gen Pop) N	%	Rosser [in 2 Items] Covers	Euroqol [in 5 Items] Covers	NHP [in 45 Items] Covers	Qlty of Wellbng [in 43 Items] Covers	SIP [in 136 items] Covers
Usual Activities	01	118	8.32	*	*	*	*	*
Gen Well/Ill	02	65	4.58				*	
Dying	23	0	0				*	
Gen Not W/Ill	24	16	1.12				*	
Pain	25	96	6.77		*	*	*	
Named diseases	26	45	3.17				*	
Appetite	27	43	3.03					*
Resp. symptoms	28	37	2.61				*	
Other symptoms	29	41	2.89				*	
[Illn/Symptoms]	SUM	343	24.2					
Feelings	03	157	11.0	*	*	*	*	*
Finance	04	6	0.42			*		
Gen Hlth Behvr	05	8	0.56					
Diet	51	22	1.55					*
Smoking	52	15	1.05					
Exercise	53	22	1.55			*		*
Alcohol	54	11	0.77					
[Behaviours]	SUM	78	5.50					
Energy	06	150	10.5			*		*
Appearance	07	115	8.11					
Stress/coping	08	44	3.10					
Lucky	09	6	0.42					
Med/Doctors	10	34	2.39			*	*	
Mobility	11	107	7.55	*	*	*	*	*
Strngth/Rstnce	12	54	3.81					
Fitness	13	81	5.71			*		
Sens. Impairment	14	12	0.84				*	*
Sleep	15	15	1.05			*	*	*
Weight	16	44	3.10				*	
Enj usual acts	17	3	0.21			*		*
Dependence	18	33	2.32		*	*	*	*
Cogn Impairment	19	13	0.91				*	*
Lonely/helpful	20	4	0.28			*		*
	SUM	1417	100	26.9	35.9	58.07	58.62	49.11

ANNEXE D A BRIEF DESCRIPTION OF THE VALUATION METHODS

Magnitude Estimation

It was a variant of this technique that was used by Rosser in her original work as a means for obtaining direct valuations of health states. Subjects are asked to judge each health state (H_i) in terms of its perceived severity compared with the reference state of "no disability, no distress", which was assigned a value of 1. Subjects were told that each health state would last for a period of 20 years after which time they would die. They were then asked whether each state was better or worse than the reference state (1A), and then how many times better or worse. The utility values for each state are calculated using the following formula:

$$\frac{\text{(The value attached to the state } H_i \text{ minus the value for "dead")}}{\text{(1 minus the value for "dead")}}$$

A modified form of magnitude estimation was used in the scaling of the Rosser Index

Time Trade-Off

When the Time Trade-Off method is applied to states considered by the subject to be better than death, subjects are asked to make a decision between two alternatives: either to remain in health state (H_i) for a period of time (t=20years) followed by death, or to be healthy for a shorter period of time (x) followed by death. The duration of x is varied until the subject is indifferent between the two alternatives and the utility value of the individual's preference for health state (H_i) is given by the ration x : 20. For the purposes of this study we simply observed the numbers of states considered by each subject to be worse than dead under the TTO method, but did not generate any actual ratings for such states. Time trade-off has been used to examine valuations for health states in the general population in Canada, and was being used in the UK by the Brunel Group.

Category Rating: Three Variants

This is an indirect method of obtaining health state preferences in which ordinal data are generated. Three variants were used in the present study: (i) a graphical visual analogue scale in the form of a thermometer where 100 represented "best imaginable health state" and 0 represented "worse imaginable health state" (CRT); (ii) a column of numbered boxes in which the box numbered 1 represents the "worst imaginable health state" and the box numbered 9 represented "best imaginable health state" (CRN); and (iii) a version which used nine labelled boxes with the following descriptions opposite each respective box: best imaginable health state; very good; good; fairly good; neither good nor bad; fairly bad; bad; very bad; worst imaginable health state (CRL).

Respondents were asked to rate each health state according to the categories shown. In order to make comparisons with the direct methods the values generated using the thermometer version (CRT) were transformed to a 0-1 scale where 0 represented death and 1 represented "no disability, no distress" using the following formula:

$$\frac{\text{(value for state } H_i \text{ - value for dead)}}{\text{(value for state 1A - value for dead)}}$$

In the later stages of our survey work, when using the CRT (or "VAS") approach we adopted the "bisection" procedure. In this, after first rating the best and worst states on the "thermometer", respondents are then asked to select the state which came closest to being half-way on the scale between where they had rated the best, and where they had rated the worst state. After rating this state wherever they thought it should go, the process is repeated for the state which falls roughly halfway between the middle state and the best state, and then for the state which falls roughly halfway between the middle state and the worst state.

Pairwise Comparison

With this method individuals are asked to make judgements about pairs of health states by indicating which of the two states is worse, or whether the two states are considered equal in severity. The method enables measures of internal consistency to be calculated, and it can be used to assess the quality of each respondent's performance as well as the extent of agreement between individuals. Paired comparisons methods have been used in deriving valuations for the Nottingham Health Profile

Standard Gamble

When the Standard Gamble method is applied to states considered by the subject to be better than being dead, individuals are asked to compare the certainty of remaining in that state with a lottery in which they risk immediate death in order to be restored to full health. The risk of immediate death is varied until the subject is indifferent between the lottery and remaining in the state in question. The greater the risk of death in this situation, the worse the state must be. For states considered worse than death by the subject, the comparison is between the certainty of immediate death and a lottery in which there is a chance of being restored to full health rather than remaining in the state in question. The chance of being restored to full health is varied until the subject is indifferent between the lottery and immediate death. The greater the chance of being restored to full health that is needed to get the subject to this situation, the worse the state in question must be.

ANNEXE E THE EUROQOL CLASSIFICATION SYSTEM

Mobility

1. No problems walking about
2. Some problems walking about
3. Confined to bed

Self-Care

1. No problems with self-care
2. Some problems washing or dressing self
3. Unable to wash or dress self

Usual Activities

1. No problems with performing usual activities (e.g. work, study, housework, family or leisure activities)
2. Some problems with performing usual activities
3. Unable to perform usual activities

Pain/Discomfort

1. No pain or discomfort
2. Moderate pain or discomfort
3. Extreme pain or discomfort

Anxiety/Depression

1. Not anxious or depressed
2. Moderately anxious or depressed
3. Extremely anxious or depressed

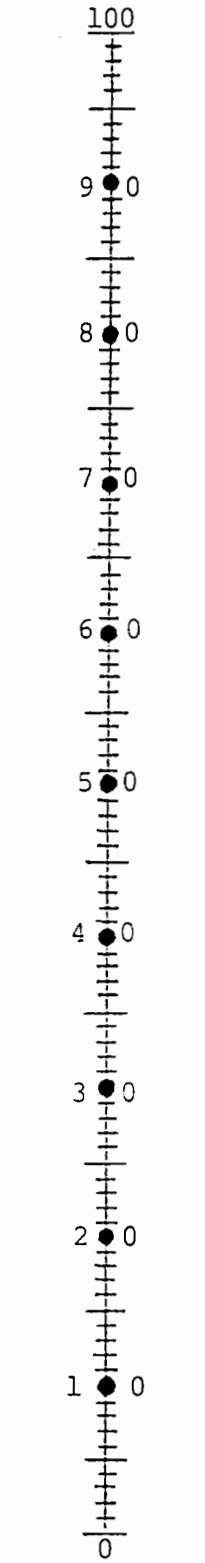
Note: For convenience each composite health state has a five digit code number relating to the relevant level of each dimension, with the dimensions always listed in the order given above. Thus 11223 means:

1 No problems walking about
1 No problems with self-care
2 Some problems with performing usual activities
2 Moderate pain or discomfort
3 Extremely anxious or depressed

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked by 100 and the worst state you can imagine is marked by 0.

We would like you to indicate on this scale how good or bad is your own health today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your current health state is.

Your own health state today



ANNEXE G EUROQOL STATES VALUED IN THE MAIN SURVEY

A. Each respondent valued all four of these key states:

11111
33333
unconscious
immediate death

B. Each respondent also valued 2 of the following (very mild) states (selected at random):

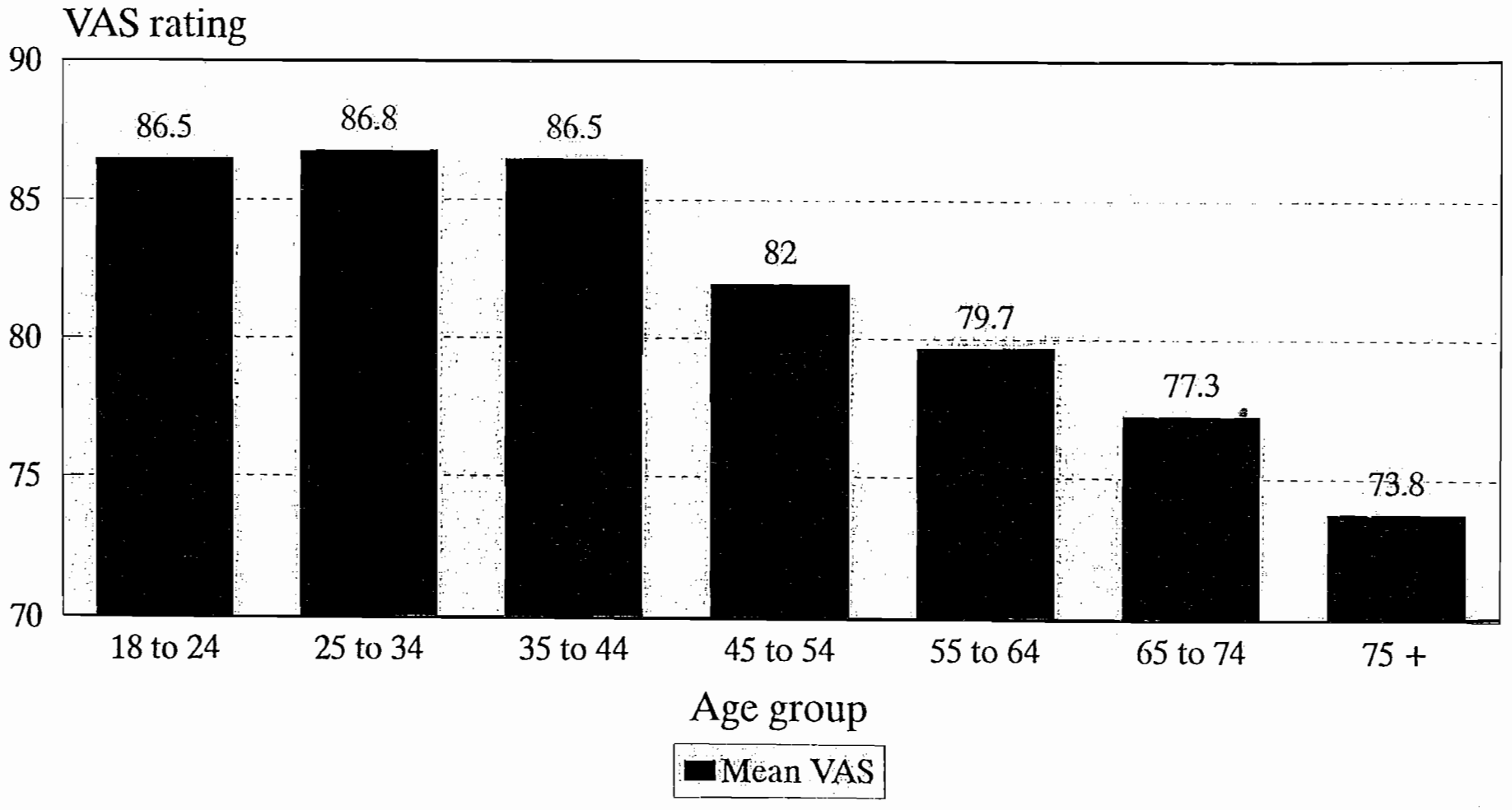
11112
11121
11211
12111
21111

C. Each respondent valued 3 randomly selected states from Set 1, which are the mildest ones. Each respondent also valued 3 randomly selected states from Set 2 (the moderately valued ones), and 3 randomly selected states from Set 3 (the most severe ones).

SET 1	SET 2	SET 3
12211	13212	33232
11133	32331	23232
22121	13311	23321
12121	22122	13332
22112	12222	22233
11122	21323	22323
11312	32211	32223
21312	12223	32232
21222	22331	33321
21133	21232	33323
11113	32313	23313
11131	22222	33212

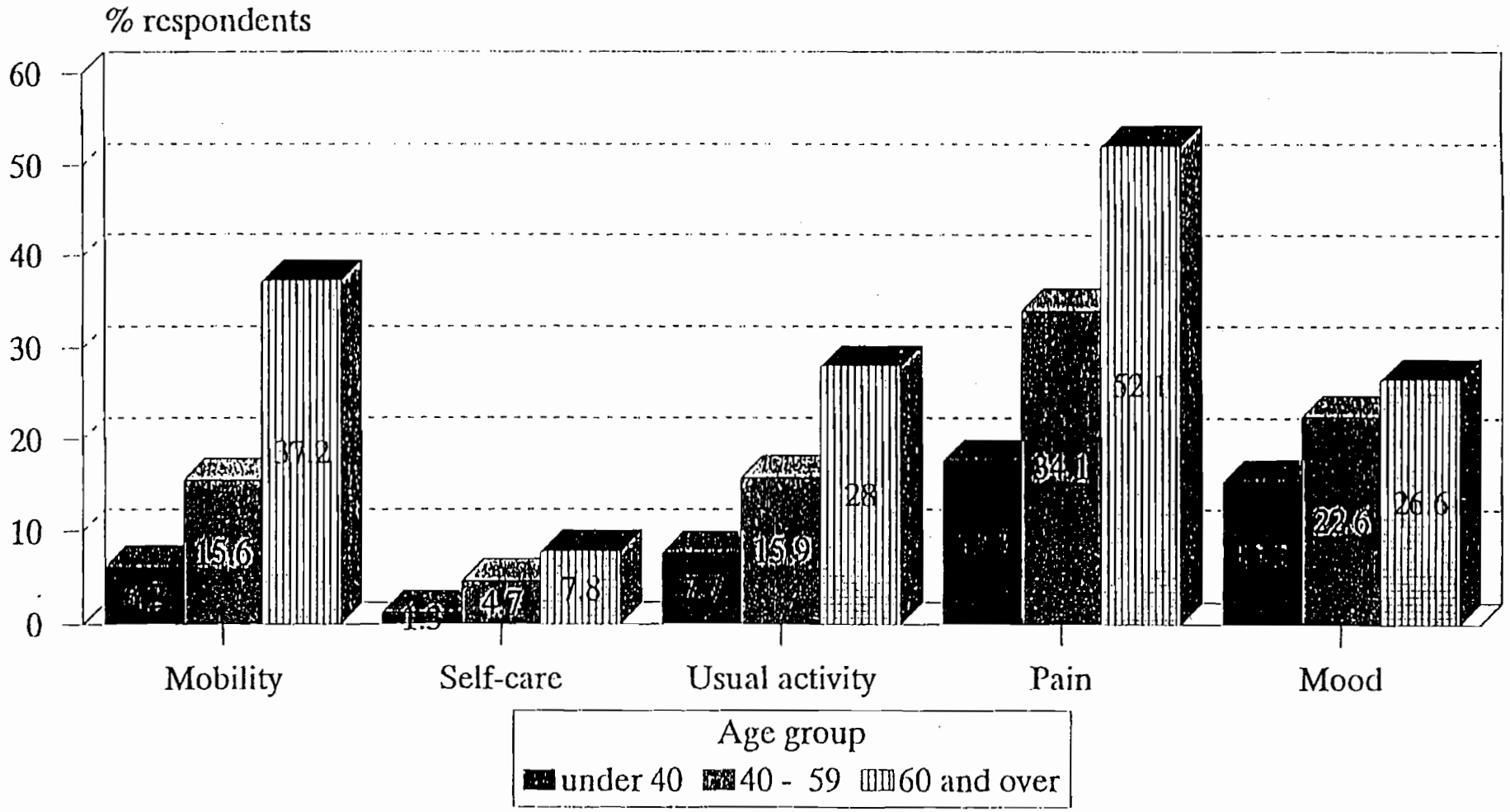
ANNEXE H

Self-rated health status
by age group



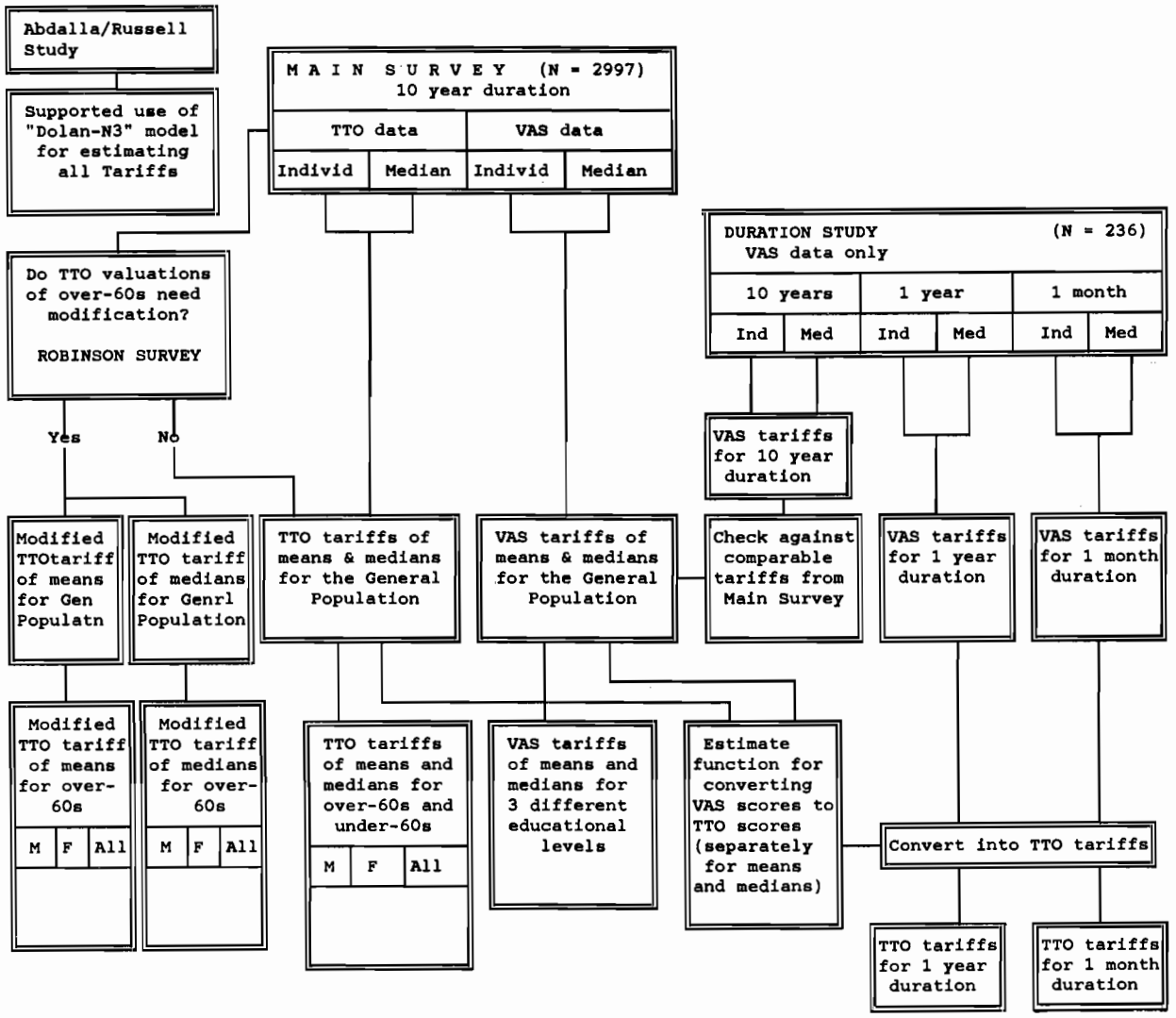
Frequency of reported problem

Distribution by age group



n = 3395

ANNEXE I STRUCTURE OF THE ANALYTICAL WORK



ANNEXE J THE BASIC TARIFF

A TARIFF OF VALUES FROM THE GENERAL POPULATION
(for health states of 10 years duration)
Means based on time-trade-off valuations

Dimension	Coefficient
Constant	0.081
Mobility	
level 2	0.069
level 3	0.314
Self-care	
level 2	0.104
level 3	0.214
Usual activity	
level 2	0.036
level 3	0.094
Pain/discomfort	
level 2	0.123
level 3	0.386
Anxiety/depression	
level 2	0.071
level 3	0.236
N3	0.269

Note: Unconscious = -0.402

The arithmetic needed to calculate the value for any health state from this table of decrements is given by the following example:

Taking health state 1 1 2 2 3

Full health = 1.0		
Constant term (for any dysfunctional state)		(subtract 0.081)
Mobility	.. level 1	(subtract 0)
Self-care	.. level 1	(subtract 0)
Usual activity	.. level 2	(subtract 0.036)
Pain / discomfort	.. level 2	(subtract 0.123)
Anxiety / depression	.. level 3	(subtract 0.236)
Level 3 occurs within at least 1 dimension		(subtract 0.269)

Hence **the estimated value for state 1 1 2 2 3 =**

1.0 - 0.081 - 0.036 - 0.123 - 0.236 - 0.269 = **.255**

Note: Some severe states will have negative values if they last 10 years, indicating that the general public regards such a prospect as worse than being dead.

TARIFF A1: TTO TARIFF OF MEANS: WHOLE POPULATION - 10 year duration

	Level 2	Level 3
Mobility	0.069	0.314
Self-care	0.104	0.214
Usual activity	0.036	0.094
Pain/discomfort	0.123	0.386
Anxiety/depression	0.071	0.236
Constant = 0.081		N3 = 0.269

1 1 1 1 1	1.000	1 2 1 3 2	0.089	1 3 2 2 3	0.041
1 1 1 1 2	0.848	1 2 1 3 3	-0.076	1 3 2 3 1	0.014
1 1 1 1 3	0.414	1 2 2 1 1	0.779	1 3 2 3 2	-0.057
1 1 1 2 1	0.796	1 2 2 1 2	0.708	1 3 2 3 3	-0.222
1 1 1 2 2	0.725	1 2 2 1 3	0.274	1 3 3 1 1	0.342
1 1 1 2 3	0.291	1 2 2 2 1	0.656	1 3 3 1 2	0.271
1 1 1 3 1	0.264	1 2 2 2 2	0.585	1 3 3 1 3	0.106
1 1 1 3 2	0.193	1 2 2 2 3	0.151	1 3 3 2 1	0.219
1 1 1 3 3	0.028	1 2 2 3 1	0.124	1 3 3 2 2	0.148
1 1 2 1 1	0.883	1 2 2 3 2	0.053	1 3 3 2 3	-0.017
1 1 2 1 2	0.812	1 2 2 3 3	-0.112	1 3 3 3 1	-0.044
1 1 2 1 3	0.378	1 2 3 1 1	0.452	1 3 3 3 2	-0.115
1 1 2 2 1	0.760	1 2 3 1 2	0.381	1 3 3 3 3	-0.280
1 1 2 2 2	0.689	1 2 3 1 3	0.216	2 1 1 1 1	0.850
1 1 2 2 3	0.255	1 2 3 2 1	0.329	2 1 1 1 2	0.779
1 1 2 3 1	0.228	1 2 3 2 2	0.258	2 1 1 1 3	0.345
1 1 2 3 2	0.157	1 2 3 2 3	0.093	2 1 1 2 1	0.727
1 1 2 3 3	-0.008	1 2 3 3 1	0.066	2 1 1 2 2	0.656
1 1 3 1 1	0.556	1 2 3 3 2	-0.005	2 1 1 2 3	0.222
1 1 3 1 2	0.485	1 2 3 3 3	-0.170	2 1 1 3 1	0.195
1 1 3 1 3	0.320	1 3 1 1 1	0.436	2 1 1 3 2	0.124
1 1 3 2 1	0.433	1 3 1 1 2	0.365	2 1 1 3 3	-0.041
1 1 3 2 2	0.362	1 3 1 1 3	0.200	2 1 2 1 1	0.814
1 1 3 2 3	0.197	1 3 1 2 1	0.313	2 1 2 1 2	0.743
1 1 3 3 1	0.170	1 3 1 2 2	0.242	2 1 2 1 3	0.309
1 1 3 3 2	0.099	1 3 1 2 3	0.077	2 1 2 2 1	0.691
1 1 3 3 3	-0.066	1 3 1 3 1	0.050	2 1 2 2 2	0.620
1 2 1 1 1	0.815	1 3 1 3 2	-0.021	2 1 2 2 3	0.186
1 2 1 1 2	0.744	1 3 1 3 3	-0.186	2 1 2 3 1	0.159
1 2 1 1 3	0.310	1 3 2 1 1	0.400	2 1 2 3 2	0.088
1 2 1 2 1	0.692	1 3 2 1 2	0.329	2 1 2 3 3	-0.077
1 2 1 2 2	0.621	1 3 2 1 3	0.164	2 1 3 1 1	0.487
1 2 1 2 3	0.187	1 3 2 2 1	0.277	2 1 3 1 2	0.416
1 2 1 3 1	0.160	1 3 2 2 2	0.206	2 1 3 1 3	0.251

2	1	3	2	1	0.364	2	3	2	3	2	-0.126	3	2	2	1	3	-0.040
2	1	3	2	2	0.293	2	3	2	3	3	-0.291	3	2	2	2	1	0.073
2	1	3	2	3	0.128	2	3	3	1	1	0.273	3	2	2	2	2	0.002
2	1	3	3	1	0.101	2	3	3	1	2	0.202	3	2	2	2	3	-0.163
2	1	3	3	2	0.030	2	3	3	1	3	0.037	3	2	2	3	1	-0.190
2	1	3	3	3	-0.135	2	3	3	2	1	0.150	3	2	2	3	2	-0.261
2	2	1	1	1	0.746	2	3	3	2	2	0.079	3	2	2	3	3	-0.426
2	2	1	1	2	0.675	2	3	3	2	3	-0.086	3	2	3	1	1	0.138
2	2	1	1	3	0.241	2	3	3	3	1	-0.113	3	2	3	1	2	0.067
2	2	1	2	1	0.623	2	3	3	3	2	-0.184	3	2	3	1	3	-0.098
2	2	1	2	2	0.552	2	3	3	3	3	-0.349	3	2	3	2	1	0.015
2	2	1	2	3	0.118	3	1	1	1	1	0.336	3	2	3	2	2	-0.056
2	2	1	3	1	0.091	3	1	1	1	2	0.265	3	2	3	2	3	-0.221
2	2	1	3	2	0.020	3	1	1	1	3	0.100	3	2	3	3	1	-0.248
2	2	1	3	3	-0.145	3	1	1	2	1	0.213	3	2	3	3	2	-0.319
2	2	2	1	1	0.710	3	1	1	2	2	0.142	3	2	3	3	3	-0.484
2	2	2	1	2	0.639	3	1	1	2	3	-0.023	3	3	1	1	1	0.122
2	2	2	1	3	0.205	3	1	1	3	1	-0.050	3	3	1	1	2	0.051
2	2	2	2	1	0.587	3	1	1	3	2	-0.121	3	3	1	1	3	-0.114
2	2	2	2	2	0.516	3	1	1	3	3	-0.286	3	3	1	2	1	-0.001
2	2	2	2	3	0.082	3	1	2	1	1	0.300	3	3	1	2	2	-0.072
2	2	2	3	1	0.055	3	1	2	1	2	0.229	3	3	1	2	3	-0.237
2	2	2	3	2	-0.016	3	1	2	1	3	0.064	3	3	1	3	1	-0.264
2	2	2	3	3	-0.181	3	1	2	2	1	0.177	3	3	1	3	2	-0.335
2	2	3	1	1	0.383	3	1	2	2	2	0.106	3	3	1	3	3	-0.500
2	2	3	1	2	0.312	3	1	2	2	3	-0.059	3	3	2	1	1	0.086
2	2	3	1	3	0.147	3	1	2	3	1	-0.086	3	3	2	1	2	0.015
2	2	3	2	1	0.260	3	1	2	3	2	-0.157	3	3	2	1	3	-0.150
2	2	3	2	2	0.189	3	1	2	3	3	-0.322	3	3	2	2	1	-0.037
2	2	3	2	3	0.024	3	1	3	1	1	0.242	3	3	2	2	2	-0.108
2	2	3	3	1	-0.003	3	1	3	1	2	0.171	3	3	2	2	3	-0.273
2	2	3	3	2	-0.074	3	1	3	1	3	0.006	3	3	2	3	1	-0.300
2	2	3	3	3	-0.239	3	1	3	2	1	0.119	3	3	2	3	2	-0.371
2	3	1	1	1	0.367	3	1	3	2	2	0.048	3	3	2	3	3	-0.536
2	3	1	1	2	0.296	3	1	3	2	3	-0.117	3	3	3	1	1	0.028
2	3	1	1	3	0.131	3	1	3	3	1	-0.144	3	3	3	1	2	-0.043
2	3	1	2	1	0.244	3	1	3	3	2	-0.215	3	3	3	1	3	-0.208
2	3	1	2	2	0.173	3	1	3	3	3	-0.380	3	3	3	2	1	-0.095
2	3	1	2	3	0.008	3	2	1	1	1	0.232	3	3	3	2	2	-0.166
2	3	1	3	1	-0.019	3	2	1	1	2	0.161	3	3	3	2	3	-0.331
2	3	1	3	2	-0.090	3	2	1	1	3	-0.004	3	3	3	3	1	-0.358
2	3	1	3	3	-0.255	3	2	1	2	1	0.109	3	3	3	3	2	-0.429
2	3	2	1	1	0.331	3	2	1	2	2	0.038	3	3	3	3	3	-0.594
2	3	2	1	2	0.260	3	2	1	2	3	-0.127	Unconscious [-0.402]					
2	3	2	1	3	0.095	3	2	1	3	1	-0.154						
2	3	2	2	1	0.208	3	2	1	3	2	-0.225						
2	3	2	2	2	0.137	3	2	1	3	3	-0.390						
2	3	2	2	3	-0.028	3	2	2	1	1	0.196						
2	3	2	3	1	-0.055	3	2	2	1	2	0.125						

ANNEXE K CONVERTING VAS SCORES INTO TTO SCORES

Various functional forms were used to examine the nature of the relationship between the VAS and TTO valuations. A general equation of the following form resulted from this study

$$TTO_i = a_0 + a_1 . VAS_i + a_2 . VAS_i^2$$

where
VAS_i is the `observed' VAS score for health state i
TTO_i is the predicted TTO value for health state i
a_0 a_1 a_2 are coefficients with different values assigned when individual-level or median VAS data are modelled

The values of coefficients for the two forms of equation is as follows

	Estimated values based on means	Estimated values based on medians
a_0	-0.445	-0.704
a_1	2.112	3.313
a_2	-0.580	-1.604
r^2	0.99	0.98