ISSUES IN THE VALUATION OF HEALTH OUTCOMES

Paul Dolan

Thesis submitted for D.Phil

University of York

Department of Economics

January 1997

ABSTRACT

The measurement of preferences concerning health outcomes has the potential to provide important information on the benefits associated with alternative allocations of the health care budget. However, there are a number of important and controversial issues which must be addressed when measuring such preferences and this thesis addresses some of these.

It begins by briefly outlining some of the problems associated with measuring benefits in monetary units and then concentrates on issues relating to the measurement of health-related quality-of-life (HRQoL). It then focuses upon the following issues: the relationships between, and the appropriateness of, different valuation methods; the methodologies that can be adopted to estimate a full set of health state valuations from a subset of direct valuations; the factors that may influence or bias stated responses; and the appropriateness of using individual valuations of HRQoL as measures of social value.

In addressing each of these issues, valuation data gathered principally from members of the general public but also from convenience samples of students are used to test different hypotheses. The data analysed here form one of the most comprehensive datasets ever gathered on respondent preferences concerning HRQoL and thus enable much additional light to be shed on the issues outlined above. Of course, such data itself generates additional questions and the thesis also looks at the ways in which future research might be directed towards addressing them.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX

ACKNOWLEDGEMENTS

The main objective of the Measurement and Valuation of Health (MVH) Group at the Centre for Health Economics, University of York since the Group's inception in 1987 has been the measurement of preferences over HRQoL. I was a member of the MVH Group from 1991 to 1994 and Chapters 2 and 3 of this thesis are based upon research I was involved in during this period whilst Chapters 4 and 5 develop issues that were raised out of my time with the Group. Therefore, a considerable debt is owed to the other members of the Group during this period - Alan Williams, Paul Kind and Claire Gudex.

In addition, I would like to thank my co-authors on the papers which appear in this thesis: Michael Jones-Lee, Matthew Sutton, Colin Green, Nigel Rice and Angela Robinson. Their contribution, together with the comments of referees, enhance its quality immeasurebly. Special thanks are also due to my thesis advisory group - Graham Loomes, Mike Drummond and Alan Williams - who took the time and effort to read earlier drafts of this thesis and to point me in the right directions.

Over and above his 'official' contribution, I would like to acknowledge the guidance and support of Alan Williams over the last five years. Finally, I would like to thank my parents, Josie and John Dolan, and my partner, Maisie Rowe, who have given me the support, encouragement and love that made this thesis possible.

AUTHOR'S DECLARATION

This thesis is based upon a number of articles that have appeared or are about to appear in peer-reviewed journals:

Chapter 2.1 is based upon Dolan P and Sutton M, Mapping visual analogue scale scores onto time trade-off and standard gamble ones, Social Science and Medicine, 44, 1519-1530, 1997.

Chapter 2.2 is based upon Dolan P, Gudex C, Kind P and Williams A, Valuing health states: a comparison of health states, Journal of Health Economics, 15, 209-231, 1996.

Chapter 3.1 is based upon Dolan P, Gudex C, Kind P and Williams A, The time trade-off method: results from a general population study, Health Economics, 5, 141-154, 1996.

Chapter 3.2 is based upon Dolan P, Modelling valuations for EuroQol health states, Medical Care, forthcoming.

Chapter 3.4 is based upon Dolan P, Aggregating health state valuations, Journal of Health Service Research and Policy, forthcoming.

Chapter 4.1 is based upon Dolan P and Jones-Lee M, The effect of lifetime reallocation of consumption and discounting on TTO responses, Journal of Health Economics, forthcoming.

Chapter 4.2 is based upon Dolan P and Gudex C, Time preference, duration and health state valuations, Health Economics, 4, 289-299, 1995.

Chapter 4.3 is based upon Dolan P, Modelling valuations for health states: the effect of duration, Health Policy, 38, 189-203, 1996.

Chapter 5.1 is based upon Dolan P, The measurement of individual utility and social welfare, Journal of Health Economics, forthcoming.

And upon articles that are currently under consideration by journals:

Chapter 4.3 is based upon Dolan P and Rice N, The effect of duration on health state valuations, resubmitted to Medical Decision Making.

Chapter 5.2 is based upon Dolan P and Robinson A, The allocation of benefits in health care: does equity matter?, submitted to Medical Decision Making.

Chapter 5.3 is based upon Dolan P and Green C, Using the person trade-off approach to measure differences between individual and social value, submitted to Health Economics.

CHAPTER 1: AN INTRODUCTION TO THE ISSUES

CHAPTER 1.1: VALUING THE BENEFITS OF HEALTH CARE

Introduction

Under conditions of perfect competition, price exactly equals marginal cost. In other words, the value consumers place on the last unit of production is exactly equal to its opportunity cost of production; a situation of allocative efficiency. In addition, in can be shown that in the long-run a firm in a perfectly competitive market is forced to the point of maximum technical efficiency. Neither is true of any other market structure. It is these propositions which underlie the "predilection of the market" on the part of many economists from Adam Smith onwards.

The reliance on market forces to allocate goods implies that the prevailing income distribution (or, more strictly, the income distribution after any transfers of wealth and income) is considered acceptable. It also rests on the acceptance of additional value judgements: for example, that individuals are the best judges of their own well-being, and that social welfare depends only on the welfare of persons in society. Although these latter value judgements are of very general appeal, there are conditions (for example, concerning the consumption of hard drugs) under which many people might think it justifiable to override individual preferences.

Whilst it is generally accepted that no market works perfectly, it may still be that leaving the resource allocation process to be determined by market forces remains the best way of getting as close as possible to the ideal outcomes (given the value judgements cited above) of the perfect market. However, the basic rationale underlying government intervention in health care is that none of the assumptions of perfect markets (fully informed consumers, a large number of suppliers etc.)

work in the case of health care. In other words, there are important characteristics of the commodity health care which render it more susceptible than other commodities to government intervention (see McGuire *et al* [1988] pp182-93).

If market forces are no longer relied upon to promote an efficient allocation of resources, it is necessary to find alternative ways in which to value the costs and benefits of health care, so that governments will be better able to deploy the limited resources at their disposal where they will be of greatest benefit. By leaving resource allocation to the market, this benefit is defined in terms of consumer preferences which are expressed by their willingness to pay, matched by their ability to pay. But when there is government intervention in the market, there are questions relating to how benefit is defined and how it is to be allocated amongst the population.

This thesis will concentrate on how the measurement of preferences concerning health-related quality-of-life (HRQoL) can be used to provide information on the benefits associated with alternative allocations of the health care budget. It begins, though, by considering an alternative methodology; namely, cost-benefit analysis, which is well-grounded in economic theory and which attempts to capture all the benefits associated with health care, but which suffers from a number of methodological problems. After briefly outlining the limitations of cost-effectiveness analysis, the thesis will then consider cost-utility analysis which has been widely used by health economists in the evaluation of health care programmes. This section will look at the problems associated with using more narrow measures of benefit like HRQoL before discussing the process of developing a measure of HRQoL that will be of use to economists wishing to compare a wide range of diverse health care interventions and policies.

Cost-benefit analysis

In general terms, a cost-benefit analysis (CBA) of a programme requires the identification of all the costs and benefits derived by all members of the community affected by the programme. It then requires these costs and benefits to be measured in some common unit so that aggregate benefits can be compared with aggregate costs. In this way, CBA provides an estimate of the value of resources used up by each programme compared to the value of resources the programme might save or benefits it might produce. Since costs are typically measured in monetary units, it is usual to express benefits likewise.

There are essentially three methods that can be adopted in order to attach monetary values to commodities that are not traded on the market. First, the human capital approach measures the value of preventing someone's death, injury or illness as the gain in the value of their present earnings. Second, the revealed preference approach attempts to observe decisions individuals make concerning risks to their health and then to infer their willingness to trade money for changes in these risks. Third, the contingent valuation method (CVM) which uses answers to survey questions to estimate an individual's willingness-to-pay (WTP) for a particular good, risk reduction or health improvement. For a more detailed discussion of these approaches see Jones-Lee [1976] and Sugden and Williams [1978]).

Because of problems associated with the first two approaches (for example, the human capital method does not take individual preferences into account and the revealed preferences method relies on (imperfect) market data), it is now generally accepted that in principle the CVM offers the most direct and effective means of establishing preference-based monetary values. The method was originally developed to estimate the value of environmental changes, such as the preservation of a recreational area (for a review, see Cummings *et al* [1986]). The CVM was probably first used in health care by Acton [1973] but most of the

studies is this area have taken place in the last 10 years (see the review by Donaldson [1993]).

It is essential in any CVM study that the scenario being valued is both plausible and meaningful to respondents. Needless to say, these requirements are not unproblematic. In their typology of potential response effect biases in CV studies, Mitchell and Carson [1989] identify three main types of bias relating to the design of the CV instrument: 1) incentives to misrepresent responses: this includes strategic bias, whereby a respondent (perhaps through acting out of self-interest) has an incentive to under- or over-state their true WTP; 2) implied value cues: for example, starting point bias in bidding games, whereby a respondent's stated WTP may be influenced by the first bid they receive; and 3) scenario misspecification: this occurs when the respondent does not respond to the correct scenario, usually because the question is formulated incorrectly.

Another potentially important bias arises from the fact that the CVM provides hypothetical answers to hypothetical questions. Thus, we cannot be certain that the respondent would behave in the same way in a real situation as they do in an experimental one. Mitchell and Carson [1989] argue that there is no evidence that the results from CV studies are biased in any systematic way but some studies have shown that the WTP based on the CVM is higher than the WTP based on actual decisions (see Cummings *et al* [1995]).

Against this background, the CVM has become the subject of heated controversy in the literature. For example, in the environmental context, possibly the most wide-ranging and fundamental debate, involving some fierce exchanges of views, followed the Exxon Valdez oil spillage (see Carson *et al* [1992], Arrow *et al* [1993] and Hausman [1993]). Much of this debate centred around the measurement of non-use values, due to the public good properties of environmental change. Since health is primarily (though not exclusively) a private good, the problems in using CVM in health care might be smaller. However, in

the context of safety, a recent study suggested that responses to CV questions designed to estimate the value of a statistical life are insufficiently sensitive to the size of the reduction in the risk of death faced by the respondent. As a result, there are now serious doubts as to the validity of monetary values of non-market goods established on the basis of responses to CV questions.

In addition, the appropriateness of eliciting (WTP) valuations for commodities like health care, for which government intervention indicates that willingness-to-pay should not be the main criterion for allocating resources, is questionable. Of course, this criticism stems from the fact that WTP is constrained by ability to pay and it is possible to construct hypothetical experiments which mitigate against the respondents' actual ability to pay influencing their responses. However, this raises additional questions about the validity of WTP responses that are not subject to an effective budget constraint. In any event, respondents, particularly those in societies with public provision of health care, may be hostile to questions that are couched in monetary terms and some policy-makers may be hesitant to base decisions on health benefits denominated in monetary units.

Alternative approaches

In view of all this, it is not surprising that many researchers have sought alternative methods of measuring the benefits of health care programmes. Since CBA is the only methodology that, at least in theory, provides information on the *absolute* benefit of programmes, the alternatives have concentrated only on the assessment of their *relative* performance. In this way, they can be seen as addressing a second-best optimisation problem i.e. assessing how best to allocate health care resources once the level of these resources have been determined. Although this is certainly a departure from the first-best world where we can say something about the appropriate level of resources devoted to health care relative to other demands on the public purse, Weinstein and Zeckhauser [1973] have shown that resource allocation problems involving a fixed budget may be solved by looking at the ratio

of benefits to costs for each project, and all projects with a ratio greater than an (endogenously determined) critical ratio are undertaken.

One of the alternative methods that can be used to evaluate health care programmes is cost-effectiveness analysis (CEA). This includes both the costs and consequences of the alternatives, using a single outcome measure expressed in natural units, for example, life years gained. It allows comparisons between treatments where the effectiveness is not equal. However, the results are meaningful only when the alternatives compared results in a change in the same outcome measure. It is not possible to rank therapies with different outcomes, for example, hip replacements where the primary outcome is mobility with surgery for cataracts, where the primary outcome is sight. In other words, CEA informs choices about the economic merits of interventions within but not between therapeutic categories. In addition, it permits inclusion of only one outcome measure, whereas many interventions have several potential benefits.

Cost-utility analysis

Unlike CEA, cost-utility analysis (CUA) combines multiple outcomes into a single measure and thus allows comparisons of the efficiency of interventions between different conditions. Here benefits are measured in terms of the utility associated with different levels of, or improvements in, health status. These utilities, or more accurately measures of HRQoL since preferences are elicited within the domain of health, are typically expressed on a scale where 'full health' is assigned an index value of 1 and 'death' an index value of 0 (see Torrance [1986]). To allow for the possibility that some states may be regarded as being worse than death, negative indices are often also allowed for. The indices attached to different states of health then provide the basis of a common denominator which allows the costs and benefits of different health care programmes to be compared. This common denominator is often expressed in terms of quality-adjusted life-years (or QALYs), which attempt to combine the value of quality of life with the value of length of

life into a single index number, which may then be used as a currency in which the benefits of health care interventions can be expressed.

It is worth noting that this thesis is not exclusively about QALYs - after the brief discussion here, the term does not reappear again until Chapter 5 - but about issues relating to the measurement of preferences concerning HRQoL more generally. However, since the QALY is one of the best known tools for aiding decision-making in the domain of health care, and are derived principally from the elicitation of preferences concerning HRQoL, it is worth considering their scope and limitations. Because one of the main advantages of CBA is that in principle it can account for all the benefits associated with a particular health care programme, attention will be focused on the extent to which QALYs can in principle do the same. Issues relating to the measurement of HRQoL (whether for use in QALY calculations or not) are discussed in more detail in the main body of text in this thesis.

As alluded to above, QALYs essentially measure health. The question that arises, then, is to what extent does improved health encompass the benefits of health care? Whilst it is reasonable to assume that the objectives of health care (whether decided upon by politicians, doctors, patients or the general public) will be dominated by considerations about health, is it also reasonable to assume that these considerations will strictly dominate all others all of the time? Mooney [1994] has suggested that these other benefits may relate to the extent to which the provision of health care is equitable, the value of information that doctors and others can provide, and the autonomy that patient's have in the decision-making process (pp16-20). Let us take each of these in turn.

There has been considerable debate in the health economics literature about what type of equity we want our health care system to promote; for example, do we want a system that promotes equal access for equal need (however need is defined), as argued by Mooney [1994] or one that promotes equal health, as

suggested by Culyer [1995]. In so far as we are concerned with inequalities in health, QALYs can in principle take such concerns into account and how this might be done in practice is discussed in Chapters 5.1 and 5.2. This author suspects that this is what we (be it the general public, patients, special interest groups or health care professionals) would be most concerned about but this is, of course, an empirical question; and one that he intends to address in future work.

In a number of screening programmes, it has been shown that patients value information *per se* (see Lange *et al* [1991] and Mooney and Lange [1993]), although the weight attached to information is likely to be considerably less in treatment programmes. Another argument likely to be in the patient's utility function is autonomy, which Mooney [1994] defines as the right to make a consumption decision, as well as the right not to make it. The question here is whether or not information and autonomy can be taken account of within the QALY framework.

It is the view of this author that in principle information (at least about one's current or future health status) can be taken account of within an appropriately defined 'mental health' dimension of a HRQoL instrument but that to incorporate autonomy is more problematic since QALYs are concerned with outcomes rather than the processes by which the outcomes come about. However, until empirical evidence shows that such factors as autonomy are sufficiently important for respondents to trade health off for them, it seems entirely justifiable that CUA proceeds in a piecemeal fashion assuming that things other than health are held constant.

The discussion in this chapter has highlighted the main issues associated with the use of CBA and CUA in the evaluation of health care programmes. In assessing the appropriateness of the two approaches, this author agrees with Johansson [1995] when he suggests that "there seems to be no strong reason for arguing emphatically in favour of one or the other approach to benefits measures or in

favour of cost-benefit analysis over cost-effectiveness analysis [which he takes to include CUA]; there is no measure which works in all possible circumstances; and cost-effectiveness analysis and cost-benefit analysis are complementary rather than mutually exclusive approaches" (p162-3). Against this background, issues related to the measurement of HRQoL are now discussed.

CHAPTER 1.2: THE MEASUREMENT OF HRQOL FOR USE IN CUA

Introduction

There are two main stages in the development of any measure of HRQoL. The first is to describe health status, preferably in such a way that different states of health can be identified. The second stage involves determining the numerical index value to be attached to the health states so described. Whilst the approach taken by the MVH Group will be discussed in this way, it must be remembered that the distinction between the two stages is not quite so clear-cut, since any decision made during the first stage to leave a particular aspect of health out of the descriptive system implies a value of zero in the second stage.

Description

There exist a number of different types of health state descriptive systems, each designed for a specific purpose. For example, condition-specific instruments are designed to measure HRQoL within a particular condition or disease group. They typically contain very detailed descriptions of a limited number (often only one) of the dimensions of health, since they are designed to be sensitive to small changes within the dimension(s) relevant to the particular condition. Attempts have been made to establish an 'exchange rate' between the different condition-specific measures so as to facilitate comparisons across disease groups but this has proved problematic (see Cairns *et al* [1991]).

Generic measures have been developed which measure health status across a range of different dimensions and as a result are typically less sensitive than condition-specific measures. Most generic measures consist of a health profile which allow a comparison of health within each dimension independently but do not combine the different dimensions to form an overall single index. For many health profiles, it would be impractical to try to do so since the combination of the various levels of the different dimensions would typically generate a universe of health states that is too large to elicit indices for; for example, the SF-36 (see Ware and Sherbourne [1992]) generates over ten million possible health states.

Those profile measures for which it is conceivably possible to generate an overall score often avoid the question of preferences by assuming that the levels within each dimension are equi-distant and even those that do have preference-based indices within each dimension often assume that each dimension contributes equally to the overall score (for example, see Spitzer *et al* [1981]). Therefore, neither condition-specific nor profile measures are suitable for use in informing resource allocation decisions across a range of diverse interventions.

In fact, at the time the MVH Group began work, there were remarkably few instruments that were suitable. The exceptions were the McMaster Health State Classification System (see Torrance [1982]), the Sickness Impact Profile (see Bergner *et al* [1981]) and the Quality of Well-Being Scale (see Patrick *et al* [1973]). Because these instruments are all of North American origin, the problems associated with cross-cultural comparisons of health (see Hunt *et al* [1991]) meant that a new generic instrument which is capable of being reduced to single index was required.

There are essentially two ways in which such an instrument can be developed (see Kind [1990]). One is a 'top-down' approach in which the researcher makes a judgement about the relevant aspects of health, based either on their personal views or upon an existing definition; for example, that of the World Health

Organisation (WHO) which expresses health in terms of overall physical, emotional and social well-being. The problem with this approach is that without reference to a wider set of judgements, there can be no guarantee that all relevant aspects of health have been included.

The other is the 'bottom-up' approach which, by asking a wider population to provide the relevant aspects, partially overcomes the problem of judging what should be included in the descriptive system. For example, the general public and/or patients can be surveyed (typically using a very 'open-ended' questionnaire format) in order to generate the descriptive material necessary to generate the health states. As with the 'top-down' approach, though, the researcher must make some judgement; this time about how this descriptive material should be organised so as to provide a viable set of descriptions.

The MVH Group in collaboration with researchers from Northern Europe (together calling themselves the EuroQol Group) decided to draw on both approaches. From published literature and their own experiences, the Group drew up a number of dimensions which they called the 'common core' since they were the items which most instruments contained, and then surveyed a number of different population groups in the UK to assess the extent which these were considered the most salient dimensions of health (see van Dalen *et al* [1994] for details).

After some revisions, and heated debate within the Group as to how best to balance considerations of coverage, importance of items and complexity against one another, the current EuroQol descriptive system evolved consisting of five dimensions, each of which comprises three levels of severity (see Figure 1.2.1), thus generating 3^5=243 possible health states. Of course, some of these states, for example, 33331, might be considered to be highly implausible but circumstances where they exist cannot be ruled out *ex ante*. For completeness, two further states were added, unconscious and dead, making 245 in all.

Valuation

For the EuroQol to be used in evaluating the health benefits associated with different health care interventions, it is important to derive a single index number for each of the 245 states. In determining these indices, there are three broad strategies: 1) use expert judgement, 2) use indices obtained from relevant literature, or c) use direct measurement of the preferences of an appropriate population (see Torrance [1986]). Because of the potential sources of bias associated with the first two (for example, judgements may be wrong, or published literature may be inappropriate) the third strategy is generally seen as the most appropriate and was the one adopted by the MVH Group.

In eliciting valuations for health states, two important issues must be considered: *whose* valuations should be sought, and *how* should these be derived? In answering the first question, we could elicit the preferences of doctors and other health care professionals, patients, or the general public. Since doctors might be thought of as having a broader and more objective view of the relative severity of different states of health, it might be appropriate to give greatest weight to their preferences. But, in a sense, greatest weight is already given to their views in that they hold a powerful position in the provision of health care.

It is sometimes argued that it is most appropriate to elicit valuations from those people who are currently experiencing the health states for which index values are sought. The argument seems to be that these are the only people who know what it is really like to be in these states but there are a number of points that need to be made here. First, the proposition itself may be flawed since, apart from the more severe health states, the distinction between those with current experience of illness and those without is really very blurred. Even in supposedly 'healthy' populations (for example, the general public), there is a substantial degree of 'ill' health and many currently 'healthy' people have experienced 'ill' health at some

time in their lives (even if they may sometimes have difficulty recalling it; see Christensen-Szalanski [1984]).

Second, it is possible that patient preferences may be susceptible to strategic bias, similar to that outlined in the discussion of the CVM. Of course, this may be true of any group of respondents but is potentially more likely amongst patients who may feel that their treatment will be directly affected by their responses. This suggests that when making comparisons of interventions that affect many different population sub-groups, it is likely that the views of the whole population will be most relevant. Third, even if we could 'trust' patient responses, there is the question of whether or not we would want to. Since it is well-established that there is a direct positive link between the time spent in ill health and adaptation to that ill health (see Meyerowitz [1983] and Cassileth *et al* [1984]), the question that arises is whether or not such adaptation should be taken into account when allocating resources which will deal with the treatment of prospective patients.

In general terms, the answer to this question will turn largely on whether social welfare is looked at *ex ante* or *ex post*. The *ex ante* approach means that social welfare is viewed as a function of the expected levels of utility attained by different individuals. The *ex post* approach means that utility is calculated conditional on everybody experiencing the same state of the world, and then to arrive at the overall level of social welfare, the utility of all the possible states of the world is weighted by the probability that these states occur.

The two approaches will only yield the same result when the social welfare function is of the utilitarian form (see Milne and Shefrin [1987]). Whilst the *ex post* approach has its proponents, most notably Broome [1991], it is fair to say that most economists, particularly those involved in empirical research, have adopted the *ex ante* approach. This may in part be due to the fact that the *ex ante* approach lends itself much more readily to empirical investigation. This author feels that the *ex post* approach has much to commend it but, since resource

allocation decisions do principally affect future (rather than current) patients, not least because they may be the one denied treatment in the future, it seems legitimate to give weight to the *ex ante* preferences of potential patients when making *ex ante* resource allocation decisions.

Finally, it could be argued that, since the general public pay for health care, their preferences should be the chosen basis for the weights used in the resource allocation process. This is consistent with conventional welfare economic theory which suggests that public sector decisions should, so far as possible, reflect the preferences of all those who will be affected by these decisions.

For these reasons, the MVH Group decided that its main fieldwork would concentrate on eliciting the preferences of the UK population and in a large-scale general population study (henceforth referred to as the Main Study) we went to considerable lengths to ensure that the sample was as representative of the wider population as possible. In this thesis, the empirical results presented in Chapters 2, 3 and 4 are based on the valuations of the general public whilst the exploratory studies on measuring attitudes towards equity in Chapter 5 are based on valuations from convenience samples. It is important, however, that valuations are elicited from as many different population sub-groups as possible since empirical evidence on inter-rater differences will illuminate the issues surrounding whose values should count. Ultimately, though, this decision will be a political not a scientific or empirical one.

How valuations should be derived raises two important questions: one concerns the choice of valuation method(s) and the other concerns the approach adopted to valuing all 245 EuroQol health states. An important consideration in the choice of method is the level of measurement that is required. Valuation methods can produce scales that are ordinal, interval or ratio (see Froberg and Kane [1989]). An ordinal scale is one in which health states are ranked in order of severity but there is no indication of how much more severe one state is compared to another.

An interval scale provides information on how far apart those states are in terms of severity but it does not indicate the absolute magnitude of severity. A ratio scale is achieved when the distance from zero is known for at least one state, and thus the absolute severity can be determined for all states. CUA requires that cardinal indices be assigned to each health state on an interval scale (see Lipscomb [1982] and Chapter 1.1 above).

Three methods that in principle generate valuations that lie on an interval scale have been widely used in a number of studies; the visual analogue scale (VAS), the standard gamble (SG) and the time trade-off (TTO). For this reason, these methods have been used in the studies conducted by the MVH Group. Chapter 2.1 outlines the relative merits of the three methods and discusses how valuations might differ between them. Without wishing to pre-empt this discussion too much, suffice it to say here that because no systematic relationship between the methods could be established and in the face of limited resources, the MVH Group was committed to choosing between the SG and TTO for use in the Main Study. Reasons for removing the VAS from this choice can also be found in Chapters 2.1 and Chapter 2.2 considers the criteria that can be used to make a choice between the SG and the TTO. In the light of the empirical evidence presented in Chapter 2, a variant of the TTO was the chosen for use in the Main Study and the results are presented in Chapter 3.1.

Since it is not feasible to elicit direct valuations for all 245 EuroQol health states, a choice has to be made about how best to interpolate some of the indices. There are essentially two different approaches that can be adopted here: one I will refer to as the decomposed approach, the other as the composite approach. The former involves asking the respondent to value each level within a particular dimension assuming that the levels of all other dimensions are held constant. Thus, the decomposed approach requires few (and in some cases no) valuations for composite health states, although most studies that have adopted this approach have elicited valuations for a small subset of composite states (see Torrance *et al*

[1982]). Valuations for composite health states can then be generated by specifying a multi-attribute function (MAUF). The problem with this approach is that the conditions that the MAUF must satisfy are stringent; the least-restrictive model (in which the MAUF is multilinear) requires utility independence which means that preferences for various level of each dimension do not depend upon the particular levels at which the other dimensions are fixed.

Because of the restrictions imposed on preferences by the decomposed approach, the MVH Group decided to adopt the composite approach whereby each respondents is asked to value a subset of composite EuroQol health states. An important consideration when choosing these states - and when choosing a larger subset from which to sample if the number that each respondent can value is deemed to be too small - is that they should be widely spread over the valuation space so as to include as many combinations of levels across the five dimensions as possible. This is subject to the constraint that the states are likely to be considered plausible by respondents; for example, respondents would probably have difficulty imagining a state in which they were confined to bed or were unable to wash or dress themselves yet had no problems with their usual activities. The next step, using appropriate regression or statistical techniques, is to estimate a model which allows valuations for all 245 EuroQol states to be interpolated from direct valuations on a subset of these. The advantage of this approach is that fewer restrictions need to be placed on the resultant model. The particular model which best describes the data from the Main Study is presented in Chapter 3.2.

From the results presented in Chapter 3.1, it appears that TTO valuations are affected by the age and the sex of the respondent; those aged 18-59 have higher valuations than those aged 60 or over and men have higher valuations than women. Given that a large proportion of health care expenditure is directed towards elderly populations, and to a lesser extent towards women, it is possible that policy-makers may wish to give more weight to the valuations of the appropriate sub-group when making resource allocation decisions within that sub-

group than they do to valuations from the general population. Against this background, Chapter 3.3 presents different valuation 'tariffs' according to the age and sex of the respondent using the model described in Chapter 3.2.

Whilst the issue of whose values should count has received a great deal of attention in the literature, the question of how these individual responses should be aggregated has received much less attention. Economists have generally advocated that the theoretically correct way to aggregate individual preferences is to calculate the mean value from any given distribution but, in the realm of public policy, an alternative view is that group preferences should be expressed in terms of the median. Because of the lower and, particularly, upper bound on health state valuations, TTO valuations for less severe health states are negatively-skewed whilst those for more severe states are positively-skewed. Thus, the transition from a more severe state to a less severe one is valued less according to the mean than the median and hence the choice of the measure of central tendency may have important implications for resource allocation decisions. Because of the methodology employed, the 'tariff' presented in Chapter 3.2 provides a good approximation of mean values. The purpose of Chapter 3.4 is to present a tariff based on median values.

Further methodological questions

Since valuations in the Main Study and the sets of 'tariff' values reported in Chapter 3 were based upon responses to TTO questions, it is important to consider the extent to which such responses can be interpreted in the way that many researchers interpret them; namely, as HRQoL indices which lie on an interval scale. Chapter 4 looks at the effects of three potentially important sources of bias in interpreting TTO responses in this way: i) lifetime reallocation of consumption; ii) time preference; and iii) duration. Chapter 4.1 is directed towards explicitly setting out the likely magnitude of the first two of these biases and shows that for responses to TTO questions to provide unbiased estimates of

the HRQoL associated with different health states, it is necessary that there is a zero rate of time preference. Chapter 4.2 reports on a pilot study designed to test the plausibility of this assumption using the TTO method, and also looks at the impact that the time spent in a particular state can have on its subsequent value. The results show that using the TTO to assess the effect of duration on health state valuations (whether time preference can be accounted for or not) is problematic and, therefore, Chapter 4.3 reports on a much larger general population study in which valuations for three different durations were elicited using the VAS method.

Since the primary purpose for eliciting individual valuations is to inform decisions at the social level, it is important to consider the extent to which individual valuations can be used to express social preferences. This is the purpose of Chapter 5. Chapter 5.1 looks at how distributional issues that are known to be important in the context of health care can be incorporated within the QALY framework. An approach is suggested which uses a particular class of social welfare functions (SWF) that allows efficiency and equity to be considered independently and which is sufficiently flexible to represent a wide range of social preferences. Particular attention is given to the log-linear form which implies inequality aversion and which the results from a preliminary experiment suggest might indeed represent the average preferences of groups of respondents. Although the approach presented in Chapter 5.1 appears to be a feasible one, it is necessary to explore different approaches, not least because very few empirical studies have addressed the efficiency-equity trade-off in health care. In Chapter 5.2, the method first described by Atkinson [1970] in order to measure the shape of the SWF with respect to income distribution is used to allow the shape of the SWF with respect to the distribution of health gain to be measured. In Chapter 5.3, another approach is considered. Rather than, as in Chapters 5.1 and 5.2, look at how equity might be taken account of after an efficient allocation has been determined, this chapter looks at the use of the person-trade-off (PTO) method, which, its proponents argue, captures (efficiency and equity) concerns that are

relevant to social decision-making in one question. Specifically, the chapter reports on a pilot study which was designed to assess whether two treatments that yield the same benefit to the individual are also considered, by that individual, to yield the same benefit to society (as measured by responses to PTO questions).

The thesis concludes with Chapter 6 which considers three important themes to have emerged from the thesis in the context of how they may provide a general strategy for future research into the measurement and valuation of health. The themes are: 1) the nature of individual preferences; 2) the measurement of social welfare; and 3) the use of different tariffs. Under each of these of headings, suffice it to say here that: 1) attempts must be made to establish the cognitive processes that respondents use to arrive at their responses in order to get a better understanding of *why* valuations differ in addition to *how* they differ; 2) given that the aggregated health state utilities are, for the purposes of public policy, interpreted as measures of social value, research efforts should be directed towards assessing the shape of and, perhaps more fundamentally, the arguments in the health-related SWF; and 3) it should now be possible in future cost-utility analyses that use EuroQol 'tariff' values to consider the extent to which the cost-utility ratios are affected by the choice of tariff and, crucially, to consider the implications for resource allocation decisions.

CHAPTER 2: METHODS FOR VALUING HEALTH STATES

CHAPTER 2.1: THE RELATIONSHIP BETWEEN VAS, SG AND TTO

Introduction

Most empirical studies to date have shown that the VAS, SG and TTO yield different valuations from the same respondents for identical descriptions of health. Torrance [1976] and Read *et al* [1984] found correlations of 0.65 between the scores elicited by the SG and the TTO methods. Torrance concluded that these two methods are equivalent, but Read *et al* emphasised that high correlations can coexist with systematic differences between sets of scale values. Both studies found that VAS valuations were lower than SG and TTO ones and. in a comparison of mean VAS and TTO values, Torrance concluded that "the two techniques exhibit a systematic relationship [that] can be approximated by a number of different functions. Two that fit well ... are a logarithmic function and a power function" (p.134), although these relationships did not hold at the individual level.

Wolfson *et al* [1982] obtained somewhat different results. Having estimated linear relationships, they concluded that standard gamble values were much higher than VAS ones, with TTO values generally lying somewhere in between. In a more recent study, Hornberger *et al* [1992] found poor correlation between any of the methods at the individual level and, in a comparison of mean values, found the TTO produced the highest values, the VAS the next highest, and the SG yielded the lowest values. This contradicts much of the earlier work and may be in part a function of the fact that, whilst other studies invoked hypothetical scenarios, this study elicited patients' valuations of their own health. In a review of the literature to 1988, Froberg and Kane [1989] concluded that "while correlations between methods are usually moderately high, the different methods do not necessarily produce equivalent scale values".

Because valuations from the VAS are elicited in a choiceless context, and thus do not require people to make trade-offs between different arguments in their utility function, the method is commonly regarded by economists as theoretically inferior to the choice-based SG and TTO methods. A notable exception is Broome [1993] who regards a method he describes as identical to the VAS (although he does not use the term himself) as "uncontaminated" by factors which he considers to be irrelevant to the measurement of "goodness"; like risk attitude in the SG or time preference in the TTO. However, the VAS has the practical advantages of being simpler to complete and cheaper to administer than either the SG or the TTO. Consequently, it is widely used in clinical and evaluative studies. If an algorithm can be found which maps VAS values onto SG and/or TTO ones and if, crucially, the relationship is stable, then it might be possible to elicit valuations via (cheap and simple) VAS methods and "convert" them into (theoretically superior) SG and/or TTO values. Furthermore, the nature of these algorithms may provide useful insights into *why* different methods yield different valuations.

Whilst the studies highlighted above have shown beyond doubt that different methods can be expected to yield different sets of valuations, it is questionable to what extent their results and conclusions are generalisable. This is because: 1) all were based on small samples of convenient populations (none consisted of more than 67 people); 2) the analysis was performed on aggregate- rather than individual-level data, thus making the choice between competing models more difficult as well as making inefficient use of the data; and 3) all the results were generated using ordinary-least squares (OLS) regression analysis which is inappropriate given the (censored) nature of the data being analysed. This chapter considers whether there exists an empirical relationship between the valuations from the VAS and those from two variants of the SG and TTO: one using specially designed boards and cards (Props); and one using a self-completion booklet (No Props). In contrast to previous studies, the results are from a large scale study of the general population, and are based on individual-level analysis using a Tobit model which takes account of the type of data typically dealt with in the health state measurement field.

Hypotheses

Despite their limitations, the results of earlier studies, together with some theoretical work, suggest two hypotheses. Firstly, for identical descriptions of health status, valuations from the VAS will be lower than those from the SG or TTO. This may result from respondents' use of different reference points in the valuation of health states on the VAS than on the SG and TTO. For example, in the VAS, a respondent may take full health as their reference and value dysfunctional health states as losses from this state. In a qualitative study, Morris and Durand [1989] suggest that VAS responses are indeed constructed in this way. In the SG and TTO, on the other hand, the respondent is asked to imagine already being in a dysfunctional health state and, consequently, this state becomes the reference point.

It seems entirely plausible that the value given to a particular health state will be a function both of its severity and of the state that it is viewed from, just as our perception of how fast a car is travelling is a function both of its actual speed and the speed it had previously been travelling at, and that people may give some special status to their current position, and react asymmetrically to movements away from that position, placing greater weight on what they perceive as losses *vis-à-vis* the reference point than on what they perceive as gains. These ideas are developed by Kahneman and Tversky [1979] who propose that an individual's value function is concave with respect to gains, convex with respect to losses, and steeper at each level of loss than at the corresponding level of gain, as depicted in Figure 2.1.1. The frequently observed substantial disparities between what people say they would be willing to pay (WTP) for some marginal benefit, and what they would be willing to accept as monetary compensation for a comparable marginal disbenefit, is often taken as evidence of such an effect (see Kahneman, Knetsch and Thaler [1990]). A (non-stochastic) value function with these properties will result in lower valuations from the VAS than from either the SG or the TTO.

Secondly, based on two assumptions about individual preferences, mapping functions for SG variants are expected to be different from those for TTO variants. The first assumption is that people are risk averse: if utility on the ordinate is plotted against length of life on the abscissa, the resulting utility function is concave to the origin. This implies that people will be less willing to accept the gamble outcomes in the SG and more willing to accept the certain outcome. The second assumption is that people have positive time preference: they value years of life in the near future more highly than they value years of life in the more distant future. This implies that people will be more willing to give up years of life at the end of a profile, as in the TTO. Thus, both assumptions imply that, for the same health states, SG values will be higher than TTO ones. Significant differences across variants of the same method are not anticipated.

Study design

The Questionnaires

Previous piloting indicated that when asked to value the same EuroQol health states using three methods, respondents could not effectively assess more than six states on each method (plus two anchor states on the VAS: full health and death). The states were chosen to be widely spread over the valuation space and, in the SG and TTO tasks, were presented to respondents in a standard order.

Each respondent was first asked to describe their own health using the EuroQol descriptive system. They were then asked to rank all 8 health states. It was explained that each state was to be regarded as lasting for 10 years without change, followed by death. The respondent was asked to indicate where they would rate their own health on a vertical VAS, with endpoints of 100 (best imaginable health state) and 0 (worst imaginable health state). They were then asked to rate the 8 health states on an identical VAS. Once full health and dead had been removed, the remaining six states (which were always presented in the same order), were valued using one variant of the

SG and one variant of the TTO. At the end of the interview, personal background data were collected from each respondent.

The SG asks the respondent to choose between the certainty of an intermediate health state and the uncertainty of a treatment with two possible outcomes, one of which is better than the certain outcome and one of which is worse. For a state, h_b, rated as better than dead, the intermediate state is h_b and the treatment outcomes are full health and death, respectively. For a state, h_w, rated as worse than death, the intermediate state is death and the treatment outcomes are full health and h_w. In both cases, the object is to find the probability, p, at which the respondent is indifferent between the two alternatives.

For SGP, a sliding scale on a specially designed board showed the varying chances of success and failure of treatment. For each health state, the respondent was initially asked to choose between living for 10 years in that state and a treatment which would return them to full health for 10 years (i.e. a 100% chance of success). Then, to determine whether the state was considered to be better or worse than death, the respondent was asked to choose between 10 years in that state and a treatment which would result in immediate death (i.e. a 0% chance of success). If they preferred the former, the protocol for states rated as better than dead was used and if they preferred the latter, the protocol for states rated as worse than dead was used. In both cases, the chances of success were presented in intervals of 10% in a "ping-pong" fashion i.e. 90% success, 10% success, 80% success, etc. The question was complete when either a) preferences changed over a 10% interval (e.g. the treatment was preferred when it had a 70% chance of success but the certain health state was preferred when the treatment had a 60% of success), in which case the state would be valued at half-way between the two probabilities (i.e. 0.65 in this example), or b) indifference was reached.

The SGNP variant consisted of a self-completion booklet which showed the (certain and uncertain) alternatives on the left-hand page and the chances of success and failure relating to the uncertain treatment on the right-hand page, which were listed in 10% intervals ranging from a 100% chance of success at the top to a 0% chance of success at the bottom. The respondent was initially shown the protocol for states rated as better than dead and was asked to place a tick alongside all those probabilities where they would prefer the treatment, a cross alongside all those probabilities where they would prefer the certain health state, and an equals sign alongside the probability at which they would find it hardest to choose between the certain state and the treatment. If the treatment was preferred when it was certain to result in immediate death (i.e. when a tick was placed alongside a 0% chance of success), the respondent turned over the page and was presented with the protocol for states rated as worse than dead, which was competed in a similar fashion. Details of both SG protocols can be found in Gudex [1994] and examples of the protocols used are shown in the appendix.

The TTO asks the respondent to choose between two alternatives. For a state, h_b, rated as better than dead, the first alternative is to live for a defined period of time, t, in h_b and then die. The second alternative is to live for a shorter period of time in full health and then die. For a state, h_w, rated as worse than dead, the first alternative is to die immediately and the second alternative is a number of years in h_w followed by a number of years in full health (which combined sum to t). In both cases, the time in full health, x, is varied until the respondent is indifferent between the two alternatives.

For TTOP a sliding-scale on a double-sided board showed the number of years spent in each alternative: one side was used for states rated as better than dead and the other for states rated as worse than dead. Using the former side of the board, for each health state, the respondent was first asked to choose between living for 10 years in that state and living for 10 years in full health. Then, to determine whether the state was considered to be better or worse than death, the respondent was asked to choose between 10 years in that state and immediate death. If they preferred the former, they continued to use the side of the board for states rated as better than dead and if they

preferred the latter, the side of the board for states rated as worse than dead was used. In both cases, the number of years spent in full health were presented in units of one year in a "ping-pong" fashion i.e. nine years, one year, eight years, etc. The question was complete a) when preferences changed over a one year period (e.g. full health was preferred when it was for 7 years but 10 years in the particular state was preferred when full health lasted for 6 years), in which case the state would be valued at half-way between the two lengths of time (i.e. 0.65 in this example), or b) when indifference was reached.

The TTONP variant consisted of a self-completion booklet which showed the two health profiles on the left-hand page and the number of years spent in the profiles on the right-hand page, which were listed in units of one year, ranging from ten years in full health at the top to zero years in full health at the bottom. The respondent was initially shown the protocol for states rated as better than dead and was asked to place a tick alongside the cases where they preferred a certain number of years in full health to 10 years in the particular state, a cross alongside the cases where they preferred 10 years in the particular state to the number of years in full health, and an equals sign alongside the case where they consider a certain number of years in full health to be equivalent to 10 years in the particular state. If immediate death was preferred (i.e. when a tick was placed alongside zero years in full health), the respondent turned over the page and was presented with the protocol for states rated as worse than dead, which was competed in a similar fashion. Details of both TTO protocols can be found in Gudex [1994] and examples of the protocols used are shown in the appendix.

Ten years was chosen as the time horizon because it was considered long enough for respondents to be able to make meaningful sacrifices and to be able to distinguish between states but not too long so as to be unrealistic for older respondents. It is recognised that this time horizon would have been unrealistically short for many younger respondents but it was felt that other alternatives (such as variable time horizons based on a person's own expected life expectancy) would have created even greater problems of measurement and interpretation.

The sample

The sample was drawn from adults aged 18 and over in the general population. A random sample of 700 addresses was drawn from 13 regional areas in the U.K. by Social and Community Planning Research (SCPR) using the postcode address file. The main fieldwork was carried out by 25 specially trained interviewers between March and May 1992. In order to study the two variants of the SG and TTO (from now on referred to as P and NP, to denote Props and No Props, respectively) and to test whether the order of presentation of these tasks influences valuations, each interviewer was randomly allocated to one of eight experimental groups (2 methods x 2 variants x 2 orders of presentation).

Of the 700 addresses selected for sampling, 88 (13%) were found to be 'out of scope', being non-residential, empty/derelict, untraceable, or not yet built. Of the remaining 612 addresses, 335 interviews were achieved, giving a 55% response rate on in-scope addresses. The main reason for unsuccessful interviews was a refusal by the selected person. Table 2.1.1 shows that the sample was similar to the general population in terms of age and sex, although there was some response bias in favour of the more educated. Table 2.1.2 shows that, by chance, more respondents were in the four groups containing the TTONP variant than in the four groups containing TTOP. No statistically significant differences were found between the groups on the basis of their socio-demographic characteristics.

Table 2.1.2 also shows that 14 interviews were incomplete (defined as one or both of the main valuation methods being missed out entirely) and these respondents have been excluded from subsequent analysis. These respondents were scattered amongst the eight experimental groups and did not differ in terms of background characteristics from the remainder of the sample. Finally, Table 2.1.2 shows that the number of missing observations from the remaining respondents is very small, particularly for TTOP.

Methods

Re-calibration of scores

In CUA, aggregation across respondents is achieved by measuring all individual valuations on a common 0 to 1 (dead to healthy) scale, as noted in Chapter 1.1. Because respondents could locate full health and death anywhere on the VAS, it is necessary to re-calibrate raw VAS scores so that full health and death are of equal value to everybody, and hence "the unit of health" is the same across all respondents. If full health and dead are assigned scores of 1 and 0 respectively, then, using notation introduced above, h_b=p on the SG and x/t on the TTO, whilst h_w=-p/(1-p) on the SG and -x/(10-x) on the TTO. Thus, negative scores lie on a ratio (not an interval) scale and, unlike the case for states rated better than dead, are theoretically unbounded (though, given the response categories available to respondents, in this study they are bounded by -19 and in the Main Study reported on in Chapter 3 they are bounded by -39).

The asymmetry between positive and negative values poses problems since those respondents rating a state as worse than death will have a much greater impact on the measures of central tendency than those respondents rating it as better than death. As Torrance [1984] noted, "this issue of large negative values and what to do about them needs much more study". Patrick *et al* [1994] transformed their negative values so that scores for states rated as worse than dead were bounded by -1 i.e. symmetrical to the upper bound of +1 for states which are rated as better than dead. This transformation was justified on statistical grounds but there is possibly a psychometric justification as well; namely that respondents may treat the scale for states worse than dead in the same way as they are assumed to treat the scale for states better than dead i.e. as an interval (not a ratio) scale. For these reasons, then, valuations for states worse than dead have been transformed such that h_w=-p on the SG and (x/10)-1 on the TTO.

For the purposes of the analysis in this chapter, all SG and TTO scores have then been re-calibrated so that death equals 0.5 for each respondent. This is so that mathematical manipulations of the VAS scores are well-behaved functions, and so that the SG and TTO scores can be transformed onto a logarithmic scale for the analysis of power functions. Denoting the raw VAS scores for full health, death and some dysfunctional health state h by v_f, v_d and v_h respectively, a health state index, h, can be derived according to the following decision rule:

if $v_f \leq v_h$	$h = 1$
if $v_d < v_h < v_f$	$h = 0.5 + [\ 0.5*(v_h - v_d)/(v_f - v_d)\ ]$
if $[v_d - (v_f - v_d)] < v_h < v_d$	$h = 0.5 - [\ 0.5*(v_h - v_d)/(v_f - v_d)\ ]$
if $v_h \leq [v_d - (v_f - v_d)]$	$h = 0$

Diagnostic tests for heteroskedasticity and functional form are included in the analysis, which should indicate any problems with this approach.

Independent variables

Previous studies of the relationship between scores elicited by different techniques have concentrated on the relationships between aggregate (mean) scores. However, an alternative approach, for which a larger amount of data are available, is the estimation of mapping functions based on the individual-level data. Initially, a model using all the data was estimated which contained binary variables for variant (P or NP) and method (SG or TTO). This model failed the specification tests outlined in the next section and thus the relationship is estimated separately for both variants. This produces four mapping functions to be estimated. The aim of these functions is to show the relationship between the VAS and the variants of the SG and TTO, irrespective of the health state being valued.

With SG and TTO scores as regressands, the regressors of interest are different transformations of VAS scores. However, in the analysis of the individual-level data, information is available for a number of respondent background characteristics which have been found to significantly affect health state valuations (for a review of the literature to date see Froberg and Kane [1989]). If these independent variables affect scores derived from the methods differently, then these factors should be taken into account in estimating any mapping function. The characteristics considered in this analysis are sex, age (considered by the introduction of two dummy variables representing three age-groups: 18 - 29 years; 30 -59 years; and 60 years or more), and the individual's rating of their own health status on the VAS. Each of these variables is entered as an independent factor and as a cross-product term with the visual analogue score. Details of the abbreviations used for the independent variables are given in Table 2.1.3.

Models

Five functional forms were estimated to represent the relationship between VAS and SG/TTO scores: a linear, a quadratic and a cubic model plus two log-linear models. The linear model included only the VAS score, (V_{ix}), and the other independent variables. In the quadratic model $(V_{ix})^2$ was added, and in the cubic model $(V_{ix})^3$ was included:

Model 1: $$Y_{ix} = a + \beta V_{ix} + \Phi_k Z_{ixk} + e_{ix}$$

Model 2: $$Y_{ix} = a + \beta V_{ix} + \Gamma V_{ix}^2 + \Phi_k Z_{ixk} + e_{ix}$$

Model 3: $$Y_{ix} = a + \beta V_{ix} + \Gamma V_{ix}^2 + \Theta V_{ix}^3 + \Phi_k Z_{ixk} + e_{ix}$$

in which Y_{ix} is the re-calibrated score on the SG/TTO for state x from individual i, V_{ix} is the score for health state x elicited from individual i by the visual analogue method, Z_{ixk} is a vector of k independent variables for individual i, and e_{ix} is an error term.

Two log-linear models were estimated:

Model 4: $$Y_{ix} = A\, V_{ix}^{\beta}\, Z_{ixk}^{\emptyset k}\, e_{ix}$$

Model 5: $$(1-Y)_{ix} = A\, (1-V_{ix})^{\beta}\, Z_{ixk}^{\emptyset k}\, e_{ix}$$

The difference between the models relate to the testing of two assumptions: 1) a health state rated as good as full health on the VAS will also be rated as good as full health on the SG/TTO; and 2) a health state rated as far below dead as full health is above dead on the VAS will also be rated as far below dead as full health is above dead on the SG/TTO. Model (4) makes the first assumption whilst the size of the constant in the regression can be used to test for the second assumption. Model 5 makes the second assumption whilst the first can be tested from the regression results. In the estimation of his power function, Torrance [1976] assumed equality between valuations from different methods at both end-points of the scale.

Estimation and Testing

To take account of the fact that valuations on both the SG and the TTO were within the range 0.025 to 0.975, Tobit estimation was undertaken with censoring at both the top and bottom ends. The models were estimated using maximum likelihood within LIMDEP (see Greene [1992]). Two specification tests have been included in this analysis: a modified RESET test, and a test for heteroskedasticity in the error terms. The modified RESET test is undertaken in a two-step process. In the first stage, the model is estimated and the linear function is calculated:

$$¶_{ix} = ß_j X_{ixj}$$

in which: ßj are the j coefficients estimated in the Tobit model, where j refers to all independent variables in the model; and X_{ixj} is the vector of all independent variables which includes value(s) on the VAS. In the second stage, the square of the linear function is added to the equation, and the t-statistic on this variable can be used as a test of the functional form of the original model.

All models were initially estimated on the assumption that the variance term is constant across all observations i.e. $\sigma_i = \sigma$. The models were then estimated by allowing the variance term to be an exponential function of either the VAS (V_{ix}), the square of the VAS $(V_{ix})^2$, or the square of the linear function from the unadjusted Tobit estimation $(¶ix)^2$ i.e. $\sigma_i = \sigma e^{\gamma Zi}$. These homoscedastic and heteroscedastic models, respectively, were compared using the likelihood ratio (LR) test.

Calculation of mapping functions

Because the data was analysed using a Tobit model, the predicted values from the regressions have to be transformed according to the following equations:

$$E[y_i] = L\,Ø_L + U\,(1 - Ø_U) + (Ø_U - Ø_L)\,ß'\,x_i + \mu_i\,(\tilde{O}_L - \tilde{O}_U)$$

in which: L and U are the lower and upper bounds of the range within which y_i are restricted; ß is the corresponding vector of coefficients for the set of independent variables x_i; μ_i are the standard errors of the (possibly heteroskedastic) error terms; Ø is the standard normal cumulative distribution; Õ is the standard normal distribution; and:

$$\tilde{O}_j = \tilde{O}\,[(j - ß'\,x_i)\,/\,\mu_i]$$

and $\quad \emptyset_j = \emptyset\,[(j - \beta' x_i) / \mu_i]$

in which: j = L, U. In contrast to OLS methods, the predicted values calculated from these equations are constrained within the desired range [L ,U].

Results

The results of the various linear, quadratic and cubic models for each of the four variants are shown in Table 2.1.4. Where possible, results are shown for models which do not show evidence of functional form misspecifications or heteroskedastic disturbances at the 99% level. As it turned out, the LR test for all models suggested that there was no significant heteroscedasticity present. Whilst it would then be justified to use the homoscedastic models, some models adjusted for heteroscedasticity offered improvements in the RESET statistics. Therefore, results of Tobit models with adjustment for heteroskedasticity are shown for all valuation methods except SGNP, for which no adjustment made any significant improvement in this regard.

Few of the independent variables associated with respondent background characteristics are significantly different from zero at even the 5% level. For example, there is little evidence of any systematic effect of gender or self-rated health on the difference between scores from the VAS and the other methods. However, for both SG variants, the negative coefficient on YOUNG and the positive coefficient on YOUNG*VAS are both significant at the 5% level and, although not statistically significant, the sign on these coefficients is the same for the TTO variants. These results imply that, given the same VAS score, younger respondents tend to give lower valuations to more severe health states on the choice-based methods, particularly the SG. This divergence between age-groups, however, decreases as the VAS rating increases.

The results for the two log-linear models are not very encouraging and hence the results are not presented here. For the logarithmic model (Model 4), the likelihood ratio indices are below 0.05 for all methods, there is evidence of heteroskedastic disturbances in all four equations and the estimated models for all but TTONP fail functional form tests. Similarly, with regard to the estimation of a power function (model 5), the likelihood ratio indices are less than 0.01 for all methods, there is evidence of heteroskedasticity in all simple and adjusted equations and no models pass functional form tests at the 5% level. The estimated likelihood ratio indices illustrate that the ability of power functions to explain variations in this date-set is negligible. Moreover, although unlikely to be valid, the estimated coefficients on the constant terms are not close to one as assumed by Torrance [1976].

Mapping functions can be derived based on the predicted value formulae described above, a simulated index of possible VAS scores, and the average values of other independent variables. The average values of the other independent variables for which the predicted value functions have been estimated are given in Table 2.1.5. Average values of even the discrete variables are used, since this may reflect the characteristics of a general sample of respondents (e.g. the TTO Props case in which 29% are aged under 30 years and 20% are aged over 60 years). In contrast to the log-linear models, the linear functional forms pass misspecification and heteroskedasticity tests at the 99% level of significance for the majority of methods and, as indicated by the likelihood ratio index, explain a significant proportion of the variation in the data. As a result, mapping functions have been generated based on the coefficient estimates from the linear models only. These functions are shown in Figures 2.1.2 to 2.1.5.

As well as highlighting systematic differences across methods and variants, the mapping functions suggest that the relationship with VAS scores is a function of the severity of the state. Figures 2.1.3 and 2.1.5 both show that values from the no props variants are higher than those for the VAS and that this difference increases as the severity of the state increases. Figure 2.1.2 shows that TTOP valuations are higher than VAS ones for mild states and lower for more severe states, although SGP

valuations are broadly similar to VAS ones for a wide range of scores (see Figure 2.1.4). Note that the functions estimated for the props variants imply that for states worse than dead, a marginal increase in the VAS score is associated with a marginal decrease in the corresponding TTO or SG score. This finding is counter-intuitive, and thought to be related to both the lack of observations in this quadrant and the dominating effect of the function in other parts of the valuation space.

Discussion

This chapter has used health state valuation data from a large-scale general population study to estimate an empirical relationship between VAS scores and scores elicited from two variants of the SG and the TTO. The analysis was based on individual level data using the Tobit model which takes account of the (censored) nature of the data. A number of different functional forms were tested and a range of diagnostic tests were applied to the competing models. Logarithmic and power function formulations were outperformed by more flexible (linear, quadratic or cubic) functional forms, both in terms of specification and ability to explain variations in the data. In particular, the results do not lend support to the hypothesis that VAS and SG/TTO values can be related by a concave power function, since this functional form failed all diagnostic tests.

In contrast to expectations, differences between the mapping functions are more pronounced across variant than across method. That valuations differ by method variant is an important finding which has been noted elsewhere (for example, see Nord [1992]). This result offers an explanation for the lack of consensus regarding the comparability of the different methods, since although the studies referred to in this chapter used the same methods, the way in which they were administered differed enormously. This suggests that comparisons can only be made between valuations from different studies if both the same method and the same variant were used.

One of the underlying hypotheses was that, for identical descriptions of health status, VAS valuations would be lower than SG/TTO ones. Thus, for high valuations, the intercept of any mapping function between VAS and SG/TTO scores is expected to be greater than one, whilst the gradient will be less than one. The gradients of all mapping functions for the linear models are indeed shallower than the 45° line. However, the mapping functions imply that a VAS score greater than about 0.8 is associated with a lower SG and TTO score. This suggests that the disutility associated with the very mild health states is greater for methods in which the reference point is the dysfunctional health state itself than for a method in which the reference point is likely to be full health. This result appears to cast doubt on the value function hypothesised by Kahneman and Tversky's Prospect Theory [1979].

However, inspection of the distribution of health state values suggests that the majority of respondents do behave in accordance with Prospect Theory i.e. have SG and TTO values that are higher than VAS ones for very mild states. Therefore, the fact that mapping functions cross the 45° line at about 0.8 appears to be explained by a small number of 'outliers', who give high VAS scores yet give very low SG/TTO scores. Although there were no grounds for excluding such responses from the data set, this finding suggests that the functions presented here should be treated with caution, and their interpretation made clear. Because of the estimation technique used, the results represent mapping functions that approximate how mean VAS scores can be converted into mean SG and TTO ones.

Another unexpected result (at least in the presence of a reference point effect) is the suggestion from Figure 2.1.2 that for the more severe health states, TTOP values are lower than VAS ones (notwithstanding the non-monotonicity at very low values discussed above). Given that this is not the case for any of the other methods, it is likely to be explained in terms of the differential effect that TTOP has on valuations. One possible explanation is that TTOP (which has a ten-year scale on a board) is the method which focuses the respondent's attention explicitly on the length of time spent in a health state. If the disutility associated with a severe health state increases as the

time spent in that state increases (see the discussion in Chapter 4), then lower valuations will be elicited for such states from methods which focus more explicitly on the time dimension.

Of course, if this is the reason why TTOP valuations are lower than VAS ones, then one would expect to observe a similar (although perhaps less powerful) effect with respect to TTONP which (unlike both SG variants) also asks respondents to think in terms of time. That this is not the case raises the more general issue of why valuations from the no props variants are higher than those from the props variants, particularly for lower VAS values. A possible explanation relates to the different ways in which response categories were presented to respondents.

In the props variants, respondents were presented with choices in a "ping-pong" fashion, moving back and forth between higher and lower probabilities of success in the SG and longer and shorter life expectancy in the TTO. In the no props variants, on the other hand, respondents were presented with all possible responses at once. These were listed from high to low probability of success in the SG, and from long to short life expectancy in the TTO. It is likely that respondents would have started from the top of the page and worked their way down. This may have resulted in an analogue of the reference point effect, in which respondents gave special status to favourable outcomes and hence to higher (inferred) health state valuations (as evidenced by the no props mapping functions which are above the 45° line and 'fan-in' from above). This suggests that no props variants may introduce systematic bias into valuations.

It was also hypothesised that mapping functions for the SG would be different from those for the TTO; specifically, that, for the same VAS values, SG values would be higher than TTO ones. In terms of the relationship between the different choice-based methods, two patterns emerge from the mapping functions: 1) for VAS scores above about 0.4, SGP values are lower than TTOP ones, whilst SGNP and TTONP values are very similar; and 2) for VAS scores below 0.4, SGP values are higher than TTOP

values and SGNP values are higher than TTONP ones. That TTOP valuations are higher than SGP valuations for high VAS values goes against *a priori* expectations, but might be explained in terms of the relative weight respondents' attach to the numeraire they are asked to sacrifice in order to gain an improvement in health. In the SG, health improvements are valued in terms of the level of risk (usually of immediate death) a respondent is prepared to accept, whilst in the TTO they are valued in terms of the amount of life expectancy a respondent is prepared to sacrifice. Thus, the results may indicate that sacrificing an extra six months of life expectancy is more valuable to respondents in this study than taking an extra 5% risk of death. That the expected relationship between SG and TTO holds for lower VAS values might be explained in terms of the explicit reference to the time spent in the health state in the TTO exercise.

With respect to the impact of respondent background characteristics, it is found that only age appears to have a significant impact on the resultant mapping functions. For both SG variants, it was found that those aged 18-29 years had significantly lower intercepts and steeper slopes than older respondents, suggesting that, for the same (low) VAS score, younger respondents tend to give lower valuations on the SG, but that this difference decreases as the VAS score increases. Although the same pattern is observed for the TTO variants, the coefficients on the relevant variables (YOUNG and YOUNG*VAS) fail to reach conventional levels of significance. It is unclear why younger respondents should have different mapping functions from other respondents, or why this difference should be more pronounced at the lower end, or why it should be more pronounced on the SG than on the TTO. The literature to date does not help to shed much light on this subject since, although age is generally regarded as having negligible a impact on health state valuations (see Carter *et al* [1976], Rosser and Kind [1978] and Kaplan *et al* [1978]), much of the analysis has concentrated on differences *within* valuation methods and not *across* them.

Nonetheless, the results suggest that, at least when considering severe states of health, younger respondents (for the same VAS score) are more willing to sacrifice life

expectancy and, relatively speaking, even more willing to risk death than are older respondents. It might be that younger respondents differ from older respondents more in their attitude to risk (as measured by different SG values) than in their attitudes towards time (as measured by different TTO values). However, that both these differences decrease as the VAS value increases suggest that either risk attitude and time preference are not independent of the health state being valued or that something else is being picked up here. These would appear to be important issues that future research efforts should be directed towards addressing.

This chapter has attempted to assess whether VAS valuations can be mapped into SG and/or TTO ones. If robust mapping functions could be estimated, then this would have the practical advantage of allowing valuations elicited from the cheap and simple VAS to be converted into theoretically superior SG and/or TTO ones. In addition, models could be developed that might explain the mapping functions, in the same way that Loomes [1993] has shown that Regret Theory in its non-stochastic form can explain the relationship between aggregate values in the Torrance [1976] and Wolfson et al [1982] data.

However, the results presented in this chapter suggest that the way the methods are administered is as important a determinant of the resultant mapping functions as the methods are themselves. This suggests that the way in which a question is framed can have a significant effect on responses: a fact which is increasingly recognised by many economists who now accept that changes in questionnaire design can bias a respondent's stated preferences. A fuller discussion of the possible reasons why this should be so is beyond the scope of the present chapter. However, the results in this chapter do suggest that an important consequence is that no single set of mapping functions is likely to explain the observed disparities between health state valuation methods.

CHAPTER 2.2: CHOOSING BETWEEN SG AND TTO

Introduction

Given that estimating a robust relationship between the valuations from different methods does not appear to be possible, it is necessary from the study described in Chapter 2.1 to choose one method to use in the Main Study. The VAS was not one of the contenders in the choice of "best" method because, as noted in Chapter 2.1, valuations from this technique are elicited in a choiceless context. Thus, they do not reflect the importance of health relative to other arguments in an individuals utility function and are not regarded as measures of utility, defined in its broadest sense.

The SG and TTO methods, on the other hand, both start from the premise that, given that health is an important argument in an individual's utility function, we can estimate the welfare change associated with a change in health if we can determine the compensating change in one of the remaining arguments in an individual's utility function that leaves utility unchanged. In the SG, health improvements are valued in terms of the level of risk (usually of immediate death) an individual is prepared to accept, which means assuming utility to be a negative function of such a risk. In the TTO, health improvements are valued in terms of the amount of life expectancy an individual is prepared to sacrifice by assuming utility to be a positive function of longevity. In this way, both the SG and the TTO can be viewed as sharing a common theoretical background.

In developing Expected Utility Theory (EUT), Von Neumann and Morgenstern [1953] showed that if a cardinal utility could be expressed as equivalent to a gamble, under certain assumptions, it would be a linear function of the risk involved in the gamble. In other words, the level of risk involved in standard gamble questions is linear in utility. This led many to regard the SG as the "gold standard" for health status measurement (see Torrance [1976] and Gafni [1994]). However, doubt has been cast on EUT both as a positive and as a normative theory. First, there is evidence that people systematically violate the axioms of EUT (see Llewellyn-Thomas *et al* [1982] and

Schoemaker [1982]). Thus, much of the appeal of the SG is lost since it will only be an accurate measure of utility if the axioms of EUT apply. Second, EUT focuses only on the expected utility of different outcomes, and there is increasing evidence that many people consider this to be an *irrational* basis on which to make decisions, preferring instead to take account of the process by which the outcomes were arrived at. This has led to a number of new theories which relax the independence axiom (for example, Regret Theory as developed by Loomes and Sugden [1982]).

The literature often distinguishes between utility, which results from decisions under uncertainty (as measured by the SG, for example), and value, which results from decisions based on certainty (see Gafni, Birch and Mehrez [1993]). Because in the TTO both of the alternatives presented to the respondents have outcomes that are known with certainty, it is said to produce a value, not a utility, function (see Pliskin *et al* [1979], Dyer and Sarin [1982] and Bennett *et al* [1991]). However, this is based on a very narrow definition of utility, one that has arisen as a direct result of Von Neuman Morgenstern EUT. In its broader sense, and one which is perhaps more relevant to the measurement of quality-of-life, utility is defined as a (cardinal) index of strength of preference. It is possible to measure this under conditions of uncertainty or certainty.

The SG is also advocated on the grounds that almost all decisions about health care are made under conditions of uncertainty (see Mehrez and Gafni [1991]). Whilst this is indeed the case, the appropriateness or otherwise of a valuation method is determined by its ability to act as a proxy for utility and not by its capacity to model the situation being valued (see Buckingham and Drummond [1993]). In this respect, the TTO may be considered more appropriate since, by definition, it gives the number of years in full health which are valued equally to a (longer) period in the health state being measured. In this respect, it collapses the relationship between the health state, its duration and its value into one single measure. Nevertheless, there is doubt about the validity of the underlying assumption of the TTO method that individuals are prepared to trade-off a constant proportion of their remaining years of life in order to improve their health status, irrespective of the number of years that remain (see Sackett and Torrance [1978] and Sutherland *et al* [1982]).

Criteria for choice

It is therefore difficult to choose between SG and TTO on theoretical grounds since valuations from neither method can automatically be assumed to map directly onto utility. This is an important point since it implies rejecting the idea that the SG should be regarded as the "gold standard" for measuring health state values. Instead, a choice between the SG and the TTO needs to be informed by their respective performance on empirical grounds. The evidence here is limited since relatively few studies have obtained within-respondent comparisons of the different valuation methods (see Chapter 2.1). Empirical assessment of the different techniques involves considerations of feasibility, consistency, validity and reliability.

Feasibility means that the method must be capable of being carried out in practice and be acceptable to respondents. This last point would appear to be satisfied by the high response rates and even higher levels of complete data that most studies have reported (Froberg and Kane [1989]). Consistency refers to the extent to which the health states used in a study are given a logical ordering within a method. This might be seen as construct validation in the sense that it tests the construct that "better" states of health should be given higher scores but since this has rarely been considered (in fact, inconsistent respondents have generally been excluded from data analysis; see Martin and Elliot [1992] and Torrance *et al* [1992]) it is treated here as a criterion in its own right.

Essentially a measure is valid if it accurately reflects the concept or phenomenon it claims to measure. In establishing the validity of different methods, most studies have examined the extent to which the different methods yield similar results. This test, often referred to in the literature as concurrent validity, has been predicated on the notion that the SG represents the gold standard against which different methods are compared. Indeed, Torrance [1976] advocated the use of the TTO primarily *because* he found it to be correlated with the SG. The above discussion argues that the theoretical justification for according the SG such status is questionable. In this context, concurrent validity is an almost meaningless concept since it tells us nothing

about which method is more valid if the methods yield different results, nor whether both or neither method is valid if the methods yield similar results. However, if one method yields very different results from a number of other methods, then doubt may be cast on its validity.

In the absence of a gold standard, the most rigorous approach to establishing validity is testing construct validity. A construct is a theoretically derived notion of what the method is intended to measure. An understanding of the construct allows the extent to which the method fulfils its predictions to be examined. Construct validity can be assessed by examining (a) the extent to which the valuations from the different methods are correlated with factors for which there is an *a priori* expectation of good correlation (sometimes referred to as convergent validity) and (b) the extent to which the valuations are not correlated with factors for which there is expected to be poor correlation (sometimes referred to as discriminant validity).

The evidence currently available suggests that variation among population subgroups is not explained by the different demographic characteristics of respondents, such as age, sex, or socio-economic status. There is, however, some evidence to suggest that experience of illness may influence respondents' valuations of health states. For example, Sackett and Torrance [1978] reported that home dialysis patients assigned higher utility to kidney dialysis than did the general public. In addition, Rosser and Kind [1978], from comparisons of patients, nurses, physicians and the general public found significant differences between medical patients and physicians and between medical patients and psychiatric patients. The possibility that valuations differ according to illness experience has been noted by Froberg and Kane [1989] who state that "We have seen that patients with a particular condition often assign a higher utility than do patients without the condition".

The reliability of a valuation method can be investigated in two ways; a) Split-test reliability which assesses an individual respondent's consistency when an item is presented more than once and b) Test-retest reliability which assesses the stability of values over short periods of time. Torrance [1976] found the SG and TTO to have

similar split-test correlation coefficients (between 0.80 and 0.90) and these results have been considered to be "acceptable" (see Froberg and Kane [1989]). O'Connor *et al* [1985] reported correlations of 0.80 to 0.87 for a one week retest of SG and TTO respectively, although some respondents may have remembered their initial valuations given the relatively short time interval between test and retest.

Methods

Against this background, four principal criteria were selected as the basis for choosing between competing valuation methods. These concerned the quality of the data elicited from the respondents, rather than the practical aspects of administering the different tasks (such as the burden placed upon respondents and interviewers) and were as follows:

1) Completeness (as a measure of feasibility): the extent to which each method produces a complete data set.
2) Logical Consistency: the extent to which the health states used were given a logical ordering within each method.
3) Construct Validity: the extent to which valuations differ in accordance with prior expectations.
4) Test-retest Reliability: the extent to which respondents' responses are stable *within* each method over a relatively short time interval.

Results

Completeness

Table 2.2.1 shows that at both test and retest TTOP task was the most complete. In the test data, TTOP was significantly more complete than any of the other main methods (all at $p<0.01$). In the retest data, TTOP was more complete than SGNP ($p<0.05$), with no missing values.

Logical Consistency

Given the ordinal structure of the component dimensions in the EuroQol descriptive system, some states are logically ordered with respect to others. For example, it would be expected that 21111 should be given a higher score (to indicate less severity) than 21221 because it is better on at least one dimension and no worse on any of the other dimensions. For some pairwise comparisons, there are no *a priori* expectations of this kind, e.g. between 21111 and 11122. Where an *a priori* expectation holds, it is termed a logical consistency.

With the states used here, 12 such comparisons are possible. A calculation has been made of the number of logically consistent rankings made by each respondent, expressed as a percentage consistency rate. Because the number of possible pairwise comparisons drops substantially when a respondent fails to value a state, the data of those respondents with more than one missing value on the SG or TTO were considered to be unusable in the calculation of consistency rates. In addition, the few respondents who gave the same score to five or all six states on the same method were also excluded from this analysis. The distribution of consistency rates was highly skewed, with the majority of respondents having rates close to 100% and a few respondents having rates below 50%. For this reason, the median was chosen as the appropriate measure of central tendency. Table 2.2.2 shows that the TTO variants have higher strong consistency rates than SG variants both at test and at retest but there are no statistically significant differences between any of the four main methods.

Consistency rates on the VAS (in the region of 95%) were the same across the eight experimental groups, suggesting that differences in consistency rates between main methods were not attributable to a response bias. Consistency rates for each of the main methods when they were done first i.e. immediately after the VAS, showed no statistically significant differences, either at test or retest. Also, there was little difference between test and retest consistency rates since subtracting each respondent's consistency rate at retest from their rate at test yielded a median difference of zero for all methods. With respect to respondent characteristics, it appeared that level of

education and consistency rate were positively related, particularly for the SG variants where those with a minimum education had significantly lower consistency rates ($p<0.05$ on both variants). With respect to possible interviewer effects, a few interviewers had respondents with lower than average consistency rates, but results were not affected when data from these interviewers were removed from the analysis. Similarly, no 'learning effect' was identified when each interviewer's first three interviews they conducted were compared with their remaining interviews.

Valuation Results

Since there were no differences found for any of the methods according to the order of presentation of the task or according to whether the preceding task was a props or a no props variant, Table 2.2.3 shows the valuations for each health state (at test) from the four main methods. The predominant order of states is 21111, 11122, 21221, 21232, 22323, 33333 but SGP produces a 'reversed' order for 21232 and 22323, although the valuations given to these two states are close together for all methods anyway. In general, it appears that the no props variants yield higher values than the props ones and that TTO values are higher than SG ones although, interestingly, TTOP is the only method which gives a negative median score to state 33333. Table 2.2.4 shows the valuations elicited at retest where the predominant order of states is the same as that at test and again 33333 is, on average, considered to be worse than dead on TTOP.

Table 2.2.5 shows the results of a within-respondent comparison of valuations using the Wilcoxon matched-pairs signed-rank test. The results confirm those indicated in Tables 2.2.3 and 2.2.4 and the mapping functions in Chapter 2.1 i.e. that: 1) TTONP values are significantly higher than SGP ones for all states except 33333, 2) SGNP values are higher than TTOP ones for the three most severe states, 3) TTOP values are higher than SGP ones for the three least severe states and lower for 33333, and 4) there are no significant differences between TTOP and SGP valuations. There are fewer significant differences between methods at retest than at test due partly to the smaller number of respondents at retest.

Construct Validity

Construct validity relates to the background characteristics of respondents that are (and are not) expected to account for variance in valuations. The constructs tested here are that those in poor health should give higher valuations than those in good health but that valuations should not differ by any other background characteristic. Table 2.2.6 shows that respondents who were themselves in a dysfunctional health state (i.e. reported being in either level 2 or 3 on a dimension) did give significantly higher scores to some but not all states. Other background characteristics, such as age, gender and employment status, showed no systematic influence on valuations.

Test-retest Reliability

The interval between the first and second interview varied from 6 to 16 weeks (median of ten and a half weeks). At retest respondents were asked "Has anything important happened to you since the last interview a few months/weeks ago?". 29 of the 110 test-retest respondents (26%) reported that they had experienced an important event of whom all but three reported a deterioration in the own or someone else's health. As a group these people reported significantly more impairment of mobility and usual activities than the other respondents (both $p<0.05$), and also reported more pain ($p<0.01$) and anxiety/depression ($p<0.05$). Reflecting this, they also reported more personal experience of illness ($p<0.05$). Since this greater experience of very recent illness may affect the respondents' valuations of health states, those re-interviewed were separated into two groups on the basis of whether or not they reported that they had experienced an important event since the first interview.

Treating the data first as ordinal and then as cardinal, Spearman's rank coefficient and Pearson's r coefficient were calculated. The mean correlations for those without and with important events are shown in Table 2.2.7 which shows that TTOP has the highest correlation coefficients and for those without important events performs significantly better than both SGP and TTONP ($p<0.05$). Table 2.2.8 shows the correlation coefficients for each method separated according to the time interval

between test and retest. TTOP and SGNP have the highest correlations for respondents re-interviewed 'early' i.e. within the median time interval of 73 days. While both the Spearman and Pearson correlations for SGNP fall as the time between test and retest increases, the corresponding values for TTOP remain at high levels. In terms of median differences in scores between test and retest, there are no significant differences between test and retest for any state within any method for those <u>not</u> reporting important events. For those with an important life event, only two differences are significant at the 5% level (both on SGNP), suggesting that important life events have negligible effects on state-by-state valuations.

Discussion

This chapter has looked at the criteria that can be used in order to make a choice between the SG and TTO methods. Since a choice could not be made between SG and TTO on theoretical grounds or on the basis previous empirical work, the study reported in this chapter was designed to allow a direct comparison between two variants of each of these two methods. On the grounds of completeness, there is evidence in favour of TTOP since it was significantly the most complete of the main methods at test and had <u>no</u> missing data at retest. No clear 'winner' emerged from a test of logical consistency but TTOP would be given a slight preference. This issue has rarely been considered in valuation studies and would be unimportant if all methods generated similar (high) levels of consistency. However, if consistency rates are low then doubt is cast on the feasibility of valuing health states in this way. Our experience here is that there is a 'threshold' level of consistency of somewhere in the region of 85% for SG and 90% for TTO and I consider these rates to be acceptable.

The construct validity of the methods was assessed according to the extent to which valuations differed by the background characteristics that previous literature had shown to be important (and unimportant) determinants of valuations. It was hypothesised that valuations would not differ according to the age, gender and employment status of the respondent but that higher valuations would be elicited from respondents with experience of illness. All methods yielded valuations which

supported the former construct whilst tentative support was lent to the latter construct. Of course, if the construct is not supported it does not necessarily invalidate the method as it may be that the construct itself is misspecified. More research is needed before the constructs hypothesised in this chapter can be considered to be absolute standards.

The validity of health state valuations may also be assessed by considering the extent to which the valuations elicited by these methods are valid representations of individual preferences. One way to test this would be to examine the robustness or otherwise of the valuations. Less confidence would be placed in valuations that are sensitive to seemingly irrelevant changes in problem structure or question format, for example. It is encouraging that valuations from all methods appeared to be unaffected by the order of presentation i.e. valuations were no different whether that task was administered first or second, or whether it was preceded by a props or a no props variant. This finding contradicts that of Llewellyn-Thomas *et al* [1984] who found the existence of an anchoring effect when riskless methods such as the TTO were preceded by lottery questions, as in the SG.

It is also encouraging that all the methods produce a similar ordinal ranking of health states, which suggests that they all allow respondents to differentiate between states of differing severity. However, differences in cardinal values are observed and possible reasons for this have been discussed in Chapter 2.1. Whatever the reasons, it appears that valuations from TTOP are the most central in that they are generally higher than SGP ones and lower than SGNP and TTONP ones. The exception appears to be state 33333 which has a lower score on TTOP than on any of the other methods. Indeed, TTOP is the only method which results in a negative median value for this state. In other words, at least half the people valuing this state on this method consider it to be worse than death. This may be because the TTO method forces respondents to think more closely about the consequences of being in an extremely dysfunctional state for 10 years without any change. In the SG, the duration element may be given less prominence by respondents. There is evidence to suggest that more states are regarded as worse than death the longer they last (see Chapter 4). In this respect,

valuations to TTOP may more accurately represent individual preferences for states that last 10 years without any change.

Before definite conclusions can be reached on the issue of which method most accurately represents individual preferences, it is important to gain a better understanding of the reasons why valuations differ (both within and between sub-groups). This issue will be discussed more fully in Chapter 6 but is complicated by the fact that health state valuations from choice-based methods are likely to be a function of both the severity of the health state and the context of the choice. For example, responses to standard gamble questions are likely to be influenced by attitudes to risk; and responses to time trade-off questions are likely to be influenced by life expectancy and time preference.

Test-retest reliability gave a more definitive answer in that TTOP valuations showed the most stability across time, performing significantly better than both TTONP and SGP and similarly to SGNP. It has been conventional in this field to assess test-retest reliability by calculating the correlation coefficient between the first and the second sets of scores obtained from each respondent. The coefficient of 0.81 for TTOP compares well with those from other studies (see Churchill *et al* [1984], [1987] and O'Connor *et al* [1985], particularly as the time between test and retest was longer at a median of ten and a half weeks.

Bland and Altman [1986] have argued that use of correlation is misleading since it measures only the strength of a relation between two variables and not the agreement between them. Instead, they suggest plotting the differences in scores between two methods (or in this case the difference between test and retest scores) against their mean. By calculating 'limits of agreement' between the two sets of scores (defined by Bland and Altman as the mean plus and minus two standard deviations) and the confidence intervals associated with them, the degree of agreement between the sets of scores can be summarised. However, given that six states were valued using four methods, there would be twenty-four graphical representations of the differences between test and retest scores. This would mean that unless one method produced the

greatest agreement between test and retest for all six states (which is not the case), then it would be extremely difficult to determine the overall performance of each method. For this reason we feel that the correlation coefficient provides the best summary statistic available.

On the basis of the results reported here, TTOP has been chosen as the valuation method to be used in the Main Study. This choice has been made in the context of a study conducted with a random sample from the British general population, using a particular descriptive tool for health status, and with specially designed boards and protocols. Although the method performed well, it was clear from interviewers' comments that improvements could be made to ease the handling of scripts, cards and boards in an often confined space. This would be particularly important in a clinical setting although it is encouraging that TTO has been found to be relatively easy in practice (see Torrance [1987]) and has been used fairly widely to generate valuations for health states (see Laupacis *et al* [1992] and Singer *et al* [1991]). It is not known whether the choice of the EuroQol descriptive system affected the outcome, although it is unlikely to have had a differential effect on the SG and TTO methods.

It is recognised that no clear cut 'winner' emerged from this study and, in particular, there is little to choose between TTOP and SGNP. The need to select one method, however, has pushed the balance in favour of TTOP as the method which performed significantly better on completeness, marginally better on logical consistency, and significantly better than SGNP and TTONP on test-retest reliability. In addition, the possibility that questionnaire framing may bias responses to the no props variants casts some doubt on the validity of the valuations from these methods.

CHAPTER 3: GENERATING A SET OF VALUATIONS

CHAPTER 3.1: THE TTO RESULTS FROM THE MAIN STUDY

Study Design

Choice of health states

For the EuroQol to be used in evaluating the health benefits associated with different health care interventions, it is important to derive a single index value for each of the 243 health states it generates. Two pilot studies conducted prior to the Main Study suggested that no one respondent can be expected to value more than about 13 states using TTO in any one interview but this number was deemed to be too small to interpolate valuations for all possible EuroQol states from. Therefore, a larger set of 43 states was chosen in total and each respondent was asked to value a subset of these.

In choosing the states both for use in the study itself and for each respondent, the most important consideration was that they should be widely spread over the valuation space so as to include as many combinations of levels across the five dimensions as possible. This was subject to the constraint that the states were likely to be considered plausible by respondents. Therefore, level 1 on usual activities (no problems) was not combined with level 3 on mobility (confined to bed) or with level 3 on self-care (unable to wash or dress self). Figure 3.1.1 shows the set of states chosen for direct valuation and how a subset of these were chosen for each respondent.

Structure of the interview

Each respondent was first asked to describe their own health using the EuroQol descriptive system. They were then asked to rank a predetermined set of 15 health states (the 13 to be used in the TTO plus 11111 and "Immediate Death"), which were printed on cards, in order from best to worst. It was explained that each state was to be regarded as lasting for 10 years without change, followed by death. The respondent was then asked to indicate where on a vertical VAS with endpoints of 100 (best imaginable health state) and 0 (worst imaginable health state) they would rate each of the states.

The 13 states were then valued by the TTO method using a specially-designed double-sided board similar to that described in Chapter 2.1 and shown in the appendix. For states that were regarded by the respondent as better than dead, respondents were led by a process of "bracketing" to select a length of time in the 11111 state that they regarded as equivalent to 10 years in the target state and were given an opportunity to refuse to trade-off any length of life in order to improve its quality. In the case of states worse than dead, the choice was between dying immediately and spending a length of time (10-x) in the target state followed by x years in the 11111 state. At the end of the interview, personal background data were collected from each respondent.

Retest interview

In order to test the reliability of the TTO valuations, a sub-sample of 221 respondents that were representative of the full sample in terms of sex, age, and qualifications were taken through exactly the same interview by the same interviewer about 10 weeks after the original interview.

The Sample

In determining the size of the sample, there was the need for enough observations to be obtained so as to detect differences between the valuations given to different states and to be able to detect differences in valuations between different sub-groups of the population (e.g. by age, or social class, or geographical location). Although there is little evidence in the literature regarding what size difference is required to be considered meaningful (see O'Brien and Drummond [1994]), it was decided that a .05 difference between health states and between different sub-groups is likely to be considered important in many contexts. A sample size of 3235 enabled such a difference to be detected between health states and between four equally-sized subgroups at the .05 level of significance with 80% power. This required the selection of 6080 addresses; thus allowing for a response rate of 53%. The sample was drawn up by Social and Community Planning Research (SCPR) using the postcode address file. The main fieldwork was carried out by 92 trained interviewers between August and December 1993.

Study population and exclusions

Of the 6080 addresses selected for sampling, 706 (12%) were found to be 'out of scope', being non-residential, empty/derelict, untraceable, or not yet built. Of the remaining 5324 addresses, 3395 interviews were achieved, giving a response rate of 64% on in-scope addresses. The main reasons for unsuccessful interviews were a refusal by the selected person. Table 3.1.1 shows that the sample had broadly similar characteristics in terms of age, sex, marital status. educational attainment and social class as the general population. Table 3.1.1 also shows the number (and background characteristics) of respondents excluded from subsequent data analysis. Because the criteria for excluding respondents were as stringent as possible, in total only 58 (1.3%) of respondents were excluded: 42 had insufficient data for further analysis; 7 had rated all states as worse than death; and 9 did not

understand the TTO task. It can be seen that excluded respondents were more likely to be aged 60 and over, to have no qualifications, to be in social classes III-V and to report problems on the EuroQol dimensions. However, given such a small number of exclusions, the 3337 respondents remaining in the data set were still broadly representative of the general population.

Valuation results

Distribution of scores

Table 3.1.2 shows the transformed mean and median scores for all 43 states. Inspection of the range of health state values suggests that respondents were more prepared to sacrifice life expectancy for states that include "extreme problems" with any of the dimensions. Level 2 (which involves "some problems") on the dimensions appears to be much more tolerable. For example, state 22222 has a median valuation that is 0.13 higher than 11113 and 0.25 higher than 11131. This results in most states that include level 3 on two or more dimensions having values that imply they are, on average, perceived to be worse than death. In total, 17 states have a negative mean score and 13 states have a negative median score (a further 4 had median values of 0.0). Kolmogorov-Smirnov tests indicated that the distribution of scores for each state was non-normal: the distributions were generally negatively-skewed for less severe states (indicated by higher median than mean values for such states) and positively-skewed for more severe states (as evidenced by higher mean values for such states).

The effect of background characteristics

Before addressing this issue, it was determined that the valuations were not susceptible to interviewer bias nor to regional effects. OLS regression analysis was used to assess the impact of a number of respondent characteristics on health

state valuations. The dependent variable was taken to be the TTO valuation and the independent variables were the different background characteristics (see Table 3.1.3 for a description of these variables). Most of the variables are categorical except for age which is a continuous variable and age-squared which of the various transformations of age tested was found to be the most significant. To allow for the possibility that the impact of one or more of these variables may not be uniform across the entire range of EuroQol states, the regression was performed separately on the 'mild' (which included the set of five 'very mild' states), 'moderate' and 'severe' states, as defined in Figure 3.1.1.

The results are shown in Table 3.1.4. That the adjusted-R^2s are so low is not in itself a cause for much concern since the object of this analysis is to assess the relative effect of different respondent characteristics on valuations rather than to find the model(s) which explains all the variance in valuations. Given the large number of observations in each regression, a particular variable is considered to be significant if the (absolute) t-statistic associated with it is greater than 3.29 (which corresponds to a probability value of 0.001).

The results suggest that TTO valuations are primarily affected by the age and sex of the respondent. Figure 3.1.2 shows the effect of age on the valuations given by men and women, respectively, when all other dummies take a value of zero. They suggest that TTO valuations increase slowly from the age of 18 to about 40, then begin to fall slowly from about 40 to 60 and then fall sharply in later years. Although this pattern is observed for all three sets of states, it is more marked for moderate and severe states than for the mild states. The effect of gender is also more pronounced for more severe states: for the set of mild states, women give valuations that are, on average, 0.03 lower than those given by men but the difference increases to 0.06 for moderate states and to 0.07 for severe states.

In addition, the marital status of the respondent appears to be a statistically significant explanatory variable for the set of mild states (where, on average, the valuations of single people are 0.006 higher than the valuations of married people and 0.005 lower than those who are separated, divorced or widowed) and for the set of moderate states (where single people have valuations that are 0.008 lower than married people). However, it can be seen that the value of these coefficients is very small suggesting that, although statistically significant in the regression equation, the effect of marital status is negligible and unlikely to be meaningful in any practical sense.

Quality versus quantity?

The question of whether TTO valuations differ by sub-group is essentially about whether some people are more or less prepared to sacrifice life expectancy in order to avoid poor health than other people. In this context, there is another important question; namely, are some people more or less willing to sacrifice any life expectancy in order to avoid poor health than others? This draws a distinction between those willing to trade quantity (in terms of life expectancy) for quality (in terms of improvements in health), irrespective of the rate of exchange, and those unwilling to "play the game". In other words, there exists a qualitative difference between an implied health state value of 1.00 and any other value.

46% of respondents were willing to sacrifice life expectancy to avoid all of the dysfunctional states they were presented with, and thus had no health state values of 1.00. A further 29% were willing to sacrifice life expectancy for all but one or two of the states. In such cases, the unwillingness to trade-off time was almost exclusively associated with one or both of the very mild states. In all, 95% of respondents were prepared to sacrifice life expectancy for 6 or more states. The 25% of respondents who were unwilling to trade off time for three or more states were older and less educated than the remainder of respondents; 33.7% were aged

60 or over compared to 28.4% and 41.8% had no qualifications compared to 34.7% in the group of respondents more willing to sacrifice life expectancy. Interestingly, the 5% of respondents who were unwilling to sacrifice any life expectancy in order to avoid more than half of the states they valued were no older than the remainder of respondents. Instead, such respondents were found to be less educated (45.6% had no qualifications compared with 33.9% of the other respondents).

Test-retest reliability

The 221 respondents in the retest were representative of those in the test in all respects except educational level, where 28.6% of retest respondents had no qualifications compared with 37.0% of respondents not in the retest (Chi=6.26, d.f.=1, $p<0.05$). For the purposes of group analysis, 4 respondents were excluded from the retest data set: 1 previously excluded from the test data set; 1 with all states missing at retest; 1 with all states rated as worse than dead; and 1 with the same score given to all states. At test, respondents taking part in the retest gave a significantly higher TTO score to state 33323 than respondents who did not go on to do the retest ($p<0.01$), but this was the only significant difference in the valuations given by the two groups of respondents. The results of Wilcoxon matched-pairs signed-rank tests showed that no health state valuation at retest was significantly different from its corresponding value at test. However, Figure 3.1.3, which graphically represents the differences in median scores between test and retest, shows that for 2 (3) states the difference between the median at test is more than 0.20 higher (lower) than the median at retest.

For comparisons on an individual-by-individual basis an intra-class correlation coefficient (ICC) was calculated for each respondent for each of the valuation methods. This statistic is calculated using the following formula;

$$ICC = (A^2+B^2-C^2) / (A^2+B^2+D^2-C^2)$$

where A is the SD of the difference between each score at test and the mean score at test

B is the SD of the difference between each score at retest and the mean score at retest

C is the SD of the difference between each score at test and each score at retest

D is the mean difference between each score at test and each score at retest

The closer the ICC is to 1, the greater the reliability. Figure 3.1.4 shows the distribution of ICCs. The majority of respondents had an ICC that was close to 1 and only 24 (10.9%) had an ICC that was less than or equal to 0.5. The mean ICC was 0.73 (S.D.=0.22) and the median was 0.79 (IQR=0.64-0.88). ICCs appeared to be negatively related to educational attainment; those with a degree or equivalent had higher ICCs as a group than those with no qualifications at all ($p<0.05$).

Discussion

The group valuations elicited for the 43 health states suggest that members of the general public can distinguish between states of health that involve different degrees of severity. However, the measures of dispersion (SDs and IQRs) were much higher than expected which casts doubt on the assertion made by Torrance [1986] that "the mean utility value for a health state can be made as precise as desired by increasing the group size". Rather than reflecting the degree of consensus about the value that should be attached to a particular health state, it is possible that the large SDs and IQRs reflect the difficulties respondents encountered in imagining themselves being in the health states so described. That

the variance around the central tendency values increases as the severity of the health state increases, lends some support to this hypothesis.

However, the interpretation of measures of dispersion does not tell the whole story, because it is quite plausible that respondents rank adjacent states in the same way, but some do so using high values, while others do so using low values. Analysing pairwise relationships between states (using the Wilcoxon matched-pairs signed-rank test), revealed that there were no more than 4 states adjacent to any particular state which were not significantly different from it at the 1% level. Thus, it appears that the large SDs and IQRs obtained in this study, particularly for the more severe health states, are more likely a reflection of the fact that different people have very different views about the same health state, rather than an indication of respondent confusion.

For TTO valuations to be interpreted on an interval scale requires each year of life to be valued equally. However, if people discount future years of life because of a positive rate of time preference (i.e. because they give greater value to years of life in the near future than to those in the distant future), then it is no longer valid to treat TTO valuations in this way. Moreover, if people are not prepared to trade-off a constant proportion of their remaining life expectancy in order to avoid a dysfunctional health state, then valuations elicited for states lasting ten years cannot be assumed to hold for states lasting for longer or shorter durations irrespective of the impact of duration. These issues are discussed more fully in Chapters 4.1 and 4.2.

It is unclear how generalisable these results are since it is likely that are in part a function of the duration of the states. Since respondents were told to imagine that each state would last for ten years without any change, it is likely that some felt they could not tolerate extreme dysfunction (particularly pain) for this long. Whilst the finding that some states were considered worse than death is not unique

(they have appeared in several countries for several valuation methods; for example, see Rosser and Kind [1978] and Read *et al* [1984]), there is evidence to suggest that fewer states would be regarded as worse than death were they to last for less time and this issue is addressed in Chapter 4.3.

There is also the issue of whether the order of presentation for states rated as worse than death may have had an effect on valuations: respondents may value a scenario in which a bad state is followed by a good state (as in this study) differently from one in which a good state is followed by a bad state (as suggested by Torrance [1986]), even though the time spent in each of the states may be identical. This is an empirical question which needs addressing. In addition, there is the question of how to interpret scores for states worse than dead. As discussed in Chapter 2.1, given the standard health preference scale, states preferred to death have an upper bound of one but there is no comparable lower bound for states rated worse than death. The asymmetry results from the TTO (as well as for the SG) producing an interval scale for positive scores and a ratio scale for negative scores. It seems reasonable to treat positive and negative scores in the same way i.e. to convert the ratio scores into interval ones, thus setting a lower bound of -1, and this adjustment finds support in the literature.

One of the most important findings is the effect that the age and, to a lesser extent, the gender of the respondent has on health state valuations. Other background variables, such as social class and education, were found to be insignificant, and others (such as marital status), whilst statistically significant, are unlikely to be meaningful in programme evaluation. The importance of age and sex contradicts the findings of other studies: in their review of the literature to 1988, Froberg and Kane [1989, p586) find "little compelling evidence of population differences due to demographic characteristics". However, most of the previous studies of health state preferences have contained small numbers of respondents, and, as Froberg

and Kane readily admit [1989, p586] "low statistical power may be obscuring differences".

The valuations of 'middle-aged' respondents appear to be higher than those of younger respondents, whilst older respondents have much lower valuations than those in the other two age 'groups'. This may lend support to the notion that the middle-aged have the lowest rates of time preference and thus place relatively more weight on years in the future (i.e. the ones they are being asked to sacrifice) than younger or older respondents. However, the fact that the effect of age is not uniform across all states, being more pronounced for moderate and severe states than for mild ones, may suggest that the effect of time preference for health is <u>not</u> independent of the severity of the health state. Again, see Chapter 4 for a fuller discussion of this issue.

It may be that the much lower valuations of the older respondents in this study are an artefact of the TTO method. For states that were rated as better than dead, respondents were asked to imagine that each state would last for 10 years without any change, after which they would die. If they did not believe that they actually had 10 years life expectancy, they might willingly give up these "excess" life years, thereby depressing the apparent value attached to the health states. However, the effect of age appears to be more pronounced for the more severe states (which were much more likely to be rated as worse than dead) than for the less severe ones. It is unclear how and why an argument of this kind would apply with greater force to the worse than dead scenario than to the better than dead one.

An alternative explanation is that, as people's life expectancy shortens, they see less reason to tolerate suffering during their remaining years. Conventional wisdom suggests that people become <u>more</u> tolerant of poor health as they get older, either through adapting to a general deterioration in health or through a lowering of expectations, and there is some empirical evidence to support this hypothesis (see

Sackett and Torrance [1978]). However, it is entirely plausible that somebody who has limited life expectancy and is possibly in a poor health state, may be prepared to sacrifice a great deal (either life expectancy for states rated better than dead or time in full health for states rated worse than dead) in order to avoid severe health states. In a study of cancer and renal patients with limited life expectancy, Shiell, King and Briggs [1993] found that TTO results were polarised; some would not trade off any life years, while others would trade off almost everything to have their final years as healthy ones. The older respondents in this general population study may have held similar views about the (severe) states as this latter group.

In addition, older respondents may be more conscious of the burden that serious chronic illness can place on their family or close friends, particularly if they have experienced the suffering of someone close to them. This might explain why the valuations of older respondents were closer to those of other respondents for states they considered "tolerable" (both for themselves and for those close to them) yet much lower for states they considered would be "intolerable" for themselves and their family. It may also go some way towards explaining why women had lower valuations than men for the more severe states: women may be more concerned about the burden they would be to others than men are, particularly as they may be likely to have experience of caring for someone with serious illness.

Before definite conclusions can be reached about the effect of certain background characteristics it is important to gain a better understanding of the reasons <u>why</u> valuations differ (both within and between sub-groups) but, as noted in Chapter 2.2, this issue is complicated by the fact that health state valuations from choice-based methods are likely to be a function of both the severity of the health state <u>and</u> the context of the choice. Therefore, some of the differences in health state

valuations reported in this chapter may be the result of different perceptions of time rather than differences perceptions of severity of illness *per se*.

The premise of the discussion so far has been that differences in valuations according to the age and, to a lesser extent, gender of the respondent are real differences. However, it could be argued that this relationship is a spurious one; that the large number of variables being assessed and the large sample size, by chance, account for the results. Whilst this possibility cannot be completely ruled out, the fact that in the three regressions (one for each set of mild, moderate and severe states) the effects of age and sex are systematic (though not constant) suggests that genuine effects are being picked up. In addition, the use of regression analysis should isolate the effects the age and gender, and thus reduce the possibility that they are acting as proxies for other (more important) explanatory variables.

Although almost half of the respondents were prepared to sacrifice life expectancy in order to avoid all of the dysfunctional states they were asked to consider, one-quarter were unwilling to sacrifice even a couple of weeks at the end of 10 years for 3 or more states. Such preference may be further evidence of the reference point effect discussed in Chapter 2.1. If perceived losses (in terms of life expectancy) are weighted more heavily than perceived gains (in terms of HRQoL), the effect will be to elicit a higher health state valuation than might otherwise be the case since respondents are asked to imagine that they are already in the poor health state. At the limit, the perceived loss is so great as to make any change (from 10 years in poor health) undesirable.

It seems reasonable to suppose that the reference point effect would be more prevalent the more unclear respondents are about the choices they are being asked to make; an "if in doubt, stick with what you've got" hypothesis. The hypothesis might be, therefore, that less educated respondents are more likely to suffer from

reference point effects. The fact that less educated respondents were more likely to be unwilling to sacrifice any life expectancy, even for some moderate and severe states, lends support to this hypothesis. Certainly, the possibility of a reference point effect should not be overlooked and, as already discussed, may go some way towards explaining the observed differences between valuation methods that start with different "endowments".

The results from the retest were encouraging. At the aggregate level, 32 of the 43 states had a median at retest that was within 0.1 of the median at test and there were no significant differences in the valuations of any of the states between test and retest. This finding is consistent with other studies which found group values to be remarkably stable regardless of the make-up of the group [see Boyd *et al* [1982] and Wolfson *et al* [1982]). At the individual level, only 1 in 10 of the respondents had an intra-class correlation coefficient that was below 0.5. The mean ICC was 0.73 which is considered acceptable and compares well with the results of previous studies (see Chapter 2.2).

Of course, the stability of the valuations of the general public does not necessarily imply that the valuations of all groups will be stable. For example, this study did not include test-retest measurements taken from patients before and after therapy, whose valuations of the same state might be expected to differ. Christensen-Szalanski [1984] found that women's preferences for anaesthesia during childbirth were labile; not surprisingly, perhaps, preferences for anaesthesia were more positive during labour than they were one month before or after labour. However, Llewellyn-Thomas *et al* [1993] found that patients' TTO valuations of hypothetical health states encountered during radiation therapy for laryngeal cancer remained stable when those states were experienced at a later time.

Finally, with respect to the logistics of the study itself, representativeness was achieved and the data were near complete and highly consistent, thus refuting the

claim of Froberg and Kane [1989, p681] that the TTO "probably loses its advantage in large-scale studies due to its complexity and the resulting confusion and nonresponse". Despite the many unanswered questions, the results show that eliciting the preferences of the general public is feasible and indicate possible directions for future research.

CHAPTER 3.2: MODELLING THE TTO DATA

Introduction

This chapter reports on the methodology adopted to allow valuations for all 243 EuroQol states (referred to as the "tariff") to be interpolated from direct valuations on the 43 states used in the Main Study. To enable modelling of the TTO data at the individual level, only those respondents with complete valuations data have been included in the analysis. There are 2997 such respondents. Excluding those respondents with incomplete data did not compromise the representativeness of the sample.

Methods

The modelling in this chapter uses a generalised least-squares regression technique in which the functional form is additive. The dependent variable is defined as 1-S where S is the value given to a particular health state. Besides the intercept, the specification of the remaining independent variables derive from the ordinal nature of the EuroQol descriptive system. In total, three sets of dummy variables were created:

1. Two dummy variables for each dimension; one to represent the (assumed equal) move between levels and one to represent the move from level 2 to level 3

(this allows the effect of the move from level 1 to level 2 to be different from the effect of the move from level 2 to level 3).

2. Dummies to allow for possible (first order) interactions between dimensions.

3. Dummies to count the number of times a health state contains dimension(s) which are at level 1 or at level 3.

Figure 3.2.1 shows the independent variables used in the modelling. Notice one further dummy (N3) which represents whether any of the dimensions is at level 3, of which more anon. Because the objective of this chapter was to estimate one preference-based EuroQol tariff for the whole community, respondent characteristics such as age, sex and illness experience were not entered in to the model. Differences according to these characteristics may, of course, be important in some contexts and analysis along these lines is reported in Chapter 3.3.

The approach adopted to model estimation followed the specific to general formulation in which simple models are initially estimated and new variables added if necessary. This was deemed the most appropriate methodology for this type of data since the alternative, the general to specific approach is more suited to time series data in which any known collinearity between regressors can be better accounted for (problems of multi-collinearity are discussed further below).

It was decided that analysis should take place on individual-level rather than aggregate-level data since it makes the maximum use of the available data. In addition, the results of aggregate-level analysis are likely to be uninformative in that it is possible to find different models which fit the data equally well, with no objective way of choosing between them. Analysis at the individual level is complicated by the fact that each respondent valued 12 EuroQol states and thus it is reasonable to assume that these 12 scores are related to one another. This means that if a respondent gives one valuation that is lower than the population mean, then they are more likely to give a value lower than the population mean to

the other states that they value. This means that the variance of the error term is likely to be partly determined by the individuals who value the health states and is therefore unlikely to be constant. This violates one of the key assumption underlying OLS regression and thus makes this estimation procedure inefficient for this data.

The type of generalised least-squares (GLS) model that addresses this issue is known as the random effects (RE) model in which there is an overall intercept and an error term with two components; $e_{it} + u_i$. The e_{it} is the traditional error term unique to each observation. The u_i is an error term representing the extent to which the intercept of the ith respondent differs from the overall intercept. This model assumes that the "individual specific" error term is normally independently distributed which, given the size of the sample, seems a valid assumption to make. Using the RE specification will reduce the possibility of drawing erroneous conclusions; for example, from an OLS estimation it may be concluded that a particular respondent characteristic is an important determinant of the value attached to a health state but this may simply be picking up an effect that will be nested within the RE model.

As a stringent test of the robustness of the models, each model has been estimated on a sub-sample of respondents (i.e. an internal sample) and a comparison made between the predicted values from this sub-sample with the actual values of the remaining respondents (i.e. an external sample). This method has the advantage of providing unbiased estimates of predictive ability, and avoids over-optimistic results that occur when the same data is used both to estimate and predict (see Copas [1983]). In this chapter, the internal sample was a randomly selected sample of two-thirds of respondents and the external sample constituted the remaining one-third of respondents.

The modelling has been carried out using the LIMDEP statistical package (see Greene [1992]). When estimating the RE model, LIMDEP produces the OLS equation by default. It automatically performs a Lagrange Multiplier (LM) test which is appropriate for large data sets like this. The LM test assess whether the unrestricted model (i.e. the RE one) represents an improvement on the restricted model (i.e. the OLS one). If the LM value is significant ($p<0.05$) then the RE model represents an improvement on the OLS one. The R-squared produced from the RE model may not be any greater than that produced by the OLS model, and may in some cases be lower. It must be remembered, however, that OLS is the only estimation procedure that attempts to minimise the residual sum of squares; all other GLS models (of which RE is one) have a different objective function. Therefore, the R-squareds from the different models are not strictly comparable. The models were tested for misspecification in two ways: a Ramsey RESET test, and a test for general heteroskedasticity (see Chapter 2.1 for details).

With regard to making a choice between different ways of representing the relationship between the valuations of EuroQol health states and the different dimensions and levels, the model that is ultimately chosen must predict a higher score for one state, A, than another, B, if A is logically better than B on at least one dimension and no worse on any other dimension. In choosing between the many models that satisfy this consistency condition, the one that best explains the differences in the valuations given to those states on which there is direct data was chosen. For models with comparable goodness-of-fit statistics, the ultimate choice was made according to parsimony i.e. the simplest model (both in terms of the number of independent variables and the ability to explain them) was chosen. The results presented below are from the "best" model according to these criteria.

Results

After testing many different models, one that fits the data well (in terms of goodness-of-fit statistics) and that is readily interpretable is a main effects model, in which each of the 5 dimensions is independent of others. None of the models which allowed for interactions between different dimensions improved the model significantly and many introduced inconsistencies into the estimated values. The model does, however, contain one further variable; an intercept dummy for whether any of the dimensions is at level 3. Without this additional dummy, which can be interpreted as reflecting the much greater disutility associated with "Extreme problems", the residuals are systematically related to the predicted values in that the model underestimates the values of less severe states and overestimates the values of more severe ones.

Thus, the regression equation is as follows:

$$Y = a + ß_1MO + ß_2SC + ß_3UA + ß_4PD + ß_5AD + ß_6M2 + ß_7S2 + ß_8U2 + ß_9P2 + ß_{10}A2 + ß_{11}N3$$

i.e. TTO scores are explained by 12 independent variables: two variables for each dimension (one to represent the move from level 1 to level 2 and one to represent the move from level 2 to level 3), a term which picks up whether any dimension is at level 3 and an intercept (the interpretation of which is discussed below).

The coefficients on these variables for the full and internal samples are shown in Table 3.2.1. The R^2 of 0.46 (in both cases) is very high given the type of (cross-sectional) data analysed here and the results of the LM test indicate the RE specification to be a substantial improvement over the OLS model. In addition, the remarkable similarity between the parameter estimates for the whole sample and those for the internal sample suggests that the model is robust.

However, this model (and all other models) failed the RESET test and suffered from general heteroscedasticity. That the model suffers from problems of omitted variables and/or incorrect functional form is not surprising given that the power of the RESET test increases as the sample size increases. Thus with 2997x12=35964 observations, any model with relatively few independent variables is likely to be misspecified. The problems associated with heteroscedasticity are also difficult to overcome since the conventional means of dealing with them (e.g. transformation of one or more independent variables) are not feasible given the (categorical) nature of the independent variables. Correction methods such as White's correction have been used in other contexts (for example, labour supply) but will only reduce the standard errors on the parameter estimates, not change the estimates themselves. Since the purpose of this paper is to generate point estimates of valuations, and because heteroscedasticity here will result in inefficient rather than biased parameter estimates, it is less of a problem than in contexts such as estimating labour supply functions.

Because the analysis concerns cross-sectional data, all variables from the main effects model have been left in the final equations, even those that might be considered "insignificant" (i.e. have a t-statistic whose absolute value is less than 1.96). This is to avoid any pre-test type problems where "insignificant" variables may become "significant" if sampling were to be repeated. In addition, dropping variables whose absolute t-statistic is less than 1 is likely to result in the mean square errors being higher than they should be. Moreover, given that the regressors in this modelling are collinear, the significance of parameter estimates will vary according to the other independent variables in the equation. Since not enough is known about the nature of the functional form to address problems of multi-collinearity with any degree of confidence, it is considered appropriate to include even "insignificant" main effects variables.

In computing the tariff from the model output, there is an issue relating to how the intercept, *a*, is interpreted. The strict statistical interpretation of *a* is that it represents the estimated value for (one minus) full health (i.e. when all dummies take a value of zero we have the estimated value for 11111). Thus, all estimated values should be rescaled by dividing them by 1-*a*. Alternatively, given that by definition the value of 11111 is 1, we could interpret the intercept as representing any move away from full health. Thus, *a* could represent a discontinuity in the model between level 1 and level 2 in much the same way as the 'N3' term represents a discontinuity between level 2 and level 3. In other words, we could interpret the intercept as picking up whether any dimension is at level 2, just as 'N3' picks up whether any dimension is at level 3. When predicted and actual values are compared, the algorithm in which *a* is treated in this way performs much better than when all estimated values are adjusted by 1-*a*.

Table 3.2.1 shows that the constant is highly significant suggesting that any move away from full health is associated with a substantial loss of utility. For the full sample, it can be seen that the largest decrement for a move from level 1 to level 2 is associated with pain or discomfort, some four times greater than that for the corresponding move on the usual activities dimension. Pain or discomfort continues to dominate the weighting for level 3, although mobility level 3 (confined to bed) is given a somewhat similar decrement. For the mobility, pain or discomfort, and anxiety or depression dimensions, the move from level 2 to level 3 is seen to involve a much greater decrement than the move from level 1 to level 2.

The actual (mean) and predicted values for the 42 states directly valued in the study, together with the differences between them, are given in Table 3.2.2. For only three states (21312, 23313 and 13332) does the difference between the mean and predicted value exceed 0.1 and the mean absolute difference (of 0.039) is considered acceptable. Table 3.2.3 compares the predicted values generated from the internal two-thirds of respondents with the actual (mean) values of the

remaining one-third of 'external' respondents. It can be seen that the predictive power of the model remains high; only 5 states have a predicted value that is more than 0.1 different from the actual value and the mean absolute difference is again below 0.05.

Discussion

The statistical analysis used in this chapter to interpolate valuations for all 243 EuroQol health states from direct observations on a subset of 43 states is based on regression analysis in which the dependent variable is (one minus) the score given to the health states. All independent variables are dummies and derive from the ordinal nature of the EuroQol descriptive system. The functional form estimated is a linear additive one which seems a valid approach given the assumption that valuations elicited from the TTO method for states rated as better than dead exhibit interval scale properties (i.e. the difference between 0.2 and 0.4 is the same as the difference between 0.6 and 0.8). Besides, estimating and interpreting different functional forms would be difficult given the (categorical) nature of the independent variables.

Analysis is of data at the individual level to make full use of the data available and is based on a form of GLS known as the random effects model. This specification accounts for the fact that groups of observations come from one individual. An alternative approach would have been a fixed effects (FE) model in which a dummy variable would be created for each respondent. However, models based on RE rather than on FE were deemed more appropriate for this data set for a number of reasons. Firstly, FE models, which produce results that are conditional on the units in the data set, are only reasonable if the data exhaust the population. If the data are a sample of a larger population (as is the case here), and if we wish to draw inferences regarding other members of that population (as is also the case here) then "the fixed effects model is no longer reasonable; in this context, use of

the random effects model has the advantage that it saves a lot of degrees of freedom" (Kennedy, [1992]). Secondly, given that we cannot fully account for how and why valuations differ across individuals, it is reasonable to treat this type of ignorance in a fashion similar to the general ignorance represented by the error term. Finally, there is the practical problem of estimating and interpreting 2997 coefficients.

Given that each respondent in this dataset valued a number of different health states, and thus the value a respondent gives to one state is likely to be related to the value they give to other states, it was expected that the RE models would represent an improvement over the OLS ones. The results of the LM tests confirmed this, suggesting that if a respondent values one state above (or below) the population mean, then they are more likely to value other states above (or below) the mean. It should be noted, here, that the beta coefficients (though not the standard errors) from the OLS equations are similar to those estimated by the RE models, and thus it appears that an OLS specification produces unbiased yet inefficient estimates.

Given a dataset of the kind analysed here there is, of course, a degree of uncertainty about the precise value that should be attached to any particular health state. One way of expressing this uncertainty is to calculate confidence intervals around the predicted values. The 95% confidence intervals were approximately 0.75 for all states, irrespective of their severity. Thus, the confidence intervals (even the 50% one, which is 0.26) are undoubtedly large, indicating (as the standard deviations around the mean values presented in Chapter 3.1 indicate) that different people attach very different valuations to the same health state.

However, confidence intervals do not tell the whole story. They only allow us to make inferences about the degree of consensus regarding which value to attach to each health state but say nothing about the degree of consensus regarding the

(cardinal or ordinal) differences between states. As noted in Chapter 3.1, most repondents rank adjacent states in the same way; it is just that some do so using high values, while others do so using low values. That the RE specification is much more efficient than the OLS one confirms this. Therefore, while large confidence intervals around the point estimates suggest that it is difficult to say with any degree of confidence what value any given individual will attach to any given health state, we can be reasonably confident that the individual will be in broad agreement with the (cardinal and ordinal) differences between states as that implied by the tariff values.

Besides the random effects specification, the model is as simple as they can be; the data is explained in terms of a main effects model with one additional term to account for the much greater disutility associated with having "extreme problems". On the whole, the results from this modelling appear encouraging. The R-squared (of 0.46) can be considered very good given the type of data analysed here. There is very little data with which a direct comparison of these results can be made since much of the analysis of health state valuations data has been performed on aggregate level data but, in a wider context, a number of econometric models, notably those concerning labour supply functions, report "robust" findings with R^2s as low as 0.1.

In addition, the predicted values from this model are very close to the actual ones for the majority of EuroQol states and the mean absolute difference (of 0.039) is unlikely to be considered meaningful in many contexts. Finally, when the values of a randomly chosen two-thirds of respondents are used to estimate the values of the remaining one-third, the mean absolute difference is again below 0.05. In short, the (relatively parsimonious) model presented in this chapter appears to predict the mean values of the EuroQol states for which there are direct observations reasonably well and thus can be used to interpolate mean values for the states for which no direct observations exist.

CHAPTER 3.3: MODELLING THE EFFECT OF AGE AND GENDER

Introduction

The results from the Main Study and presented in Chapter 3.1 suggest that TTO valuations are different between groups of raters; principally between gender and age groups. This means that the set of 243 EuroQol health state valuations derived for the whole community from the results of the Main Study might not be considered appropriate in all contexts. For example, if the population affected by a particular policy is exclusively elderly people, then it might be considered more appropriate to give greater weight to the preferences of such people. Thus, a 'tariff' based on the valuations of respondents over the age of, say, 60 might be required. This chapter presents valuation tariffs for all EuroQol states based on the gender and the age group of the respondent.

Methods

The regression analysis in this chapter uses the same model and modelling technique that was employed in estimating the EuroQol tariff for the whole population. The analysis is again at the individual level and the same sample of 2997 respondents is used. The question then arises of how best to incorporate the effect of age and sex into the model. To clearly differentiate the different tariffs generated and to enhance the practical and policy relevance of the tariffs, it was decided to separate respondents into two age groups: those aged 18-59 and those aged 60 or over. These two age groups represent the broad bands where valuations differ most and have some legitimacy in terms of health care interventions that might be aimed at 'young' or 'old' people. The model was estimated separately for the two age groups because the data set was too big to allow age dummies to be incorporated into the full model.

With respect to gender, it is important to allow for the fact that its effect is unlikely to be uniform across all states and that the interaction between valuations and gender may differ across levels within dimensions, or across dimensions themselves. Therefore, additional dummy variables (suffixed with gend) were created which were the product of the original 12 dummies and the dummy attached to the sex of the respondent. Thus the regression equation for both age groups is as follows:

$$Y = a + ß_1MO + ß_2SC + ß_3UA + ß_4PD + ß_5AD + ß_6M2 + ß_7S2 + ß_8U2 + ß_9P2 + ß_{10}A2 + ß_{11}N3 + ß_{12}gend + ß_{13}MOgend + ß_{14}SCgend + ß_{15}UAgend + ß_{16}PDgend + ß_{17}ADgend + ß_{18}M2gend + ß_{19}S2gend + ß_{20}U2gend + ß_{21}P2gend + ß_{22}A2gend + ß_{23}N3gend$$

Results

The coefficients on the 24 independent variables together with their associated t-statistics for the two age groups are shown in Table 3.3.1. The R^2s of around 0.47 are high but both models failed the RESET test and suffered from general heteroscedasticity. From Table 3.3.1, it can be seen that the constant for all four population sub-groups is highly significant suggesting that any move away from full health is associated with a substantial loss of utility and the size and significance of the coefficient on N3 highlights the aversion that respondents in general have to "extreme problems" on any of the dimensions.

From the significance of the coefficients on the variables designed to pick up the effect of gender, it would appear that its effect is negligible: all coefficients are insignificant (at the 5% level) in the 18-59 year-olds model and only significant for two variables (SCgend and ADgend) in the 60 or over model. However, when estimated values are calculated for the full set of EuroQol health states from these coefficients, it is found that the absolute difference between the estimated values

for men and women aged 18-59 is 0.05 or more for 133 states (55% of the total number). For the values of the 60 or over age group, the corresponding numbers are 197 (81%) and 124 (51%) respectively.

Since the model could only be estimated for the two age groups separately, it is necessary to use the t-test to compare the impact of age on the relative decrement associated with the various dimensions and levels within dimensions. The results of all the possible comparisons are shown in Table 3.3.2 and suggest that one of the biggest differences between the under and over 60s is in their attitude towards self-care. The t-ratio on SC suggests that level 2 on self-care is considered to be much worse by older respondents than by younger respondents whilst the t-ratio on S2 suggests that there is less of a difference when self-care is at level 3. In addition, the t-ratio on N3 is highly significant, suggesting a greater aversion to "extreme problems" amongst the over 60s.

To enable health state valuations to be readily calculable and to facilitate a comparison of the differential effects that age and gender have on the decrements attached to the dimensions, and to levels within the dimensions, Table 3.3.3 presents the coefficients for each population sub-group in terms of their effect on each dimension and level within dimension. It appears that all respondents are most concerned about being in pain although (younger and older) men are almost as concerned about being confined to bed (mobility level 3). Women, on the other hand, attach a higher decrement than men to having problems with self-care. Women aged 60 or over have greater decrements for all dimensions that contain 'some' problems (level 2) and greater decrements for 'extreme' problems on mobility and self-care, thus resulting in lower estimated values for most heath states. The usual activities dimension plays a consistently small part in accounting for changes in score, although level 3 attracts a greater decrement from younger than from older respondents.

Table 3.3.4 summarises how close the estimated values from the model come to the actual (mean) values of the 42 states directly valued in the study. Estimated values are very close to actual ones for the majority of states across the different population subsamples. The exception, perhaps, is the model using the values of females over 60, where the difference between the estimated value and the actual mean value is greater than 0.1 for 8 of the 42 states. Why the model should not perform so well on this particular subset of the data is unclear.

Discussion

One of the most important issues in the measurement of HRQoL is whose preferences should be given the greatest weight in constructing an index of health state valuations. If health state valuations do not differ significantly by population sub-group, then it is unlikely to matter whose preferences are used. Most studies have reported that variation in health state valuations is not explained by the different demographic characteristics of respondents, such as age or sex. However, results from the Main Study suggest that valuations for health states do differ according to both of these characteristics.

This chapter uses the same regression techniques that were used to generate a set of valuations for all EuroQol health states for the whole community to generate different 'tariffs' according to the age and sex of the respondent. The regression results confirm the patterns from the 'raw' valuations presented in Chapter 3.1: women give valuations than are, on average, lower than those given by men, and this difference increases as the severity of the state increases; and those aged 60 or over give lower valuations than those aged under 60, and this difference also increases with severity. However, by explaining valuations in terms of the different EuroQol dimensions (and levels within dimensions), it is possible to provide an insight into the relative weights that different population sub-groups attach to the different salient features of health. The results suggest that men of all

ages assign a greater disutility to being confined to bed than women do. Younger women attach a greater decrement to pain and older women to self-care than their male counterparts.

Whilst it is unclear *why* different population sub-groups attach different weights to the EuroQol dimensions, the fact that they do suggests that the question "whose values should count?" is an important one. It could be argued that it is appropriate to weight more heavily the preferences of those most directly affected by a particular policy or intervention. For example, when a comparison is being made between treatments for a particular population sub-group, there are strong grounds for using the values of that particular sub-group. However, when making comparisons of interventions that affect many different population sub-groups, it is likely that the views of the whole population will be most relevant.

The choice of tariff may have important implications for resource allocation decisions, particularly as it has been shown that the effect of age and sex is not uniform across the range of health states. Thus, when the EuroQol tariff is used consecutively to quantify the relative changes in health status, the differences between population sub-groups may be even more marked. For example, consider the evaluation of two interventions: A, which is directed exclusively at those aged 60 or over, and B, which is directed at the whole population. Both result in improvements in health from EuroQol states defined earlier as 'severe' to ones defined as 'mild'. If the tariff for the over 60s is used to evaluate A whilst the under 60s tariff is used to evaluate B, then, other things equal, A will appear more attractive since the difference in valuations between mild and severe states is greatest in the former tariff than it is in the latter one. The decision-maker's problems are not made any easier by the fact that this chapter shows that the effect of age and sex is also not uniform across dimensions. Although a decision about whose values should count is ultimately a political not a scientific one, empirical

evidence of the kind presented in this chapter can highlight the implications of the choices made.

CHAPTER 3.4: THE ISSUE OF AGGREGATION

Introduction

An important question that arises with all health state valuation methods is how to aggregate individual responses. Economists have generally advocated that the theoretically correct way to aggregate individual preferences is to calculate the mean value from any given distribution, irrespective of the nature (or skewness) of that distribution. This is based on the principles of Paretian welfare economics which takes account of the strength, or intensity, of each individual's preferences. In this way, it is possible to apply the compensation test which states that a policy change yields an improvement in social welfare if the sum of the benefit to those who gain from the policy is greater than the sum of the disbenefit to those who lose.

However, an alternative view is that, in the realm of public policy, group preferences should be expressed in terms of the median. For example, a health state value that is too high for 50% of the population yet too low for the remaining 50% is one that represents the views of the median voter and therefore might be considered the most suitable. In essence, the difference between the mean and median can be stated thus: the mean takes account of intensity of preference whilst the median treats each persons valuation as equal in a voting context.

Because regression analysis used data at the individual-level, the estimated values in the EuroQol 'tariff' presented in Chapter 3.2 provide good approximations of mean values but not good approximations of median values. Therefore, those who

consider the median to be a more appropriate measure of central tendency, may not consider that particular tariff to be the best representation of social preferences. Of course, no such problem would arise if the valuations from the Main Study were normally distributed, or, in the context of measuring health gain, if the difference between the mean and median was constant across all states. However, problems arise when distributions vary according the severity of the state. And as Figure 3.4.1 shows, the distributions were generally negatively-skewed for less severe states (indicated by higher median than mean values for such states) and positively-skewed for more severe states (as evidenced by higher mean values for such states). This means that the transition from a more severe state to a less severe one will be valued less according to the mean than the median. Therefore, this chapter uses data from the Main Study to estimate a EuroQol tariff of valuations based on median valuations.

Methods

Ideally, it is preferable to estimate a model at the individual level than at the aggregate level but the individual-level data were not readily transformed to a distribution in which individual-level data could be used to yield predicted values which approximate median ones. Therefore, it was decided that the model should be estimated on median values but that the specification of this aggregate-level model should be identical to the individual-level one. Analysis was performed using an OLS approach in which the objective is to minimise the residual sum of squares and hence maximise the R^2. The model was also tested for general heteroscedasticity in the error terms.

Results

The coefficients on the 12 independent variables are shown in Table 3.4.1. The R^2 of 0.98 is very high but the model suffers from heteroscedasticity. Table 3.4.1

shows that the (negative) constant is not significantly different from zero but the size and significance of the coefficient on N3 highlights the aversion that respondents in general have to "extreme problems" on any of the dimensions. It can be seen that the largest decrement for a move from level 1 to level 2 is associated with the self-care dimension, whilst pain or discomfort dominates the weighting for level 3. Table 3.4.2 shows the actual (median) and estimated values for the 42 EuroQol states directly valued in the study, and the differences between them. Overall, the difference between the median and the estimated values is remarkably small: the mean absolute difference is less than 0.05. The biggest discrepancies occur for states 13332 and 33333 where the median value is 0.216 lower and higher, respectively, than the estimated value, but there is only one other state (21323) for which the difference exceeds 0.1. When valuations for all 243 EuroQol states are calculated, 73 (30%) have negative values, and are thus rated as worse than dead.

Figure 3.4.2 compares the tariff of values estimated from median values with that estimated from the individual-level model. Unsurprisingly, the pattern is very similar to that in Figure 3.4.1: values estimated from medians are higher than from the individual-level model for the least severe two-thirds of states (reaching a maximum difference of 0.21 for state 23121), and lower for the most severe one-third of states (reaching a maximum difference of 0.25 for state 33333). 50% of values estimated from medians are at least 0.06 greater than the corresponding value based on individual-level data and the tariff based on median scores contains ten fewer negative values (73 compared to 83), i.e. states rated as worse than dead, than the tariff based on individual-level data.

Discussion

When eliciting the preferences of a group of people, an important consideration is how best to represent the overall preferences of that group. This issue is as

relevant to health status measurement as it is elsewhere, yet it has rarely been discussed in the literature. From the results of the Main Study, a EuroQol tariff was presented in Chapter 3.2. which, because of the methodology employed, provides a good approximation of mean values. The purpose of this chapter has been to present a set of valuations that approximate median ones. The same regression equation used in the individual-level model was adopted in order to generate the median-based tariff.

There are a number of potentially important differences between the model estimated on the basis of median values and the one generated from individual-level data. Figure 3.4.2 shows that not only do valuations differ according to the measure of central tendency chosen, but that these differences are not uniform across the range of health states. For states that are, on average, rated as better than dead, the value estimated from medians is typically higher than the value estimated from individual-level data whilst for states that are, on average, rated as worse than dead, the reverse is true.

Therefore, the benefit derived from moving between different health states will differ according to whether the set of social preferences that approximates mean or median values is chosen. Of course, the precise magnitude of the difference will depend on the initial and final health states. It would be impossible in the context of this chapter to illustrate the differential effect of movements between all possible pairs of states but for illustrative purposes, suppose that all patients with a particular condition fall into one of 5 groups of health states. Table 3.4.3 defines these groups and presents the mean health state value associated with patients in each group (there is very little difference between the mean and median value) according to whether values from the individual- or medians-based tariff are used.

For simplicity, consider treatments that only involve movements 'up' one group. It can be seen from Table 3.4.3 that moving patients from states in group 3 to those

in group 2 and from states in group 2 to those in group 1 yields approximately the same benefit in terms of health gain whether the individual- or medians-based tariff is used. However, moving patients from group 5 to group 4 and from group 4 to group 3 yields considerably more benefit when the medians-based tariff is used than when the individual-based one is used. If the only objective of the health care system were to maximise health gain, in a choice between two policies; A, which takes one group of patients from group 3 to group 2 and another from group 2 to group 1, and B, which takes one group of patients from group 5 to group 4 and another from group 4 to group 3, then, *ceteris paribus*, A would be chosen if the individual-based tariff was used whilst B would be chosen if the medians-based tariff was used.

Therefore, and in the absence of a "gold standard", whether A or B is chosen in this case is ultimately a function of one philosophical position on how preferences should be aggregated. The set of valuations presented in Chapter 3.2 are for use by those committed to the mean whilst the values presented in this chapter are for use by those who favour the median.

CHAPTER 4: INTERPRETING VALUATIONS AT THE INDIVIDUAL LEVEL

CHAPTER 4.1: POTENTIAL BIASES IN THE TTO

Introduction

Since valuations in the Main Study and hence the EuroQol tariff values were based upon responses to TTO questions, it is important to consider the extent to which such responses can provide unbiased estimates of the relative utility loss associated with different severities of illness. Although a number of authors have discussed the *sources* of potential bias in TTO responses (see, for example, Loomes and McKenzie [1989]), little attention has been directed towards explicitly setting out the likely *magnitude* of these biases. In considering the impact of time preference, this chapter could be properly regarded as a logical tidying up exercise. However, another potentially important source of bias considered in this chapter; namely, the reallocation of lifetime consumption, has not been discussed elsewhere.

Consider first a Von Neumann Morgenstern (NM) expected utility maximiser facing current-year probabilities p and q of death and a permanently disabling illness/injury respectively. For simplicity, death and disability will be treated as mutually exclusive events, so that assuming strong separability on the time dimension (see Broome [1993]), and ignoring the possibility of injury or premature death in other than the current year, the individual's lifetime expected utility is given by

$$EU = \left[(1-p-q)\left(\sum_{t=0}^{T-1}\rho^t L(C)+\rho^T D\right)\right] + \left[q\left(\sum_{t=0}^{T-1}\rho^t I(\hat{C})+\rho^T D\right)\right] + pD \quad (1)$$

where L(.) and I(.) denote the NM annual utility of consumption functions for full health and injury/illness respectively, C denotes constant consumption per annum conditional on full health, $\hat{C}$ denotes constant consumption per annum conditional on

injury/illness (where $\hat{C}$ may differ from C), ρ is a discount factor reflecting the individual's rate of time preference, T is the individual's maximum life expectancy and D denotes the NM utility associated with the prospect of death. Since the uncertainty in this expected utility framework is resolved in the initial period, it mimicks the formulation of TTO questions.

Note that for simplicity it has been assumed that consumption is time-invariant. Indeed, it is probably impossible to interpret the response to a TTO question in relative utility terms without imposing a time-invariant condition on consumption. Of course, given that the only uncertainty in a TTO question is associated with the determination of the lifetime health state at the beginning of the first period, time-invariant consumption would in fact be optimal provided that the time preference and interest rates were equal.

This can be illustrated by considering a simple two-period model. Suppose that the initial uncertainty is resolved such that an individual experiences full health. This individual's optimal life-cycle consumption decision is to maximise $L(C_1) + \rho L(C_2)$ subject to $C_1 + \mu C_2 = W$, where ρ and μ are discount factors reflecting the time preference and interest rates respectively, and W is initial wealth, including first and discounted second period income. It is then straightforward to show that the solution to this constrained maximisation problem is such that if $\rho = \mu$, then $C_1 = C_2$. Precisely the same argument applies if the initial uncertainty results in ill health. Here the optimal life-time consumption decision is to maximise $I(\hat{C}_1) + \rho I(\hat{C}_2)$ subject to $C_1 + \mu C_2 = W$, where $\hat{W}$ may differ from W to the extent that ill health (presumably adversely) affects income.

In the model developed in this chapter, it will be assumed for simplicity that ρ is independent of the health state and that D is constant, so that without loss of generality, L(.), I(.) and D can be scaled such that D=0. In addition, it will be assumed that

$$(L(C) > I(\hat{C}), \tag{2}$$

$$(L'(.), I'(.) > 0, \tag{3}$$

and

$$0 < \rho \leq 1. \tag{4}$$

Denoting the individual's marginal rate of substitution (MRS) of wealth for risk of death and wealth for risk of illness/injury by M_D and M_t respectively, with D=0 it is then straightforward to show that

$$\frac{M_I}{M_D} = \frac{\sum_{t=0}^{T-1} \rho^t L(C) - \sum_{t=0}^{T-1} \rho^t I(\hat{C})}{\sum_{t=0}^{T} \rho^t L(C)} \tag{5}$$

$$= \frac{L(C) - I(\hat{C})}{L(C)} \tag{5'}$$

$$= 1 - \frac{I(\hat{C})}{L(C)}. \tag{5''}$$

While the argument is developed for the single-period case, precisely the same result follows for the multi-period case with lifetime expected utility expressed as in equation (1), given that D has been set equal to zero (see Jones-Lee [1989]). And whilst this result has been derived on the admittedly somewhat restrictive assumption of expected utility maximisation, it can be shown that with appropriate reinterpretation of L(.) and I(.), a similar result follows in the case of a wide range of non-expected utility maximisation theories provided that the latter satisfy the betweenness axiom i.e. an individual who is indifferent between X and Y will also be indifferent between those alternatives and every probability mixture $pX + (1-p)Y$, $0 < p < 1$, ensuring that indifference curves in the Marschak-Machina Triangle are linear (see for example Jones-Lee [1989, pp 34-36] or Jones-Lee *et al* [1993, Appendix 2]).

Clearly, then, the principal focus is upon whether or not the response to a TTO question yields an unbiased estimate of $\frac{I(\hat{C})}{L(C)}$. Suppose that in response to such a question, the individual, whose lifetime expected utility is as specified in equation (1), indicates that he/she would be indifferent between the certainty of spending 10 years in the state of injury/illness referred to above, followed by death on the one hand, and the certainty of spending τ years in full health, followed by death on the other.

Clearly, from (2) we shall have $\tau<10$. In what follows, I will focus upon the case in which the injury/illness concerned is not judged to be as bad as or worse than death, so that $L(C)>I(\hat{C})>0$ and hence $\tau > 0$. With D=0, it follows that:

$$\sum_{t=0}^{\tau-1} \rho^t L(\tilde{C}) = \sum_{t=0}^{9} \rho^t I(\hat{C}), \tag{6}$$

where $\tilde{C}$ ($\geq C$) is the individual's planned annual consumption given that he expects to live for only τ (<10) years. From (6), it is immediate that

$$0 < \rho < 1 \quad \Rightarrow \quad \frac{I(\hat{C})}{L(\tilde{C})} > \frac{\tau}{10} \tag{7}$$

and

$$\rho = 1 \quad \Rightarrow \quad \frac{I(\hat{C})}{L(\tilde{C})} = \frac{\tau}{10}. \tag{8}$$

Furthermore. given $L'(.) > 0$,

$$\tilde{C} \begin{matrix} > \\ = \\ < \end{matrix} C \quad \Rightarrow \quad \frac{I(\hat{C})}{L(C)} \begin{matrix} > \\ = \\ < \end{matrix} \frac{I(\hat{C})}{L(\tilde{C})}. \tag{9}$$

Since it is highly unlikely that the individual would consume less per annum if they lived for τ (<10) years rather than the full ten years, it seems reasonable to impose the

restriction $\tilde{C} \geq C$. Given this restriction, and given that the individual does not have a negative rate of time preference (see Chapter 4.2) over life years (i.e. $0 < \rho \leq 1$), it follows from (7), (8) and (9) that

$$\frac{I(\hat{C})}{L(C)} = \frac{\tau}{10} \quad \text{iff } \tilde{C} = C \text{ and } \rho = 1. \tag{10}$$

That is, in order for the response to the TTO question to provide a direct and unbiased estimate of the ratio $\frac{I(\hat{C})}{L(C)}$, it is necessary that there should be <u>no</u> reallocation of lifetime consumption and <u>no</u> discounting of future utilities. If, by contrast, there is either reallocation (i.e. $\tilde{C} > C$) and/or discounting (i.e. $0 < \rho < 1$), then $\frac{\tau}{10}$ will unambiguously <u>under</u>estimate $\frac{I(\hat{C})}{L(C)}$.

Lifetime reallocation of consumption

If the individual has the opportunity to reallocate lifetime consumption then it seems likely that $\tilde{C}$ will exceed C. For example, an individual who is to some extent consuming out of accumulated wealth and, in the model specified above, plans to consume at a greater rate if they knew for certain that their life expectancy were to be reduced. Since the TTO is based on the comparison of two alternatives for which the respective life expectancies *are* known for certain, lifetime reallocation of consumption is clearly a source of potential bias in TTO responses.

However, it transpires that there are grounds for believing that the impact of such reallocation will be negligible. By applying a variant of the argument developed in Jones-Lee [1989, pp115-116] to equation (1), with D=0 and assuming that L(.) and I(.) are bounded above (which is necessary if the individual is to be immune to versions of the St. Petersburg Paradox), it is fairly straightforward to show that

$$\frac{L^*}{L(C)} < \frac{1-p-q}{1-p-q-\Delta p^*} \qquad (11)$$

where L* denotes supL(.) and Δp^* is the individual's "maximum acceptable increase in p" i.e. the increase in p for which the compensating variation in terms of an increase in C becomes unbounded.

Now, for most people it seems reasonable to assume that: i) their risk of death in the current period is less than 1 in 100 (i.e. $p<10^{-2}$); ii) their risk of suffering other than the most minor injury or illness is less than 1 in 10 (i.e. $q<10^{-1}$); and iii) the maximum increase in the risk of death they would be prepared to accept is also likely to be less than 1 in 10 (i.e. $\Delta p^*<10^{-1}$). It follows from equation (8) that even an unbounded increase in C will cause L(.) to increase by, at most, about 13%. Indeed, with $p=10^{-3}$, $q=10^{-2}$ and $\Delta p^*=10^{-2}$, which are not entirely implausible orders of magnitude, the increase would be only about 1%. It therefore seems clear that for the (relatively modest) increase from C to $\tilde{C}$ that might be expected from a lifetime reallocation of consumption in the context of a TTO question, $L(\tilde{C})$ would exceed L(C) only by a very small percentage. Of course, this ignores the fact that in practice most people will be consuming out of current income (not accumulated wealth) and will therefore have little scope for reallocating consumption in any case.

Discounting

If $\rho = 1$, then each year of life in constant quality (L or I) yields the same utility. However, in responding to a TTO question, the individual may be prepared to sacrifice more years of life in the future relative to years of life now, in which case they have a positive rate of time preference and will be discounting the future; hence $0 < \rho < 1$. The concept of time preference is an important one since it is one of three principal justifications for discounting the flow of future health care costs and benefits. The other two are the returns on investment, where the opportunity cost of a given outlay is assumed to vary through time due the possibility of earning interest, and the concept of diminishing marginal utility, which implies that the incremental utility derived from

health falls through time (see Warner and Luce [1982]). For these reasons, discounting is deemed appropriate if comparisons between immediate and delayed consumption are to be made.

Most economic analyses of intertemporal choice rely on Discounted Utility Theory (DUT) which assumes the greater importance of today against the lesser importance of tomorrow. The present may be seen as more important than the future for a number of reasons. First, presently available money can gain interest and thus the receipt of £100 now is likely to be preferred to the receipt of £100 delayed for one year. Second, because the enjoyment of things present is certain, whereas deferred pleasures are uncertain, people are likely to take what they can have now rather than wait for what they might not get later. Finally, people may simply derive greater utility from having good things as soon as possible and putting off bad things for as long as possible. The first reason is in part determined by the availability of financial markets, the second by people's attitudes towards risk, whilst the last reason is an example of pure time preference.

In addition to assuming that value declines through time (i.e. the discount rate is positive), DUT also assumes that value declines at an annual exponential rate i.e. the same discount rate is applied in each successive time period [see Fisher 1930]. DUT offers a descriptive model of human behaviour i.e. people act *as if* they discount. Given a choice of pleasure now or pleasure later, people prefer pleasure now. Given a choice of suffering now or suffering later, people prefer suffering later. The concept of time preference is well established in the health economics literature; for example, it has been used to explain variations in diet, exercise, and cigarette smoking amongst individuals (see Fuchs [1982]).

To get some feel for the extent to which $\frac{\tau}{10}$ might underestimate $\frac{I(\hat{C})}{L(C)}$ in the presence of positive time preference, consider integer values for τ between 1 and 9 and, in addition to $\rho = 1$, three further values for ρ of 0.95, 0.91 and 0.87 (corresponding, approximately, to annual time preference rates of 5%, 10% and 15%

respectively). In view of equations (5") and (7), Table 4.1.1 shows the values of $\frac{I(\hat{C})}{L(C)}$ that would result. Although the effect of a positive rate of time preference is to increase the ratio $\frac{I(\hat{C})}{L(C)}$ for any given value of τ, the effect is not uniform across all values of this ratio. The absolute difference between undiscounted and discounted values of $\frac{I(C)}{L(C)}$ is smallest for high and low values (i.e. for values of $\frac{I(C)}{L(C)}$ close to 1 and 0) and largest for values around 0.5. Clearly then, even setting aside the possible impact of lifetime reallocation of consumption, assuming that $\rho = 1$ (which is an assumption made in Chapter 3) will underestimate $\frac{I(\hat{C})}{L(C)}$; the extent to which will clearly depend upon how far ρ deviates from 1 but also on the severity of the permanently disabling illness/injury.

However, valuations generated by the TTO method do not have to be predicated on the assumption of no discounting. All respondents indicate in answering a TTO question is the number of years in L that are regarded as equivalent to a longer period of time in I. The value that is attached to $\frac{I(\hat{C})}{L(C)}$ (even assuming that $\tilde{C} = C$) is a separate issue. Thus, Table 4.1.1 shows how <u>different</u> values for $\frac{I(\hat{C})}{L(C)}$ can be generated from the <u>same</u> point of indifference established in a TTO question, depending on the value of ρ used.

Discussion

Despite being developed more than 20 years ago, surprisingly little attention has been directed at the extent to which TTO valuations will under- or over-estimate the required relative utility loss associated with illness. This chapter has shown that in order for a response to a TTO question to provide a direct and unbiased estimate of

the ratio $\frac{I(\hat{C})}{L(C)}$, it is necessary that: i) there is no reallocation of lifetime consumption; and ii) there is no discounting of future utilities. If there is either reallocation and/or discounting, then it is shown that a TTO response that is not adjusted for these effects will unambiguously underestimate $\frac{I(\hat{C})}{L(C)}$.

In the case of reallocation of lifetime consumption, the extent to which $\frac{I(C)}{L(C)}$ is underestimated is likely to be very small. Since most people consume out of current income, the extent to which they will be able to consume at a greater rate if their life expectancy were to be reduced is therefore highly constrained. Even for those people who are to some extent consuming out of accumulated wealth, it transpires that by making entirely plausible assumptions about the risk of death and the maximum increase in that risk that an individual would be prepared to accept, the impact of lifetime reallocation of consumption is almost certainly trivial. Therefore, it seems reasonable to assume that $\tilde{C} = C$.

The effect of discounting, however, is <u>non</u>-trivial. Assuming an annual time preference rate of 5%, an undiscounted TTO response will underestimate $\frac{I(\hat{C})}{L(C)}$ by as much as 0.06 (in relation to a true value of 0.44). Therefore, unless a reliable method of exploring the effect of time preference on health state valuations (and benefits more generally) can be constructed, then choices between alternative uses of resources that have different benefit streams are unlikely to fully represent individual or social preferences. This is an issue which is addressed more fully in the next chapter.

CHAPTER 4.2: THE EFFECT OF TIME PREFERENCE AND DURATION ON TTO RESPONSES

Introduction

The previous chapter has shown that for responses to TTO questions to provide unbiased estimates of the relative utility loss associated with different severities of illness, it is necessary that there is a zero rate of time preference i.e. that the timing of an event does not affect its relative value. The logical question to follow from this is to what extent is this a plausible assumption? There is some evidence to suggest that when a health state is experienced matters to significant numbers of people. For example, Redelmeier and Heller [1993], in a study of time preference rates over acute health states, found that the timing of identical periods of ill health mattered to a sizeable number of respondents (almost 40% of responses from a sample of medical students and doctors implied a non-zero rate of time preference). This chapter reports on a pilot study designed to test the feasibility of using the TTO to isolate the effect of pure time preference.

The study was also designed to address another important and related issue; namely the effect that the time spent in a particular health states may have on its valuation. It has been common for researchers to weight each year of added life equally, that is, to assume that the value given to a health state is linearly related to the time spent in that health state. Thus, if the health state stays constant over time then the valuation is assumed to stay constant over time. However, the value of the health state is likely to be a function of how long the state lasts for. It is entirely plausible that its value may increase over time as people adapt to illness, or adjust their expectations in the light of changes in their circumstances. Equally, its value may decrease over time as severe dysfunction becomes increasingly intolerable. If the value of a health state is indeed a function of its duration then this needs to be taken into account when we are measuring the benefits associated with different health care interventions.

There is evidence to suggest that duration can have a significant effect on health state valuations (see Christensen-Szalanski [1984], Lipscomb [1989] and Burrows and Brown [1992]). Sackett and Torrance [1978] found from a sample of about 200 members of the general public that when health states are specified for durations of three months, eight years and a lifetime, mean TTO valuations declined as the time spent in the state increased. Using the VAS for the same three time periods, Sutherland *et al* [1982] found from a convenience sample of 20 professional colleagues that the proportion preferring immediate death to varying durations in each of five health states increased as the duration of the states increased. More recently, Ohinmaa and Sintonen [1994] elicited VAS valuations from a convenience sample of 60 health economics students for states lasting one month, one year and ten years and also found valuations to be a decreasing function of duration.

Therefore, there appears to be some evidence that some poor states of health become more intolerable the longer they last. However, none of these studies attempted to separate duration from pure time preference. Simply examining the variation of the health state valuation with the time spent experiencing the state will not adequately distinguish between these two effects. Hence the need for the study reported in this chapter.

Study design

Respondents were presented with 7 cards representing 6 EuroQol states (11111, 11121, 11122, 21232, 22233 and 33333) plus "Immediate Death". They were told that each state (except "Immediate Death") would last 10 years without any change and that what happens thereafter is not known and should not be taken into account. Respondents were asked to rank the 7 states in order from best to worst and then to rate them on a VAS with endpoints of 100 (best imaginable health state) and 0 (worst imaginable health state). Respondents then valued five states (in the order of 21232, 22233, 11121, 33333, 11122) on the TTO using the protocol outlined in Chapter 3.1. After completing these "standard" TTO questions for 10 years, respondents were asked TTO questions for one year and one month. As it was felt unreasonable to

present scenarios to respondents in which they would die after one year or one month, the shorter durations were supplemented with healthy time up to a total of 10 years. It is possible that much shorter life expectancy may bias responses although the direction of this bias remains unclear (see Shiell, King and Briggs [1993]).

Given that time preference may affect valuations, it was also necessary to elicit valuations for states that last for the same duration but whose timing is different. This was achieved by asking respondents first to suppose that the one year in poor health happens at the beginning of the 10 year period and second to suppose that the one year in poor health happens at the end of the 10 year period. To avoid respondent burden, the TTO questions for one month were asked only for the one month of poor health occurring at the beginning of the 10 year period. In summary, respondents were presented with the following four scenarios for each health state, each of which was followed by death;

1) 10 years of the specified health state,

2) one year of the specified health state followed by 9 years of full health,

3) 9 years of full health state followed by one year of the specified health state,

4) one month of the specified health state followed by 9 years 11 months of full health.

The way in which the scenarios were presented to respondents is shown in the Appendix. From the responses to the scenarios it was hoped that the effects of time preference and duration on health state valuations could be identified. Relative preferences over scenarios (2) and (3) can be seen as trade-offs between outcomes occurring at different points in time and thus from these responses each respondent's time preference rate for health could be estimated. The effects of these time preference rates on the implied valuations from all four scenarios can then be investigated.

Interviews were conducted by six professional interviewers from SCPR. Each interviewer was asked to conduct either 6 or 7 TTO-based interviews in order to achieve a sample of 40. This sample size was deemed adequate both to test the feasibility of the protocol and to allow general conclusions about this type of approach to be drawn from quantitative data analysis. Because this was a Pilot Study, the sample was

not intended to be representative of the general population but was instead a quota sample of adults aged 18 or above living in a range of residential areas convenient to the interviewer. The interviews were conducted over a two-week period in September 1993.

Data analysis

Raw VAS scores have been transformed onto a 'standard' 0-1 scale in order to produce a 'unit of health' which is comparable across all respondents. To allow for time preference in the TTO valuations, an implied discount rate, r, was calculated for each state for each respondent by finding the value of r which makes the number of healthy years elicited from scenarios (2) and (3) above equivalent to one another. For example, if one year in poor health followed by 9 years in full health (scenario (2) above) is equivalent to 9 and 1/4 years in full health whilst 9 years in full health followed by one year in poor health (scenario (3) above) is equivalent to 9 and 1/2 years in full health, then:

$$1 + 1/(1+r) + 1/(1+r)^2 + \ldots + 1/(1+r)^8 + (1/4)/(1+r)^9 =$$
$$1/2 + 1/(1+r) + 1/(1+r)^2 + \ldots + 1/(1+r)^8 + 1/(1+r)^9$$

which reduces to $1 + (1/4)/(1+r)^9 = 1/2 + 1/(1+r)^9$

rearranging and solving for r gives $r = (1.5)^{1/9}-1 = 0.046$, or 4.6%.

For the "standard" 10 year TTO questions, the value of a health state that is rated as better than dead is given by the formula x/t where x is the number of years in full health that is equivalent to t years in poor health. For a state that is rated as worse than dead, its value is taken to be (x/10)-1. These calculations assume that each year is weighted equally and thus assume that there is no time preference or discounting. The same assumptions apply to the scenarios in which one year, or one month, of poor health is experienced at the beginning of a 10 year profile (scenarios (2) and (4) above). For example, consider the case where one year in poor health followed by 9 years in full

health is equivalent to 9 and 1/4 years in full health. Attaching a value of 0.25 to the poor health state is only valid if the 9 months sacrificed at the end of 10 years are strictly comparable with one year in poor health now. An analysis of time preference rates may thus indicate the validity of these assumptions.

Calculating a value for one year in poor health after 9 years in full health (scenario (3) above) from the stated number of years in full health that this is equivalent to, rests on a different assumption; namely, that the first nine years spent in full health are equivalent regardless of what is experienced in the tenth year. For example, consider the case where 9 years in full health followed by one year in poor health is equivalent to 9 and 1/2 years in full health. Attaching a value of 0.5 to the poor health state is only valid if the first 9 years (which are the same in both profiles) yield precisely the same level of utility as one another. Note that this assumption allows for any rate of time preference but nonetheless may be open to question (more anon).

Because the distribution of health state scores was highly skewed, the non-parametric Wilcoxon matched-pairs signed-ranks test was used to test for differences in the valuations of the same state lasting for different durations (significance level $p<0.05$). The analysis of whether time preference rates or valuations are affected by the background characteristics of the respondent is clearly limited by the relatively small number of respondents in the study. However, Mann Whitney U tests were carried out to see if these differed by age, sex, marital status, educational attainment or smoking behaviour.

Results

Respondent Characteristics

Because of some minor fieldwork problems, 39 interviews were conducted. Table 4.2.1 shows the background characteristics of the respondents. Despite being a quota sample, the respondents were broadly representative of the general population, although there were more females than males and the number with a degree or

equivalent was higher than the national average (of 8%). The mean age of respondents was 44.2 years (s.d.=18.1).

VAS Valuations

Table 4.2.2 shows the median transformed VAS values for the states used in this study. No statistically significant differences were found according to respondent characteristics and there was general agreement about the ordering of the states. It appears, then, that the VAS exercise performed its role of familiarising respondents with the health states.

Time Preference

Table 4.2.3 shows the respondents' preferences over the timing of poor health (i.e. their relative preferences over scenarios (2) and (3) above). Poor health later was considered to be preferred to poor health now if the respondent was not prepared to sacrifice as much life expectancy in order to avoid poor health in 9 years time as they were in order to avoid poor health now (and vice versa). This would imply positive time preference because poor health later involves less disutility than poor health now. Overall, only one-quarter of responses imply a positive discount rate whilst 39% of responses imply a negative discount rate. The remaining 36% of responses imply a zero discount rate; in other words, it did not seem to matter when the year of poor health was experienced for more than one-third of responses.

Figure 4.2.1 shows the distribution of discount rates for each of the health states. There are fewer observations presented here (particularly for the two mildest states, 11121 and 11122) because the calculation of a discount rate is conditional upon the respondent being willing to sacrifice some life expectancy in order to avoid both scenarios (2) and (3) above (see the formula for calculating discount rates shown above). For those respondents who were unwilling to sacrifice any life expectancy in order to avoid a poor health state either now or later implies that they considered the state to be equivalent to full health. Figure 4.2.1 shows that there were some

responses which imply very high (positive and negative) discount rates. Table 4.2.4 shows the mean and median discount rates calculated for each of the health states. The median rate for all states is seen to be zero whilst mean discount rates are small negative numbers for four of the health states and a small positive number for state 33333. The sign and the magnitude of the rate of time preference were not significantly related to respondent background characteristics.

A within-respondent analysis of discount rates produces equivocal results. Table 4.2.5 shows that the respondents fall into three equally-sized groups: one-third exhibit rates of time preference over the five health states that are always positive, or always negative, or (more frequently) always zero; one-third exhibit either positive and zero, or negative and zero time preference; one-third have both positive and negative discount rates. Whether a particular respondent exhibits positive, negative or zero time preference does not appear to be a function of the severity of the health state valued. In addition, these three groups of respondents are very similar to one another in terms of the background characteristics tested.

TTO Valuations

Table 4.2.6 shows the implied median TTO values for the different durations. The valuations for ten years, one year now and one month now are based on the assumption that there is no time preference. The results of the preceding section suggest that, on average, this is a valid assumption. The valuations for one year at the end of 10 years require no assumptions about time preference but instead require that equivalent periods of full health yield the same level of utility as one another despite being followed by different events. It appears from Table 4.2.6 that spending 10 years in any of the health states was considered to be much better than spending only one year or one month in them. For example, 10 years in 21232 is, on average, better than dead whilst one year or even one month in the same state is worse than dead. These, and a number of the other valuations seem implausible: Do respondents *really* feel that spending one month in moderate pain with no other health problems (i.e. 11121) is only marginally better than being dead? It seems unlikely.

Figure 4.2.2 shows the median values that are in Table 4.2.6 together with corresponding values that have been calculated as a proportion of the entire profile. For example, if one year in poor health followed by 9 years in full health is equivalent to 9 and 1/4 years in full health then this profile has a value of 0.925. What Figure 4.2.2 shows clearly is that respondents, on average, adjusted their responses to take account of the fact that shorter lengths of time in poor health are preferable (represented by higher profile scores for one year and one month than for 10 years) but failed to do so sufficiently for this to be represented in higher implied valuations.

Discussion

It appears that implied TTO valuations that are elicited in this study are more a function of questionnaire design or "framing" than they are a function of underlying preferences. Despite an increased reluctance to sacrifice life expectancy (for the profile as a whole), the algorithm used to calculate TTO valuations results in lower health state valuations. The powerful effect that experimental design can have on valuations is not a new phenomenon but it does raise questions about how the TTO technique can be used to value health states that last for short durations given that it may be deemed unreasonable to present respondents with scenarios in which they will be dead in a matter of months or even weeks. The conclusion may be that the TTO method is only feasible for valuing chronic health states that last for durations of, say, five years or more. This in itself is an important contribution of this exploratory study.

The analysis of time preference is not prone to framing effects in the same way as implied valuations are since respondents' relative preferences over experiencing poor health now or in 9 years time are in many ways isolated from such effects. For example, if a respondent prefers poor health at the end of 10 years rather than at the beginning, then they should be prepared to sacrifice more life expectancy to avoid a year of poor health now than to avoid a year of poor health in nine years time, irrespective of the valuations that may be implied from that choice. As a descriptive model of human behaviour, DUT predicts that people will wish to postpone for as long

as possible events that yield disutility and will thus prefer the profile in which poor health is delayed by 9 years. In other words, they will exhibit positive time preference.

At the aggregate level, it appears that there is indifference about the timing of poor health; median discount rates are zero for all states whilst mean rates range from -3.5% to +0.5%. These findings are broadly comparable with those of Redelmeier and Heller [1993] who, in response to questions regarding the timing of identical periods of poor health, observed discount rates of zero in 62% of cases. In addition, Cairns [1992], in a study of 29 economics undergraduates, found that the timing of an identical health state did not appear to matter as much as the timing of identical levels of wealth did. These findings suggest that the implicit assumption of the TTO method that the rate of time preference is zero is valid at the aggregate level.

These results cast doubt on the practice common to many cost-effectiveness analyses of health care interventions of assuming that the value of the discount rate is similar to the financial rate of interest (see Weinstein and Stason [1977]). They also warn against discounting benefits at the rate of 6% which is currently used by the Treasury to discount costs. Moreover, support is lent to the assertion made by Parsonage and Neuberger [1992] that "non-monetary health benefits should not be discounted at the same rate as variables expressed in monetary terms ... instead the appropriate discount rate should be at or close to zero". In short, the practice of using DUT as a basis for making intertemporal comparisons of health benefits may not represent the preferences held by individuals.

The results presented in this chapter show that there is wide variation in time preference rates at the individual level. Although the modal time preference rate is zero for all states, there are a number of responses which imply very high (positive and negative) rates (see Figure 4.2.1). The highest implied rate is 38.3%. That more responses imply negative rates of time preference than positive ones (see Table 4.2.3) contradicts the predictions of DUT. Instead of wanting to postpone poor health (as DUT would predict), it appears that more people want to get it out of the way.

The possibility that people may have negative discount rates (at least within some specified time period) is now recognised in the literature on time preference. For example, Loewenstein [1987] found that, although respondents discounted money values normally, their willingness to pay for a fleeting pleasure increased as delay increased to three days (and declined thereafter). Of more relevance to this chapter, he also found that respondents were willing to pay more to avoid receiving a fleeting unpleasantness that was delayed for three days than they were to avoid the event immediately. Knapp *et al* [1959] found that even rats, when faced with a choice between immediate and delayed pain, tended to choose the immediate pain.

Loewenstein attributes such findings to "savouring", which is the positive utility derived from the anticipation of future pleasant consumption, and to "dread", which is the negative utility resulting from contemplation of future unpleasant consumption. In this study, respondents may prefer to experience poor health now as opposed to in 9 years time because they are getting the worst outcome over with quickly and are thus eliminating dread. That people may wish to get unpleasant experiences over with as quickly as possible is evidenced in everyday life; witness how many people, for example, once having decided to visit the dentist, attempt to get an appointment as soon as possible.

In addition, if the profiles used in the TTO are seen as sequences, it is possible to explain negative time preference in terms of "adaptation". If, as Loewenstein and Prelec [1991] assert, "people tend to assimilate to ongoing stimuli and to evaluate new stimuli relative to their assimilation level" (p348), then when separate events are seen as an integral consumption package, they will choose to start with events that yield the lowest levels of utility and finish with events that yield the highest levels of utility. In this study, poor health followed by good health may be preferred to good health followed by poor health because it affords a positive departure from one's adaptation level, unlike good health followed by poor health which involves a negative departure. Again, examples of this type of behaviour can be found in everyday life; think of the number of people who when eating a meal save the most enjoyable mouthful until last.

However, savouring, dread and adaptation form no part of DUT which assumes that the utilities of different events are independent of one another. In other words, the utility attached to a particular health state is independent of the state(s) of health that precede or follow it. If this assumption is violated, then, using our earlier example, the value attached to a health state, h, when 9 years in full health followed by one year in h is equivalent to 9 and 1/2 years in full health, will not be 0.5. If the utility derived from one health state depends on what state follows, then we cannot be sure of the exact value of h because the utility of the first 9 nine years will not be the same in the two profiles even though the health states over this period are the same. There is certainly the need for additional research into the extent to which the value given to health state is influenced by the health state that precedes it, or is expected to follow it.

The discussion so far has been premised on the proposition that, in aggregate, respondents' answers reflected their true preferences more or less accurately. However, there appears to be good reasons to doubt whether this was actually the case. For example, one-third of respondents had positive discount rates for some states and negative rates for others. Whether a positive or a negative rate was elicited was unrelated to the severity of the health state. This suggests that some factor other than remoteness in time i.e. other than time preference, is being picked up here but it is hard to tell what this factor, or factors, might be.

Indeed, from a quantitative study of this kind, it is hard to tell what respondents had in mind when they valued the various profiles; specifically, we cannot be sure whether respondents were giving answers to precisely the questions that were being asked of them. For example, despite being told that each ten year profile was to be experienced with certainty, we cannot be sure that respondents treated the scenarios in this way. This may be particularly true of the profile in which poor health was to be experienced in 9 years time; for example, it is plausible that some respondents may have thought that a "cure" would be found, others that they would end their life when their time to "suffer" came. Of course, problems of this kind are true of any study which requires answers to hypothetical questions but they seem particularly relevant to questions

which ask respondents to imagine with certainty something that will happen in a number of years' time.

Whilst it is unlikely that the use of patient populations could overcome this problem, it may be that such samples could be used to better measure the extent to which savouring, dread and anticipation affect attitudes towards current and future health states. In a study of women's' attitudes towards anaesthesia during pregnancy and childbirth, Christensen-Szalanski [1984] found that discount rates changed dramatically during pregnancy, suggesting that these phenomena are important ones. Whilst there is undoubtedly the need for more research here, if one objective is to estimate the time preference rates over health for society as a whole then general population samples will also continue to be important.

This exploratory study has shown that separating out the effect of time preference and duration is a difficult and complex task and the valuations derived lack validity. However, although the results concerning time preference contain a large amount of variance, they suggest that further doubt can be cast on the axioms of DUT as a descriptive model of human behaviour. Of course, as Weinstein [1993] points out, "To abandon the normative practice of discounting in cost-effectiveness analyses, would require new arguments that it is normatively flawed, and not just evidence that individual preferences from surveys do not conform to the descriptive model" (p219). However, the motivation to search for new arguments will come from the results of studies such as this which do indeed suggest that individual preferences do not conform with the exponential discount model.

CHAPTER 4.3: THE EFFECT OF DURATION ON VAS VALUATIONS

Introduction

As previous studies (see Chapter 4.2) have shown duration can have a significant effect on valuations, it is important that we have some idea about its magnitude so that

valuations for one duration can be adjusted to better represent valuations for different durations. We could look at the results of the previous studies and then make some *ad hoc* adjustments for the effect of duration by amounts in line with those indicated in the previous studies. But there are three main problems with these studies. The first relates to their generalisability: since most studies have used relatively small (convenience) samples it is questionable to what extent the results can be considered representative of a wider population. The second problem is that their conclusions often only refer to results at the aggregate level, usually in the form of mean valuations. However, it is also important to know the extent to which results at the individual level conform to the apparently robust results at the group level. Finally, none of the studies have considered whether valuations for the same health states lasting different durations can be related to one another in a systematic way.

Or we could assess the effect that the time spent in a health state has on its subsequent valuation ourselves using more representative samples. Ideally, we would wish to elicit valuations for different durations using the same method. Thus, since the tariffs reported in Chapter 3 were based upon TTO valuations for states lasting ten years, we would like to use the TTO to elicit valuations for other shorter durations. However, the results reported in the previous chapter have shown that using the TTO to assess the effect of duration on health state valuations (whether time preference can be accounted for or not) is problematic.

Since previous studies have shown that it is feasible to use the VAS to elicit valuations for different durations, an alternative strategy might involve the following two stage process: first, to elicit VAS and TTO valuations for a long duration and to derive a functional relationship between the two sets of values; second, to elicit VAS valuations for shorter durations and to use the mapping function estimated in stage 1 to 'convert' short-duration VAS valuations into short-duration TTO ones. Given that the Main Study elicited both VAS and TTO valuations for states lasting ten years, this strategy is a feasible one. However, there are at least two problems associated with the approach: first, the results from Chapter 2.1 suggest that it is difficult to estimate a robust relationship between VAS and TTO valuations; and second, even if a robust

relationship could be found, the strategy is based on the assumption that the relationship found for the long duration will hold for other shorter durations.

Because of these problems, this strategy is not adopted in this thesis. However, there are still strong grounds for eliciting VAS valuations for different durations from a representative sample of the population. First, to test whether results generated using small convenience samples are replicated when using a larger general population sample. Second, to provide insights into the magnitude of the effect that duration has on valuations. Of course, differences in VAS valuations will not yield precise estimates of differences in TTO valuations, but they may provide information on the broad range of values that should be used when subjecting TTO valuations to sensitivity analysis. Finally, the VAS results themselves will be of interest to those using health state valuations in clinical decision-making (where the VAS is widely used) and to those generating their own valuation 'tariffs'. Considerations about the time spent in a particular health state will be important when assessing the contexts in which it is appropriate to use valuations derived from such health status measures as the McMaster health state classification system (see Feeny *et al* [1995]), which incorporates lifetime duration into the procedure used to derive valuations, and the EuroQol 'postal' questionnaire (see The EuroQol Group [1990]), in which a duration of one year is incorporated into the standard descriptive format.

This chapter reports on a large scale general population study in which valuations for three different durations were elicited using the VAS. The study was designed to test the hypotheses that: i) health state valuations are a decreasing function of duration; ii) the differences in valuations between durations will be larger, the more severe the health state; and iii) the likelihood of a health state attracting a negative value (i.e. considered to be worse than dead) increases as the duration of the state increases. In addition, VAS-based EuroQol tariffs have been estimated for the three durations for which valuations were elicited. Finally, using data at the level of the individual, an attempt is made to estimate a functional relationship between health state valuations for different durations. Analogous to the rationale for estimating a relationship between different methods, if an algorithm can be found which maps valuations from

one duration into those for a different duration, then it might be possible to elicit valuations for one duration and "convert" them into values for other durations by use of this algorithm.

Methods

Study design

The sample was drawn from respondents who expressed a willingness to be re-interviewed in the Main Study. In order to achieve a sample of 208 respondents, experience of response rates indicated that (at most) 1.5 times as many people needed to be sampled. Assuming that the standard deviations associated with the health state valuations would be similar to those found in the Main Study, a sample of this size would enable a 0.1 difference in valuations between the different durations to be detected at the 0.05 significance level with 80% power. Thus, 312 of those who expressed a willingness to be re-interviewed were sampled. The sample was chosen to be representative (in terms of age, sex, marital status and educational attainment) of the respondents in the Main Study. The interviews were carried out between March and May 1994 (about four months after the Main Study) by 20 interviewers from SCPR. Each respondent was interviewed by the same interviewer as in the Main Study.

Each respondent was presented with the same states that they valued in the Main Study. Initially they were told that each state (except "Immediate Death") would last 10 years without any change. Respondents were asked to rank the 15 states in order from best to worst. They were then asked to rate the 15 states on a VAS, with endpoints of 100 (best imaginable health state) and 0 (worst imaginable health state). The cards describing the health states were then taken up and shuffled, and presented once more to the respondent, who was then asked to rank and rate them again but this time to imagine that they last for one month. When this second cycle was complete, a third cycle was initiated in which the duration of the state was one year.

The VAS valuations were elicited using a method of "bisection", which, it has been argued, generates an interval scale (see Stevens [1971). Respondents first rate their best and worst ranked states on the VAS. They then choose from the remaining states the one whose value on the VAS is roughly halfway between the values assigned to the two extreme states, and assign a value to that state. They are then asked to rate the state whose value on the scale is roughly halfway between this mid-state and the best state, and then to rate the state whose value on the scale is roughly halfway between the mid-state and the worst state. Respondents are then left to rate the remaining 10 states in any order they chose and are allowed the same value for more than one state.

It was decided to tell respondents that what followed the specified time in a particular health state was not known and should not be taken into account. This was considered preferable to the state being followed by immediate death which, for the shorter durations, might dominate the valuations that respondents give, thus artificially driving valuations for these short durations downwards. In any event, not knowing what follows a specified period of poor health is a more plausible scenario and much more like the real world. After completing the valuation tasks, respondents were asked whether their valuations were influenced by the duration of the states. Finally, respondents were asked to give important background information, including their age, sex and educational attainment.

Data analysis

Valuations have been transformed onto a 0-1 scale in order to produce a unit of health which is comparable across all respondents and all durations. The regression techniques and specification tests used to generate EuroQol tariffs for states lasting one month, one year and ten years are identical to those reported in Chapter 3.2. Given the richness of the data available and to avoid the problems of not being able to sufficiently describe individual behaviour from the results of an aggregate level model, the relationships between valuations *across* the different durations have also been investigated at the individual level, using the Tobit specification and associated adjustments as outlined in Chapter 2.1.

An aim of this study was to assess whether differences in valuations between durations were larger for more severe health states. This could be investigated by modelling the relationships between different durations for each of the possible health states valued, and comparing the observed relationships found. If they remained the same as states became progressively worse, then one could conclude that valuations between durations were constant across health state severity. However, the large number of health states available, coupled with the fact that respondents were only required to value a random sample of health states makes this method problematic. Alternatively, the health states have been grouped into five categories on the basis of severity on the EuroQol dimensions.

In order to achieve greater specification of the models, additional independent variables have been introduced which pick up the effect of a number of background characteristics which were found to affect VAS valuations in the Main Study (see Gudex *et al* [1996]). To capture adequately the effect of age, a cubic expression has been used whilst dummy variables have been introduced to pick up the effect of educational attainment. The models themselves have been estimated separately for males and females. Table 4.3.1 defines the explanatory variables used in this analysis.

The models presented in the results refer to those that are considered the 'best' at describing the relationship between valuations for the different durations. To choose between competing models, goodness-of-fit statistics were compared and a RESET test was applied to test for evidence of misspecification. In addition, the predictive ability of the regression models has been assessed by randomly dividing the respondents into two: two-thirds to re-estimate the model and one-third to cross-validate its performance. This has been assessed using the square root of average squared prediction error, calculated as $\sqrt{[\Sigma (Y - \hat{Y})^2 / n]}$.

Results

Study population

Of the 312 people selected for sampling, 236 (76%) yielded an interview. Unsuccessful interviews were largely due to a refusal by the selected person or to the interviewer being unable to make contact with the selected person. Table 4.3.2 shows that the sample was broadly representative of the general population in terms of age, sex and educational attainment. Two respondents had to be excluded from further analysis because they had not given valuations for the one month and one year durations. The data for the remaining 234 respondents, however, was highly complete. Overall, only about 1% of the ranking and VAS data was missing and for the first duration valued (i.e. 10 years) there was no missing data on the ranking exercise and only one missing VAS valuation.

Valuations

Table 4.3.3 shows the mean VAS scores for the three durations. For 34 of the 43 states, the mean value for a state lasting one month is significantly higher statistically (paired t-test; $p<0.05$) than the mean value attached to the same state when it lasts for 10 years. The value for one month is higher than the value for one year on 18 occasions and the value for one year is higher than the value for ten years for 13 of the states. With regard to the hypothesis that dysfunctional states of health are more likely to be rated as worse than dead the longer they last, Table 4.3.4 shows the number of times each state was given a lower value than dead for each duration. Whilst very few respondents consider any of the mild states to be worse than dead irrespective of its duration, the number of times a moderate or severe state is rated as worse than dead increases as the time spent in it increases.

In response to the question "How did the differences in the length of time spent in each state affect your answers?", 49.3% explicitly stated that they thought the states (often the more severe ones) got worse (often worse than death) as they lasted for longer,

although almost one-fifth felt that their answers were not at all affected by the durations specified. However, the differences in scores for these respondents, as well as their background characteristics, were no different from the remainder of respondents.

Estimating the tariffs

The coefficients on the 12 independent variables for three durations are shown in Table 4.3.5. The R^2s (ranging from 0.55 to 0.63) are very high given the type of data analysed but all models suffered from general heteroscedasticity. The constant for all three durations is highly significant suggesting that any move away from full health is associated with a substantial loss of utility and the size and significance of the coefficient on N3 highlights the aversion that respondents in general have to "extreme problems" on any of the dimensions. It can be seen that for all three durations the largest decrement for a move from level 1 to level 2 is associated with pain or discomfort, which continues to dominate the weighting for level 3, although mobility level 3 (confined to bed) is given a somewhat similar decrement. With respect to differences across durations, the largest and most systematic shifts occur in the constant term and in the N3 term, where the decrement associated with each increases as duration increases. There is little or no systematic shift apparent for most dimensions, except perhaps for self-care where the decrements associated with both levels 2 and 3 increase marginally as the time spent in the health state increases.

The actual (mean) and estimated values for the 42 states directly valued in the study are compared in Figure 4.3.1. The difference between actual and estimated values for all three durations is remarkably small. The biggest discrepancy is for state 32211 for a duration of ten years where the estimated value is .093 greater than the mean value for this state, but for only 23 of the 126 comparisons does the difference exceed .05.

Figure 4.3.2 compares the estimated values of the same 42 states for the three durations. It is clear from the Figure that the effect of duration is not uniform across the range of health states, being more pronounced for more severe states than for less

severe ones. As would be expected, and as can be inferred from the coefficients in Table 4.3.1, the largest differences in valuations are between the 10 year and one month durations. For less severe states, the values when the states last for 10 years are about 0.05 below those when the states last for one month. This difference increases with severity, reaching about 0.15 for the more severe states. Interestingly, the differences between the values for states lasting ten years and one year are of approximately the same magnitude as the difference between one year and one month values. In both comparisons, the value for the longer duration is about 0.03 below that for the shorter duration for less severe states, and about 0.07 lower for more severe states.

Comparisons with the Main Study

Since VAS valuations for the 10 year duration were elicited in exactly the same way as in the Main Study, it is possible to compare the two sets of valuations, and hence to make some judgements about whether the one month and one year valuations elicited in this study would be likely to be those that would have been obtained had the sample been larger. A stringent test involves comparing the 243 estimated values from the 10 year valuations in this study with those derived from the Main Study. The following ordinary least-squares regression equation was used to compare the estimates:

$$y = \alpha + \beta x$$

where y is the 10 year VAS valuation from this study and x is the 10 year VAS valuation from the Main Study. The results were as follows:

$$\begin{array}{lccc} y = & 0.02 & + & 0.98\ x \\ & (6.75) & & (118.6) \end{array}$$

$R^2 = 0.98$

Since the intercept term is very close to zero and the slope term is very close to 1 and given that this simple specification did not suffer from any heteroscedasticity, it seems reasonable to conclude that the corresponding valuations are very close to each other.

The relationship between durations

The results of regressing VAS scores for states lasting one month against VAS scores for states lasting one year are given in Table 4.3.6. For both males and females, the coefficients on the values for one year confirm the hypothesis that health state valuations are a decreasing function of health state duration and the coefficients on the health state group dummies suggest that the effect of duration is greater for the more severe states. It appears that, for the same 'one month' valuations, males consistently give lower 'one year' valuations (as evidenced through the smaller constant and larger negative coefficients attached to the health state group dummy variables). For males, there appears to be a relationship between educational attainment and health state valuation that is not observed for females. Specifically, for the same 'one month' valuation, males with greater educational qualifications generally give higher 'one year' valuations those with lesser qualifications. However, neither gender exhibits a strong age effect.

Table 4.3.7 presents the results of regressing valuations for states lasting one year against on those for states lasting 10 years. Again, and as expected, the more severe the health state, the greater the (negative) effect of duration on valuations. These models contain a set of dummy interaction terms between health state group and ten year duration score. Although their inclusion significantly improves the model fit and all have positive coefficients, no consistent pattern is observed. This is most obviously observed for females where the coefficients attached to group 2 and group 5 are the same. Once again, males tend to take greater account of the duration of the states i.e. for given 'one year' scores, they give lower 'ten year' valuations than females (as evidenced by the striking difference between the intercepts for the genders). However, neither age nor educational attainment (rather surprisingly, given the results in Table 4.3.6) appear to affect the relationships between valuations for the different durations.

The results in both Tables 4.3.6 and 4.3.7 show that the majority of variation in health state valuations occurred within individuals (i.e. between health states). Having said this, approximately a quarter (24% for males and 28% for females) of the total unexplained variation in the 'one month-one year' model is due to differences between individuals, whilst in the 'one year-ten year' model the corresponding figures are 43% and 28%, respectively. No evidence of misspecification is observed in any of the models presented here. However, with regard to the performance of the models in terms of their abilities to predicted observed health state valuations from the set of explanatory variables and the valuation of the same health state for a different duration, the models do not perform very satisfactorily. As Tables 4.3.6 and 4.3.7 show, the square root of the average squared prediction error varies between 0.151 and 0.184. Accordingly, estimated health state valuations are, on average, between 15% and 18% from actual valuations.

Discussion

The results presented in this chapter confirm those of previous studies and suggest that the valuation given to a health state is a function of both its severity and its duration. It was found that the mean score for a state lasting 10 years is lower than when the same state lasts for one year which in turn is lower than when that state lasts for only one month. There is also an increasing propensity for respondents' to rate a state as worse than dead as the duration of that state increases. The hypotheses, then, that dysfunctional health states will be seen as increasingly intolerable the longer they last and that the likelihood of a health state being considered as worse than dead increases as the duration of the state increases, are both supported by these data. Moreover, these results are supported by what a large number of respondents *thought* they did when probed about whether the time spent in the health states affected their answers: almost 50% explicitly stated that the states were worse for longer durations.

The results from the estimation of tariff values are encouraging and suggest that the same functional form used in the Main Study is equally applicable to this data. The R^2s of between 0.55 and 0.63 can be considered good, the estimated values for all three

durations are in the majority of cases very close to the actual ones and the results confirm the findings outlined above. The estimated score for a state lasting 10 years is lower than when the same state lasts for one year which in turn is lower than when that state lasts for only one month and it appears that the differences between ten year and one year values are approximately equal to the differences between one year and one month values. The coefficients on the dummy variables for the different dimensions show no systematic pattern, suggesting that the effect of duration is not dimension-specific; rather that it is the severity of the health state overall that matters. Given that the estimated values for states lasting ten years in this study were very similar to the estimated values in the Main Study, it seems reasonable to conclude that, if respondents in the Main Study were asked to value states of one month and one year duration, they would have given very similar values to the corresponding ones obtained here.

This has important implications for those involved in measuring the benefits associated with health care and suggests that the results of studies in which the value given to a health state is assumed to be linearly related to the time spent in that health state should be treated with caution, and subjected to sensitivity analysis over an appropriate range of values. To give some idea about what this range of values should look like, this chapter also sought to estimate a functional relationship between valuations of the same health states for the three durations. Using Tobit regression analysis, models were estimated which related valuations for states lasting one month to those for one year, and valuations for states lasting one year to those lasting ten years. The modelling of individual level data was an attempt at assessing whether it is possible to estimate a robust relationship between the valuations elicited for one duration and those elicited for the same health states but for different durations. If such a function could be estimated, the parameters in the sensitivity analysis would be even more clearly defined.

Although the results confirm the hypotheses cited above, it was not possible to establish a *systematic* relationship between the severity of the health state and the difference in valuation across durations. The predictive capability of the models was

also not very encouraging: on the standard 0-1 (dead-healthy) scale, the value from one duration can only predict the value from another duration on average to within between 0.15 and 0.185 of the actual value. However, these results are not overly surprising given the disparate views that individuals have about health and illness and similar results were observed in Chapter 2.1 when individual data was used to predict health state preferences derived from different valuation methods.

The results do, however, provide some evidence that the relationships between the different durations are influenced by the gender of the respondent. Broadly speaking, men appear to take greater account of the time spent in the health state than females do; they give lower 'one year' valuations for the same 'one month' scores and lower 'ten year' valuations for the same 'one year' scores. It is unclear why males and females should have different mapping functions and, as noted before, the literature to date does not help to shed much light on this subject. Although gender is generally regarded as having negligible a impact on health state valuations, much of the analysis has concentrated on differences *within* a given duration and not *across* durations. Future research efforts should be directed towards addressing this issue; in the first instance, towards testing its robustness.

Given the nature of the study reported in this chapter, the findings that a) <u>all</u> health state valuations are a decreasing function of the time spent in them, and b) duration has a differential effect on the genders, are both with respect to valuations of <u>hypothetical</u> health states elicited from a broadly representative sample of the UK general population. Contrariwise, a number of studies have shown a direct positive link between time in chronic illness and adaptation to that illness (for example, see Meyerowitz [1983] and Cassileth et al [1984]). The suggestion that those in poor health successfully compensate for it may result from an adjustment or response to "cognitive dissonance" whereby people adjust their expectations in the light of changes in their circumstances (see Festinger [1957]). Therefore, it seems entirely plausible that the preferences of the general public might also differ from those of patients with regard to the effect of duration, and that valuations of the milder health states would actually <u>increase</u> as the time spent in them increases.

As with the generation of different sets of valuations according to different background characteristics, the issue that arises here is whose preferences should be used in determining priorities in health care. It could be argued that it is appropriate to weight more heavily the preferences of those who have been in poor health states for a period of time which is considered long enough for them to have adapted to their dysfunction and/or to have made the necessary adjustments to their expectations. But as discussed in Chapter 1.2, since resource allocation decisions primarily affect future (rather than current) patients, it seems legitimate to give weight to the *ex ante* preferences of potential patients when making *ex ante* resource allocation decisions. Again, this is ultimately a political issue.

CHAPTER 5: USING INDIVIDUAL VALUATIONS AT THE SOCIAL LEVEL

CHAPTER 5.1: INDIVIDUAL VERSUS SOCIAL PREFERENCES

Introduction

Chapter 4 assessed the extent to which responses to TTO questions yield unbiased estimates of an individual's preferences over health states and considered whether an individual's VAS valuations for different health states differ according the time spent in those states. The primary purpose for eliciting such valuations, though, is to inform decisions at the social level. Therefore, it is important to consider the extent to which individual valuations can be used to express social preferences. Of course, an important issue in this regard has already been addressed in Chapter 3.4 where the issue of aggregation was discussed. In fact, whether the mean or median is chosen as the most appropriate measure of central tendency (and the choice between them is ultimately a philosophical one), the preferences of each individual are given equal weight: in calculating the mean, each individual's strength of preference is weighted equally whilst each individual is counted as one 'voter' when the median is calculated. But the issue of whether the aggregation of individual preferences is a good approximation of social preferences still remains.

In addressing this issue, I will concentrate on the quality-adjusted life-year (QALY) approach which attempts to combine the value of quality-of-life with the value of length of life into a single index number, which may then be used as a currency in which the benefits of health care interventions can be expressed. Although QALYs can be used to measure the benefit derived from different therapies by an individual

patient, in this chapter they are discussed in terms of their use in the allocation of scarce health care resources among different patients.

In the simplest case, in which a person remains in the same health state for a number of years, QALYs (assuming no discounting) are calculated according to the formula H*Y, where H is the relative weight attached to a particular health state and Y is the number of years spent in that health state. If the value of H is a function of its duration (see Chapter 4.3) then the algorithm for calculating QALYs needs to be modified if they are to more accurately represent preferences and if the possibility of drawing the wrong policy conclusions is to be minimised; H_t*Y, where H varies with time. When H changes over time, the QALY algorithm assumes that the utility derived from the whole profile is equal to the sum of the QALYs derived from each health state. In other words, it is assumed that each individual's utility function is strongly separable on the time dimension i.e. $U(H^1,H^2,...H^n;Y^1,Y^2,...Y^n) = U(H^1)*Y^1 + U(H^2)*Y^2 + ... + U(H^n)*Y^n$. Whilst recognising that this assumption is a restrictive one (see Chapter 4.2), issues regarding its appropriateness are not addressed in this chapter.

It has been suggested by a number of economists that decisions about how to allocate scarce health care resources should be informed by the cost-per-QALY of the different alternatives (for example, see Williams [1985]). According to Weinstein and Stason [1977], the alternatives should be ranked according to the aggregate unweighted number of QALYs gained i.e. those that yield more QALYs are ranked higher than those that yield less. The alternatives are then "selected from the top until available resources are exhausted". This defines the objectives of the health care system in terms of the maximisation of health gain (which in this chapter is defined as efficiency) and is consistent with defining need in terms of capacity to benefit: that Weinstein and Stason and Williams define need in this way is therefore not surprising.

In this respect, it has been argued that the QALY approach fails to take account of distributional issues that are known to be important in the context of health care (see Nord *et al* [1993] and Nord [1994]). It is argued that people would want decision-makers, when choosing between alternatives, also to be concerned with how those QALYs are distributed, and again different definitions of need are relevant. For example, if need is defined in terms of ill health - those in the worst health states are those most in need of treatment - then pre-treatment health status becomes the most important consideration in determining priorities. Alternatively, if need is defined in terms of final health status, then post-treatment health status is more important. There are clearly a number of other definitions of need (for a more detailed discussion see Culyer [1995]), each with different implications for the allocation of resources, but the definitions cited above highlight the tension between efficiency (defined in terms of health gain) on the one hand, and concerns for equity (defined in terms of pre- and post-treatment health status) on the other.

It is likely, then, that people would want resource allocation decisions to be informed both by efficiency and equity considerations. In economics, we typically try to answer questions of this kind by considering the form of the social welfare function (SWF) to be employed. For example, Wagstaff [1991] has noted that "this [inequality] aversion could be incorporated into resource allocation decisions by using an appropriately specified SWF", and a number of others authors have addressed this possibility (for example, see Mooney and Olsen [1991] and the empirical analysis presented in Chapter 5.2 below). In this chapter, an approach is suggested which uses a particular class of health-related social welfare functions (HRSWF) [first proposed by Atkinson in 1970] which allows efficiency and equity to be considered independently. A particular functional form is postulated which is sufficiently flexible to represent a wide range of social preferences. Whilst Wagstaff [1991] and Jones-Lee and Loomes [1995] have employed this functional form, this chapter shows how the framework suggested allows a number of different hypotheses to be tested in a relatively straightforward way.

The social welfare function

Following Ng (1983), we may characterise social welfare by a vector of individual welfares, (W^1, W^2, ..., W^I), where W^i is the welfare (or "good") of the ith individual and I is the number of individuals. Economists have typically argued that individuals are the best judges of their own well-being, and that social welfare depends only on the welfare of persons in society. A Bergsonian SWF (Bergson 1938) may then be written as

$$W = f(W^1, W^2, \ldots, W^I) \tag{1}$$

where the precise form of f is unspecified other than that it is strictly increasing in all of its arguments. In this way, welfare economics can be said to be written largely from a consequentialist and individualistic standpoint, implying a refusal to adopt a paternalistic attitude. Note that paternalism is acceptable under the social decision-making approach to resource allocation, which suggests that policies which maximise the objectives of the decision-making unit (for example, the NHS) should be adopted (for more details of this paradigm, see Sugden and Williams [1978]). In this chapter, however, it is not only assumed that individuals are the best judges of the their own (real or hypothetical) health state but also that the objectives of government are defined in terms of individual preferences i.e. in terms of arguments in the SWF

According to the Pareto criterion (see Pareto [1935]), an increase in some W^i and decrease in no W^i is a sufficient condition for an increase in social welfare. Some economists argue that this is a sufficient and a necessary condition for an improvement in social welfare (see, for example, Gravelle and Rees [1981]). However, this very narrow interpretation implies that improvements in social

welfare can only be brought about from a policy change in the (highly unlikely) event that nobody loses (and at least one person gains) from the proposed change.

Thus, others (for example, Sugden [1981 p37]) suggest that the Pareto criterion is only sufficient for an improvement in social welfare Therefore,

$$\frac{\delta f}{\delta W^i} \geq 0 \text{ for all I} \tag{2}$$

Another more restrictive definition of social welfare is the (Benthamite) utilitarian concept of the sum total of individual happiness

$$W = U^1 + U^2 + \ldots . U^I = \sum_I U^i \tag{3}$$

where U^i is a utility index representing the preferences of individual I. In this chapter, it is assumed that *Ui* is a *cardinal* utility index, unique up to a positive proportionate transformation i.e. if *Ui* is a representation of the individual's preferences then the only admissible transformation is aUi where $a>0$. The advantage of this approach is that the SWF aggregates individual utilities in a direct and transparent manner.

In the discussion that follows, it is necessary to assume that it is possible to make interpersonal comparisons of utility. It is now well established that different SWFs require different types of comparability (see Sen [1977]). For example, maximising the sum of individual utilities requires that differences in utilities can be compared (referred to as unit comparability) whilst adoption of the Rawlsian criterion of maximising the welfare of the worst-off individual requires only that we know whether one person is better or worse off than another (referred to as level comparability). For the purposes of this chapter, full comparability, which subsumes both level and unit comparability, is

required. By going beyond individual *orderings*, problems associated with Arrow's General Possibility Theorem (see Arrow [1951]) are avoided. In addition, and without loss of generality, no distinction is made between welfarism, which is concerned with self-assessed utility, and extra-welfarism, which is typically the framework adopted in measuring health gain since it is often assumed that a given health state has the same value across all individuals.

The application of a utilitarian SWF to health care implies that HRSW is maximised when the total number of QALYs gained (subject to a budget constraint) is maximised, irrespective of how those QALYs are distributed. It is this approach that Nord *et al* [1993] appear to object to, claiming that "The rule is almost certainly defective as it ignores distributional considerations and issues of entitlement that are known to be of importance in decision-making, especially in the health sector". I agree. But the utilitarian approach is only one approach to deriving a HRSWF from individual utilities. Another might be to adopt a decision rule that gives greater weight to one individual's utility than to another's. For example, a Rawlsian "maximin" approach would require giving greatest weight to the treatment of the most seriously ill individual since maximin judges states of the world according to the level of (health-related) utility of the worst-off person. Between the utilitarian and Rawlsian formulations lies the "convex" SWF which implies that there exists a trade-off between efficiency and equity.

Therefore, taking account of equity and distributive considerations is not inconsistent with the measurement of individual utility, nor is it inconsistent with the interpersonal comparison of individual utilities. For example, assuming that an agreed unit of value applicable to each individual has been established, the conclusion that individual i is in better health than individual j, and that i's gain in health from a change from Y to Z will exceed j's loss (i.e. a positive statement) does not imply what *ought* to be done until an objective function (i.e. a normative statement) is specified. Therefore, if our objective is to maximise the sum of health utilities (i.e. the utilitarian SWF), we choose Z; if we

want to maximise the health of the worst-off individual (i.e. the Rawlsian SWF), we choose Y. Of course, there are a number of other objectives which we may wish to satisfy, and these will be considered below. The measurement of individual utilities, then, provides us with the flexibility to formulate (and even subsequently revise) any number of possible SWFs from them. And the process of collapsing individual utilities into an overall SWF makes explicit the assumptions and philosophical basis on which such aggregation is based.

Distributive justice and the social welfare function

If we assume that HRSW is a function of individual utilities i.e. that factors other than individual utilities are either regarded as irrelevant to social welfare or as being held constant, we can use the tools of welfare economics to represent a number of different SWFs. In Figure 5.1.1, the utility derived from different states of health by two (or two groups of) individuals, i and j, are shown on the x and y axes, respectively. Notice in this example that health state utilities have an upper bound of 1 (full health) and a lower bound of 0 (death). It would, however, be possible to expand this analysis to allow for states that attract negative utilities i.e. for those considered to be worse than dead. Four SWFs are postulated:

1. The utilitarian SWF (UB): a straight line drawn at right angles to the 45° line, indicating that maximising total health gain is the objective, irrespective of distributional considerations.
2. The Rawlsian SWF (UR): welfare is not increased unless the health state of the most seriously ill individual is improved.
3. The convex SWF (UC): implies that there exists a trade-off between efficiency (i.e. maximising health) and equity (i.e. greater concern for those in poor health).
4. The concave SWF (UI): implies inequality proneness.

In addition, it is possible that the objective is to equalise the health of both individuals (i.e. strict egalitarianism) which means the SWF would consist of points on the 45° line with points further from the origin being preferred to those closer to the origin. However, assuming that pathological budget lines or utility possibility frontiers (UPFs) are ruled out, preferences for strict egalitarianism will lead to the choice of the same point on the UPF as Rawlsian preferences.

All these possible formulations can be represented by a class of SWF first proposed by Atkinson [1970].

$$W = \frac{1}{A}[(u_i)^A + B(u_j)^A], \quad A \neq 0,\ 0 < B \leq 1 \tag{4}$$

$$W = \ln(u_i) + B\ln(u_j), \quad A = 0,\ 0 < B \leq 1 \tag{5}$$

The parameter A determines the curvature of the iso-welfare loci, thereby reflecting the degree of aversion (or proneness) to inequality in the distribution of health state utility between individuals (or groups) i and j. In the case of a utilitarian SWF i.e. UB in Figure 5.1.1, A will take a value of 1 and for the Rawlsian case, UR, will be equal to negative infinity. Clearly, A will lie somewhere between these values for SWF (such as UC) that represent some trade-off between efficiency and equity. In the case of inequality proneness, as shown by UI in Figure 5.1.1, A will be greater than 1. The parameter B determines the steepness of the iso-welfare loci, thereby reflecting the weight given to individual j relative to individual i. In Figure 5.1.1, B=1 which means that in every respect other than health, i and j are considered to be equal. In the original Atkinson formulation, it is assumed that each individual is treated symmetrically (by assuming two anonymous individuals whose (different) income levels are all that is known). In this way, B forms no part of the original formulation.

Cobb-Douglas preferences

For illustrative purposes, let us assume that the SWF takes the log-linear form, such that A=0. In such circumstances, the SWF is analogous to a Cobb-Douglas (CD) utility function

$$U(u_i, u_j) = u_i^{\alpha} u_j^{(1-\alpha)} \tag{6}$$

$$U(u_i, u_j) = \alpha \ln u_i + (1-\alpha) \ln u_j \tag{7}$$

where a lies in the [0,1] interval. Thus, in terms of the formulation in () and (5), B = $(1-a)/a$. CD preferences are the standard example of indifference curves that look well-behaved; in terms of a SWF they imply an aversion to inequality.

In Figure 5.1.2, a is assumed to be 0.5 (i.e. B=1) which means that the same weight is given to individual i as to individual j, perhaps because in every other respect save health they are considered to be equal. Consider an initial point such as A which results in a health state utility of 0.4 for individual i and 0.2 for j. Suppose that, given resource constraints, it is only possible to treat one person. We can either improve the health of j from 0.2 to 0.4 (i.e. move to point B) or improve the health of i from 0.4 to 0.8 (i.e. move to point C). An individual (or society) with CD preferences which give the same weight to each individual will be indifferent between these two alternatives. This is because, given a=0.5, $U(u_i,u_j) = 0.5\ln u_i + 0.5\ln u_j$. We already have one point, B, on the indifference curve which yields a utility of $0.5\ln(0.4) + 0.5\ln(0.4)$ and we know that to be on the same indifference curve, that this must be equal to $0.5\ln u_i + 0.5\ln(0.2)$. Rearranging and solving for i gives I=0.8. Note that to be indifferent between the two alternatives, the health gain of the healthier individual, i,

(0.8-0.4=0.4) has to be greater than the health gain of the sicker individual, j, (0.4-0.2=0.2), representing a distributional consideration.

The convenience of assuming CD preferences is highlighted when a (linear) budget constraint is introduced into the model. It is then straightforward to show that the optimal choices that satisfy this type of SWF are

$$u_i = \alpha(\frac{m}{p_i}), \quad u_j = (1-\alpha)(\frac{m}{p_j}) \tag{8}$$

where m is the size of the total budget and p_i and p_j are the (constant) costs per unit of utility of treating i and j, respectively. Rearranging these formulae gives

$$u_i p_i = \alpha m, \quad u_j p_j = (1-\alpha)m \tag{9}$$

This means that a fixed fraction of the health care budget is spent on each individual. The size of the fraction is determined by the exponent in the CD function (i.e. by the value of B in the class of SWF specified in (4) and (5)).

Clearly, this property of CD preferences is useful when considering the distribution of health care expenditure between two individuals who are <u>not</u> considered to be equal. Indeed, previous research has shown that the general public may wish to weight more heavily the health needs of particular groups in society; for example, the young, those with children, and those who have looked after their own health (see Williams 1988 and Charny *et al* [1989]). If we assume that individual j is given greater weight than individual I, then $a<1-a$ (i.e. $B>1$) in the specification of the SWF. For example, if $a=0.2$ (hence $1-a=0.8$) then a CD SWF takes the form $U(u_i,u_j) = 0.2\ln u_i + 0.8\ln u_j$. Thus, we know that society wishes to allocate 20% of its health care budget to the treatment of individual i and 80% to the treatment of individual j.

Because social preferences (for fixed values of p_i and p_j) are a <u>linear</u> function of m, this will be true irrespective of the overall size of the health care budget.

We know from our earlier example (assuming a=0.5) that a health gain of 0.2 to individual j was equal to a health gain of 0.4 to individual i. For expositional purposes, assume that there are no diminishing returns in the treatment of individuals I and j; in other words, the relationship between costs and health gain is linear. Under such circumstances, for the budget line, or the utility possibility frontier (UPF), to pass through these two points requires the slope of the UPF to be -0.5 i.e. we can gain 0.5 units of health for j for every one unit of health we gain for i. At this rate of transformation a society with CD preferences would be indifferent between treating i and j. Were the gradient of the UPF to be steeper than -0.5 (for example, treating individual i becomes relatively more expensive), point B would be chosen, if it were flatter (for example, treating individual i becomes relatively cheaper), point C would be chosen.

Assuming a continuous UPF between B and C, means that we can maximise HRSW by doing something for both individual i <u>and</u> individual j. In health care, the UPF is most likely discrete rather than continuous i.e. we can either improve the health of one person <u>or</u> we can improve the health of the other, but in some cases a trade-off may exist. In our example, tangency between the SWF and the UPF is where i's health improves from 0.4 to 0.6 and j's health improves from 0.2 to 0.3, as shown by point D in Figure 5.1.3. As proof, we know that u_i=0.5m/1 and u_j=0.5m/2, so, if u_i=0.4 and u_j=0.4, we know that 1 is spent on i and 2 is spent on j, and if u_i=0.8 and u_j=0.2 we know that 1.6 is spent on i and 0.8 is spent on j. In other words, m=2.4. For tangency with a higher SWF, we know with CD preferences that the budget is spent in proportion to the exponents a and 1-a, which when a=0.5 means that 1.2 is spent on i and 1.2 on j. Substituting these values back into our original equations gives u_i=0.5*1.2=0.6 and u_j=(0.5*1.2)/2=0.3

Empirical investigation

In the absence of firm empirical evidence, it could be argued that a log-linear SWF (which, *ceteris paribus*, considers a health gain of 0.4 to an individual in a pre-treatment health state valued at 0.4 to be equivalent to a health gain of 0.2 to an individual in a pre-treatment health state valued at 0.2) is preferable to a utilitarian SWF which is concerned only with the maximisation of total health gain, irrespective of whether the person in the better or worse health state gets it. But if the premise that public sector decisions should reflect the strength of preferences of those who will be affected by those decisions is accepted, then it becomes an empirical question whether society is prepared to trade efficiency and equity against each other in this (or any other) way. And in principle, the framework presented in this chapter allows us to derive the precise shape of the SWF from responses to very simple questions.

Initially, the utility a respondent attaches to different states of health can be estimated using the SG or TTO. The x and y axes (i.e. the utilities of individuals i and j) can then be calibrated with health states that the respondent has valued for themselves. The respondent could be told that individuals x and y have preferences over health states that are identical to their own. They can then be asked questions along the lines of "if you had to choose between treating individual j who is in this health state (one the respondent valued at, say, 0.2) before treatment and this health state (one the respondent valued at, say, 0.4) after treatment or individual i who is in this health state (one they valued at, say, 0.4) before treatment and this health state (one the respondent valued at, say, 0.6) after treatment, which one would you choose to treat?. The final health state of individual i can then be made better or worse depending on whether the respondent chooses to treat j or i, respectively. In this way, the value of *A* can be estimated. The value of *B* could be estimated by stating that i and j differ according to a characteristic other than health; for example, age or sex.

In addition, if it is assumed that the respondent is not inequality prone such that their iso-welfare loci over the treatment of two individuals are not concave (i.e. that *A* is less than or equal to 1), then responses to these types of questions can also be used to test the validity of individual utilities. For example, if a respondent strictly prefers moving someone from a health state utility of 0.8 to full health to moving someone else from a health state utility of 0.1 to 0.4, then it is unlikely that the health state valuations elicited from this respondent can be treated as having interval scale properties with respect to health.

Alternatively, with respondents that are familiar with the concept of health status measurement and particularly with the notion that health states may lie on a continuum from full health to (or beyond) dead, it might be possible to present them directly with health state valuations rather than using health states that have an implied value. This was an approach taken with a convenience sample of 35 undergraduate students at the University of Newcastle. The students had taken an option in Health Economics and as such were familiar both with the concept of health status measurement and the techniques that can be used to elicit valuations.

In each of four seminars, 8 or 9 students, were asked to imagine that there are two individuals, i and j, who have preferences over health states identical to their own and are the same in all relevant respects except health. They were told that, at the moment, i is in a health state valued at 0.4 and j is in health state valued at 0.2. Respondents were then asked the following question: "Imagine that there is a treatment available which could move i to a health state valued at 0.6 or move j to a health state valued at 0.4. If you could only treat i or j but not both, who would you choose to treat?"

In total, one respondent preferred to treat i, stating that she felt it was better to have one person in a 'good' health state and one person in a 'bad' state rather than to have both in 'moderate' states. Two respondents were indifferent between treating i and j, stating that the health gain was the same in both cases. The remaining 32 respondents

said they would prefer to treat j because it is 'fairer'; either because j is initially in a worse health state or because the distribution of health after treating j is more equitable than after treating i. The 32 respondents who preferred to treat j were asked a second question: "Imagine that, as before, the treatment will move j from a health state valued at 0.2 to one valued at 0.4 but that the treatment will now move i from a health state valued at 0.4 to full health (i.e. valued at 1.0). If you again had to choose between treating i and j, who were choose to treat?

8 respondents still chose to treat j for both the reasons of 'fairness' cited above. 24 respondents now chose to treat i on the grounds that the benefit to i is now much larger than the benefit to j. The 24 respondents who now chose to treat i were asked one final question: "Imagine again that the treatment will move j from a health state valued at 0.2 to one valued at 0.4 but that the treatment will move i from a health state valued at 0.4 to one valued somewhere between 0.6 and 1.0. Where between 0.6 and 1.0 would the treatment have to move i to, so that you are indifferent between treating i and j?".

The responses were as follows: 0.65 =3; 0.70 =7; 0.75 =4; 0.80 =5; 0.85 =1; 0.90 =2; 0.95 =2. The median value for the full group of 35 respondents is 0.80 (inter-quartile range = 0.70-0.95). In other words, moving one person from a health state valued at 0.2 to one valued at 0.4, on average, yields the same social value as moving another person (who is identical in all respects except health) from a health state valued at 0.4 to one valued at 0.8. For this group of respondents, then, the log-linear HRSWF described in this chapter would be a good approximation of their preferences.

Discussion

The class of HRSWF that has been postulated here is sufficiently comprehensive to encompass a wide range of prescriptions concerning the distribution of (health state) utility. In this chapter, the HRSWF is characterised by two parameters; *A*, reflecting

aversion (or proneness) to inequality in the distribution of utility, and, B, reflecting the relative weight given to the treatment of different individuals or groups.

Particular attention has been given to the log-linear form (A=0) which implies inequality aversion and is analogous to a Cobb-Douglas utility function. This is largely for expositional purposes (CD preferences are well-behaved and have a number of useful properties) but partly because they enable us to represent various kinds of trade-off that society may be prepared to make between efficiency, in terms of health gain, and equity, in terms of severity of (pre- and post-treatment) illness and other relevant characteristics. In fact, the results from a preliminary experiment suggest that the log-linear form might indeed represent the average preferences of groups of respondents. Of course, these results should in no way be considered definitive but they do suggest that the approach is a feasible one.

CHAPTER 5.2: MEASURING EQUITY USING THE ATKINSON INDEX

Introduction

As with efficiency, where there are a number of ways in which health gain (or, more accurately, values for health states) can be measured, so to with equity. Whilst the approach presented in the previous chapter does indeed appear to be a feasible one, it is not the *only* approach. Since economists have typically concentrated on defining and measuring efficiency and have been remarkably silent in providing answers to the fundamental questions concerning the definition of equity and its relationship with efficiency (notable exceptions include Broome [1991], Sen [1982] and Sugden [1981]), it is necessary to explore different approaches.

In this chapter, the method first described by Atkinson [1970] in order to measure the shape of the SWF with respect to income distribution is used to allow the

shape of the SWF with respect to the distribution of health gain to be measured. The results of a questionnaire-based study are presented which, by attempting to calculate an Atkinson Index (see below), aimed to quantify the extent of the efficiency-equity trade-off. It is important to note here that one crucial modification to the Atkinson method needs to be made. In the original formulation, one unequally- distributed income was compared with another equally-distributed income but, whilst it is possible to transfer income between individuals, it is not possible to re-distribute health in the same way. Therefore, Atkinson's method has been applied to gains in health (which result in different distributions of prospective health outcomes) rather than to health *per se*.

The framework is illustrated in Figure 5.2.1. The axes represent the value of different health states to X and Y, again assuming that an agreed unit of value applicable to each individual has been established. Consider an initial situation, represented by point A, in which two individuals, X and Y, are in health states with the same value. Assume that under the current allocation of resources, it is possible to treat both X and Y so that point B can be reached. Although both individuals have benefited from treatment, it is clear that Y has benefited more. According to Atkinson [1970], there is a level of health gain, x, that equally distributed between X and Y, has the same social value as the unequally distributed gain associated with point B.

By drawing a perpendicular from B to the 45° line (resulting in point C) the same total health gain is yielded as at point B. These two points would yield the same social value for an individual who is inequality neutral (i.e. an individual who is concerned only with size of the total gain and not with how that gain is distributed). This implies a utilitarian SWF. For an individual who is inequality averse, such that they are prepared to accept a smaller total health gain than is implied by B in order that the health gain is equally distributed, the SWF would be convex to the origin and in the diagram cuts through the 45° line at point D.

Clearly, the more convex is this contour, the greater the inequality aversion. In the case of the Rawlsian SWF, the contour is parallel to the x- and y-axes such that point E yields the same social value as point B. For an inequality seeking individual, the SWF would be concave to the origin.

Using the framework developed by Atkinson, it is possible to derive an inequality index, I, which quantifies the extent to which a respondent weighs considerations about the size of the total health gain against considerations about how that health gain is distributed. The index is calculated as 1 - x/M. For an inequality neutral individual, $I = 0$ since $x = M$. For an inequality averse individual, $I > 0$ since $x < M$. The more total health gain such an individual is prepared to sacrifice in order for that gain to be equally distributed, the higher the value of I. For an individual with preferences akin to the Rawlsian SWF, I is at its upper boundary point. The precise boundary points on the Atkinson Index are defined by the initial distribution of health gain between the two individuals: when all the gain is initially going to one individual, a Rawlsian would forego the entire amount of health gain in order to achieve equity. In such circumstances, $I = 1$. The more equitable the initial distribution, the smaller is the range of values that I may take. This highlights the descriptive element embodied in the Atkinson Index (for a fuller discussion of this issue see Sen [1982]). For an individual who is inequality seeking $I < 0$ since $x > M$.

Methods

Valuation exercise

In order to introduce respondents to the notion that health states have a 'value', they were asked to complete a VAS exercise. Respondents were first given 5 EuroQol health states which represented a spread in terms of severity. They were then presented with a 100- point scale, the top of which was marked full health

(100) and the bottom was marked dead (zero) and asked to place the health states on this scale such that the distance between states represented their relative strength of preference for one state compared with another. In this way, respondents were required to consider what 5 points (assuming no ties) on the scale meant to them personally in terms of an associated health state and in relation to the endpoints of full health and dead. In the exercises to follow, they would be required to consider points on the valuation space other than those they themselves had identified with a particular health state. Thus, respondents were asked to imagine a health state classification system which was sufficiently sensitive to result in a continuum of values covering the entire space between full health and dead. It is possible to allow for states that are rated as worse than dead but for simplicity this questionnaire concentrated on states rated as better than dead only.

Equally distributed health gain question

The remainder of the questionnaire consisted of three sections, A, B and C which will be explained in detail below. However, the questions in all three sections followed the same basic format, each based upon 'equally distributed health gain' as described above. Throughout, respondents were asked to make choices concerning the health status of two individuals, X and Y, who were assumed to have preferences over health states identical to their own. Respondents were asked to suppose that X and Y are currently in health states to which they both attach the same value (as does the respondent herself). Further, they were asked to imagine that treatments are available which will be of benefit to both individuals but that the amount of benefit they each receive differs. They were told that, because of the way in which resources are currently allocated between the treatments of the two individuals, X would end up in a state valued lower than would Y. Respondents were then asked to suppose that resources could be re-allocated between the treatment of the two individuals in such a way that X and Y

would end up in the same health state. They were then asked to think about what value they would have to attach to this health state in order to make them indifferent between this common outcome and the different outcomes brought about by the current allocation of resources.

The different prospective health outcomes currently faced by X and Y (as illustrated in the appendix) were shown on the left hand side of the page. On the right hand side of the page respondents were presented with a range of possible values for a common outcome which could be brought about by a re-allocation of resources. Although it is possible that a respondent with the blind pursuit of equity as a goal may well prefer a common health state in which both individuals are worse off than under the current allocation, the range of SWFs was restricted to those in the Paretian class and hence ruled out such pathological functions. Similarly, responses where requiring both individuals to be better off in the common health state were ruled out. Thus, in each case, the right hand scale was bounded above and below by the different prospective health outcomes faced by Y and X under the current allocation of resources. Whatever these endpoints happened to be, the right hand scales were made equal in length and calibrated such that they were equally sensitive with respect to the calculation of the Atkinson index. Respondents were asked to place a tick next to that value of the common health state which would make them indifferent between both individuals ending up in that state and X and Y ending up in different states.

What respondents are being asked to consider is how much, if any, of the total current potential health gain they would be willing to sacrifice in order to achieve equality of health gain (and, because both X and Y start off in states with the same value, hence outcome). As indicated above, inequality neutral respondents will require the *same* total gain in both the inequitable and equitable outcome situations and hence will set the equally distributed gain equal to the mean gain. Those respondents who are inequality averse will accept less total gain in order to

achieve equality and thus will set the equally distributed gain *below* the mean gain. Inequality seeking respondents will require more total health gain in order to compensate for the re-distribution of that gain and hence would set the equally distributed gain *above* the mean gain.

Section A

Section A tested respondents' attitudes towards equity when the two individuals differed only with respect to the health gain they derive under the current allocation. Thus, respondents were asked to assume that X and Y are identical in every other respect and that both would live for 50 years and then die. The questions were designed to test whether attitudes towards equity were invariant with respect to the following:

1. The mean health gain.
2. The initial health status of the two individuals.
3. The distribution of the health gain.

One assumption implicit in the Constant Elasticity of Substitution (CES) class of SWF is that of constant (relative) inequality aversion. This implies that if the distribution of gains between one pair of individuals is simply a scaled up version of that between another, then a respondent will feel the same degree of inequality aversion to both situations. Thus, aversion to inequality may be assessed independently of the mean gain. On the other hand, the degree of aversion an individual feels for any given distribution may well depend upon the size of the total gain available as well as on the 'starting point' in terms of the current level of health. Although differences in degrees of inequality aversion across different distributions of gains cannot be measured directly by the Atkinson Index, it is important to test whether or not respondents were equally likely to be averse to inequality across two different distributions of gains. Thus, respondents were

asked a series of 6 questions using the starting points, mean gains and distributions indicated in Table 5.2.1.

Table 5.2.1 indicates that a comparison of the responses to QA1 with QA2 and QA3 with QA4 will test the sensitivity of the Atkinson Index when only the mean gain is allowed to vary. There are 3 possible comparisons which can be made in order to test starting point effects, namely, QA2 vs QA5, QA2 vs QA6 and QA5 vs QA6. Differences in each pair of responses are tested using a Wilcoxon sign test at, given the sample sizes outlined below, the 10% level of significance. In addition, a comparison of responses to QA1 with QA3 and QA2 with QA4 isolates the effects of changing the distribution of the gains. As indicated above, changing the distribution of gains changes the endpoints of the Atkinson Index, making comparisons of the magnitude of I between different distributions problematic. Thus, only the direction of individual responses with respect to attitudes towards equity (using a chi-square test) will be compared in these cases.

Sections B and C

As noted in Chapter 5.1, previous studies have indicated that respondents are willing to prioritise treatment between groups on the grounds of certain non-health characteristics. To test whether these factors would influence the trade-offs respondents were willing to make for the sake of equity, in Sections B and C, X and Y were no longer to be considered identical. All questions were of the same format used in QA5 and thus a comparison between QA5 and the responses to Sections B and C will indicate the extent to which each non-health characteristic influences attitudes towards equity.

In Section B, respondents were again asked to assume that both individuals would live for 50 years and then die but now they were no longer to be considered identical in every respect other than the benefit they received from treatment. In

QB1, respondents were told that X was a smoker whilst Y had never smoked. It was hypothesised that respondents would now be less willing to give up health gain in order to achieve equity than when the behaviour of both individuals was identical. Thus, we would expect the Atkinson index to be lower than in QA5. In QB2, X had no children whilst Y had dependent children. It was hypothesised that respondents would be less willing to give up total gain in order to re-distribute some of Y's potential gain to X than when their family circumstances were identical. Again, we would expect the Atkinson index to be lower than in QA5. In QB3, X was from social class 5 whilst Y was from social class 1. Were respondents to be concerned about equity with respect to social class, then they ought to be more willing to give up some total health gain than when the respondents were from the same social class. Under these circumstances, we would expect the Atkinson Index to be higher than in QA5.

In Section C, respondents were asked to assume that the individuals were different ages. In order to make one of the individuals 60 years old, it was no longer plausible for both individuals to live for 50 years. Therefore, the questions in Section C asked respondents to assume that the individuals would live for only 20 years and then die. Respondents were first told that individual X is 60 years old whilst Y is a 25 year-old. Under these circumstances it was hypothesised that respondents would be less willing to give up total health gain in order to re-distribute some of Y's potential gain to X than when they had been the same age. Thus, we would expect a lower Atkinson Index in QC1 than in QA5. The second question did not give rise to any a priori expectations; X being a 25 years old and Y a 5 year-old.

The Sample

The sample comprised of 23 of the 35 undergraduates used in the experiment reported in Chapter 5.1 and 14 students on an MA in health service studies at the

University of Leeds. The results are presented for the entire sample since there are no significant differences in responses according to gender or whether the respondent was in the Newcastle or Leeds group. The mean age of respondents is 25 and the sample is made up of 18 men and 19 women. Five of the 37 respondents smoke and 5 have children. There was no missing data.

Results

Section A

Table 5.2.2 gives the results from Section A, both in terms of the equally distributed gain responses and the corresponding value of the Atkinson Index. The results indicate that, at the aggregate level, there is no tendency to trade-off total health gain for the sake of equity. Whilst median responses to each question indicate inequality neutrality, the mean responses to five of the six questions in this section suggest very slight inequality proneness. Thus, it appears that, if anything, respondents require more total health gain when this gain is distributed equitably than is available under the inequitable distribution. Table 5.2.3 shows the pattern of responses at the individual level and highlights the relatively large number of respondents whose responses suggest they are inequality seeking. The following pattern emerges when individual responses are analysed across the 5 questions: 6 respondents were inequality seeking throughout, 11 were either inequality seeking or neutral, 5 were either inequality averse or neutral whilst 10 exhibited both inequality seeking and averse behaviour within their set of responses. No respondent displayed inequality aversion throughout their set of responses.

As indicated above, this section was designed to test whether attitudes towards equity were invariant with respect to a number of variables. Allowing only the mean gain to vary, responses to QA1 and QA2 are statistically significantly different (p-value < 10%) from one another as are those to QA3 and QA4 (p-value

< 5%). In each case the index is higher in the first question in the pair reflecting more aversion to inequality (or at least less inequality seeking), the higher the mean gain. In the tests for 'starting point' effects, two of the three comparisons (QA2 vs QA6 and QA5 vs QA6 but not QA2 vs QA5) were statistically significantly different from one another at the 5% level. In each case the index is higher, the higher the current health status of the two individuals.

Thus, there seems to be at least some evidence to suggest that equity may be more of a concern at the 'top end' of the scale when both individuals are in a relatively good health state, than at the 'bottom end' when both are in health states to which low values are attached. This is most clearly highlighted by the fact that the only question which generates a positive mean Atkinson Index value (i.e. implying inequality aversion) is QA6 where both individuals start off in health states valued at 76 and one of them returns to full health. Of course, this finding contradicts the assumption of constant proportional inequality aversion.

A chi-square test found no statistically significant differences at the 10% level in the proportion of responses displaying inequality aversion between QA1 and QA3 and between QA2 and QA4, where only the distribution of the health gains differed. Thus, it would appear that the willingness of respondents to give up some total health gain in order to achieve equity is unrelated to the degree of inequity which exists in the distribution of gains under the current allocation of resources.

Section B

The results in Table 5.5.4 show that inequality seeking predominates when X and Y are no longer considered identical in every respect other than the benefit they receive from treatment. The Atkinson indices for QB1 and QB2 are significantly lower than for QA5, (sign test p-value < 10% in both cases). The median

responses suggest that an additional 27% of total health gain is required in order to compensate for re-distributing some health gain from a non-smoker to a smoker or from somebody with children to somebody without. Responses to QB3 are not significantly different from those to QA5, indicating that no additional weight is given to equity with respect to social class. The significantly lower index for QC1 than for QA5 (sign test p-value < 10%) indicates a greater reluctance to re-distribute health gain from the 25 year-old to the 60 year-old than when both individuals are the same age. Although the effect is not so marked in QC2, where X was a 25 year-old and Y was a 5 year-old, the Atkinson Index was lower than in QA5 (sign test p-value < 5%) indicating that respondents were again showing a preference for the younger of the two individuals.

Discussion

The purpose of this chapter has been to test whether Atkinson's equally distributed income model could be modified to measure attitudes towards inequality in the distribution of health gain. This study has shown that the questions asked posed few problems for a relatively educated sample although the true test of feasibility will come when questions of this kind are asked of a more representative sample of the population.

The results themselves suggest that when two individuals differ only with respect to the benefit they derive from treatment, respondents are, on average, indifferent between an allocation of resources such that health gain is unequally distributed between the two individuals and one in which that same health gain is distributed equally. This suggests that on the whole respondents are inequality neutral: they are concerned with the size of the gain and not with how that gain is distributed.

Although this result was robust across different distributions in the 'unequal' state of the world, statistically significantly different results were obtained when the

mean gain and/or the initial health status of the two individuals were allowed to differ. This finding casts doubt on a crucial assumption in the CES class of SWF; namely, that of constant (relative) inequality aversion. Specifically, the results suggest that there is greater aversion to inequality when both individuals (even the one who gains least in the 'unequal' state of the world) end up in health states that have values closer to full health than to dead. In other words, it would seem invalid to assume that the same degree of inequality aversion applies to two states of the world where one is simply a scaled up version of the other.

This suggests that distributional considerations may be given greater weight when both individuals can achieve a "decent" level of health (however defined), no matter how the benefits are allocated. Similarly, less weight may be accorded to an equal distribution of health gain if, in bringing this about, neither individual achieves this "decent" level of health. In this way, it might be that equity can be considered to be a "luxury" good i.e. concern for it is an increasing function of overall (or average) health status. There is certainly the need for more research here, perhaps looking at attitudes to equity in countries with low levels of overall health (however defined) compared to those with higher levels of overall health. A separate, though not unrelated issue, is whether or not it is appropriate to assume that preferences elicited over the treatment of two individuals can be used to infer preferences over the treatment of two groups of individuals. Future research effort could therefore also be directed towards considering such "scale effects".

A number of questions were also designed to test how attitudes to equity might vary with respect to certain non-health characteristics. The results from these questions were largely as expected: relative to their preferences when two individuals differ only with respect to their benefit from treatment, respondents are reluctant to re-distribute health gain from a non-smoker to a smoker; from an individual with children to one without; and from a younger to an older individual.

However, the most striking and unexpected feature of the results is the general apparent lack of aversion to inequality in any of the questions. This runs counter to the findings presented in Chapter 5.1 and is particularly striking given the considerable overlap between the two groups of respondents. Therefore, the apparently different attitudes towards inequality in this study compared to the one reported in Chapter 5.1 must be due to something other than the different samples used. Since it is now widely recognised that the way in which a question is framed can have a significant effect on responses (and this has been shown elsewhere in this thesis), the differences may lie in the nature of the tasks respondents are asked to do in each case. Specifically, the rather different attitudes uncovered here may be due to the fact that reference point effects, first introduced in Chapter 2.1, were playing a significant role. Although reference points were originally developed to describe individual decision-making, there is no reason in principle why it should not be used to describe an individual's preferences when they are asked to place themselves in the position of a social decision-maker.

Of course, the questions in this study do not involve any actual losses; X and Y each receive a gain in health status from their initial position (point A in Figure 5.2.1) and this was stated explicitly in the instructions to all respondents. However, it seems plausible that certain respondents may have adopted the potential gains available under the current allocation of resources (point B in Figure 5.2.1) as their reference point, in which case any redistribution necessarily involves a 'loss' to individual Y. Indeed, when respondents were later presented with, and invited to discuss, the results that they had collectively generated, it became apparent that for many loss aversion had played a significant role in their responses.

In the analysis above it was assumed that Y's potential losses were weighted equally to X's potential gains and the inequality neutral position was taken to be that point at which the equally distributed gain, x, was equivalent to the mean gain,

M. Consider QA1 as an example. This question sought to elicit from respondents that value for x which would set U(24,96) equal to U(x,x), where 24 and 96 are the potential gains available to X and Y respectively under the current allocation of resources. If the loss (of potential gain) to Y is weighted equally to the gain (of potential gain) to X then U(96-x) = U(x-24), and an inequality neutral respondent would set x equal to 60. Therefore, U (24,96) = U (60,60) for such a respondent. However, when losses are weighted more heavily than gains, an inequality neutral respondent would set x above M. Suppose, for example, that Y's losses were weighted twice as heavily as X's gains (dealing with wealth, Kahneman and Tversky [1979] suggest a loss-to-gains slope ratio of 2:1 whilst Fishburn and Kochenberger [1979] estimated the relationship empirically and found it to be closer to 5:1). In this case, 2U(96-x) = U(x-24) and x must now take on a value of 72. Thus, U(24,96) = U(72,72) for an inequality neutral respondent. The inequality neutral level of x has shifted up from a point halfway between the initial distribution of gains to X and Y to a point two-thirds of the way between the two. This holds for each of the questions in section A and for all other situations in which losses are weighted twice as heavily as gains.

Moving the inequality neutral level of x up in this manner results in a positive (mean and median) Atkinson Index for all questions in Section A and in more respondents appearing to be inequality averse ($I > 0$) and fewer being inequality seeking ($I < 0$). Tables 5.2.5 and 5.2.6 are the analogue of Tables 5.2.2 and 5.2.3 when losses (of potential gain) are assumed to be weighted twice as heavily as gains (of potential gain). Whilst this is purely illustrative, it does highlight that the failure to take account of aversion to losses will necessarily underestimate the extent of any aversion to inequality.

This analysis assumes the loss-to-gains slope ratio to be constant over the entire valuation space. In other words, the extent to which losses are weighted more

heavily than gains is considered to be the same at the top end of the scale as at the bottom. However, it is conceivable that the magnitude of this ratio may depend upon where in the valuation space the losses and gains occur. This offers an alternative explanation of the finding that there is a greater aversion to inequality when the individuals concerned end up in health states that have values closer to full health than to dead. In fact, the same pattern of responses would be generated by a respondents whose aversion to inequality remains constant over the entire scale but who weights losses and gains more equally at the top end than at the bottom. When designing this study, we were was insufficiently alert to the potentially important role that a reference point effect might play in generating the results. Therefore, it is impossible from responses to this questionnaire to disentangle the effect of inequality aversion and the contradictory effect (in terms of its impact on responses) of loss aversion.

The question that arises is whether or not it is appropriate to incorporate loss aversion (resulting from a reference point effect) into the HRSWF. It could be argued that loss aversion (even aversion to potential losses) is a legitimate factor which ought to be incorporated into a SWF since any real re-allocation of resources in order to promote equity will of course deny some potential health gain to the group or groups who stand to do better under the current allocation. There is no obvious reason why one and not the other of these two factors ought to be incorporated into a SWF.

However, any aversion to potential losses at the societal level will at the margin result in a pressure to maintain the status quo, presumably no matter how inequitable this status quo is. Therefore, loss aversion (which, in this case, works in precisely the opposite direction to inequality aversion) is likely to perpetuate any current or past (mis)allocation of resources. It is questionable whether we would want to take account of such a tendency. Moreover, it can be easily shown that having differential weights for potential gains and losses can result in

intransitive SWFs. It might be argued, therefore, that questions designed to formulate the SWF should control out the possible effects of a status quo bias. This might be achieved through presenting respondents with a series of pairwise choices (one where health is equally distributed, the other where it is not) and asking them to choose between the two outcomes without specifying which one is currently available. This author hopes to address these issues in future work.

CHAPTER 5.3: USING THE PERSON TRADE-OFF APPROACH

Introduction

Chapters 5.1 and 5.2 have discussed how equity might be taken account of in resource allocation decisions, but only after an efficient allocation has been determined. In other words, valuations for health states, aggregated according to the mean or median response, are used to determine the maximum amount of benefit (defined in terms of QALYs) that can be accorded given a fixed budget. If social preferences over different allocations of the health care budget can be represented by a lexicographic ordering in which health gain is the dominant argument in the HRSWF, then we need proceed no further. Given that distributional considerations, for example, are also likely to matter, the previous two chapters develop a framework in the simple QALY-maximisation approach can be adapted to take account of such considerations.

An alternative approach is proposed by Nord [1995] who argues that "weights for life years in the QALY procedure should not be derived by asking individuals to value health states *for themselves* .. [but] .. should ultimately reflect responses to person trade-off questions asked in a resource allocation context" (p202). The person trade-off (PTO) approach involves asking respondents how many outcomes of one kind they consider equivalent in terms of social value to X outcomes of another kind. The method was originally developed by Patrick *et al* [1973] who called it the 'equivalence

of numbers' procedure but Nord's terminology is used in this chapter. It is Nord's contention that PTO responses capture concerns that are relevant to social decision-making, most notably considerations about the initial severity of illness (see Nord [1994]).

This chapter reports on a pilot study which was designed to assess whether two treatments that yield the same benefit to the individual are also considered, by that individual, to yield the same benefit to society (as measured by responses to PTO questions). This is a unique approach because it is the first time that the two options in a PTO question have been chosen only <u>after</u> it has first been established that they yield the same benefit to the individual respondent. Hitherto, studies have used health states descriptions to generate PTO scenarios that are assumed to yield the same benefit for each successive move between levels of severity (for example, see Nord [1993]).

The null hypothesis in this chapter is that two treatments that have the same individual utility will also have the same social value. That is, the respondent will only be indifferent between the two treatments when the numbers receiving each treatment are identical. The alternative hypothesis is that, of two treatments with equal individual utility, the treatment with the greatest initial severity of illness will yield the greatest social value. That is, the respondent will be indifferent between the two treatments when fewer people receive the treatment with the greatest initial severity. This hypothesis is developed from results obtained by Nord [1993] which suggest that (for a Norwegian population) pre-treatment severity is a more important explanatory variable in PTO responses than treatment effect.

Study design

Respondents were presented with 15 cards, representing 14 EuroQol states (chosen to give a spread in terms of severity) plus 'Immediate Death' and were told that each state (except Immediate Death) would last for 10 years without any change after which they

would die. They were asked to sort and rank the cards so that the one they thought best was at the top and the one they thought worst was at the bottom, and to place any they felt were the same alongside each other.

Respondents were then asked to imagine that they were in state 22323 and that a treatment (referred to as T1) was available which would move them to state 12223, thus offering improved mobility and improved ability to perform usual activities. They were then asked to imagine instead that they were in state 22222 and another treatment (referred to as T2) was available which would move them to a state which they had ranked above 22222. These states were chosen for the two treatments so as they were *a priori* plausible treatment possibilities. The aim of what followed was to identify the state that when moved to from 22222 yields the same benefit as the move in T1 from 22323 to 12223. Initially T2 was defined as the move from 22222 to the state ranked 4th by the respondent. If they found this equally as good as T1, this card was taken as the `interval' state. If T2 was found to be better (worse) than T1, an iterative process was conducted using lower (higher) ranked cards until an interval state was identified. Where the respondent was unable to identify an interval state two states were identified: one referred to as BIS (a 'better than' interval state), which when moved to from 22222 was considered to yield more benefit than T1; and one referred to as WIS (a 'worse than' interval state), which when moved to from 22222 was considered to yield less benefit than T1.

Respondents then rated 8 states on a VAS with endpoints of 100 (best imaginable state) and 0 (worst imaginable state) and then valued 5 states using TTO method which asked them to compare 10 years in each of the states with shorter periods of time in full health. Neither the protocol for these methods nor the results from them are reported in detail here because the focus of the chapter is to compare the results from the comparison of treatments with those from the PTO, which followed the VAS and TTO tasks.

A specially designed board was used to present the PTO task (see appendix). Respondents were asked to imagine that there were 10 people who would spend 10 years in state 22323 (the initial state for T1), after which they would die, and that there were another 10 people who would spend 10 years in state 22222 (the initial state for T2), after which they would die. They were asked to imagine that the two groups were the same in every way they conside red relevant. Respondents were then presented with two possible treatments. Treatment 1 would move the 10 people in state 22323 to state 12223 for 10 years after which they would die. Treatment 2 would move the 10 people in state 22222 to the interval state for 10 years after which they would die. In the absence of an interval state the WIS was used wherever possible but where this was not possible (for example, in those cases where no WIS was identified) BIS was used.

Respondents were told that only one of the treatments could be provided and were asked if they would choose treatment 1, treatment 2 or whether they would not mind which one was chosen. If the respondent chose treatment 1, further questions were asked in which the number who would benefit from treatment 2 remained unchanged but the number who would benefit from treatment 1 was changed; initially to 5 and then, by an iterative process using smaller or greater numbers of people, until a point of indifference with treatment 2 was reached. Similarly, if the respondent chose treatment 2, the number who would benefit from treatment 1 remained unchanged but the number who would benefit from treatment 2 was changed until a point of indifference with treatment 1 was reached. All respondents were asked to articulate why they made the choices they did. Specifically, they were asked "Could you please tell me why you made this choice?". At the end of the interview socio-economic data were collected.

Interviews were conducted over a two-week period in late July/early August 1995. The sample was a convenience sample drawn from secretarial, administrative and academic staff of various departments at The University of Newcastle-Upon-Tyne.

Each respondent was paid £10 for participating. All interviews, with the consent of the respondents, were tape-recorded.

Results

Descriptive analysis

In total, 28 interviews were conducted. Table 5.3.1 shows the background characteristics of the respondents. As expected, the sample were more qualified and younger than a representative sample of the general public would have been. Table 5.3.2 details the states used in the study together with their mean rank. The ranking given by each respondent was analysed for any inconsistency, defined as the case where, in a comparison of two states, the state which is logically better on at least one dimension and no worse on the other dimensions is ranked below the other. The ranking data were highly consistent with 16 people having no inconsistencies at all and a further 8 having only one. Those inconsistencies that did exist appeared to be unrelated to a particular state, although as expected most occurred between the comparison of two states that were close together in the logical structure.

Table 5.3.3 shows the state that was identified as the interval state (or BIS and/or WIS) for each respondent. The most frequently chosen states are 21222, which is chosen as the interval state 4 times and as BIS 6 times, and 11122, which is chosen as the interval state 4 times and as BIS or WIS 3 times. Given that each respondent's ranking of the health states was central to the identification of an interval state, the ranking data was checked for any obvious patterns of response. For example, it was checked whether one particular ranking was chosen most frequently as the interval state and whether the difference in rankings between the two states in T1 influenced the choice of the interval state in T2. No obvious patterns were apparent from this analysis.

Comparing individual preferences with PTO responses

In the comparison of treatments exercise, each respondent identified two treatments (T1 and T2) that for them yielded the same benefit. Therefore, if an individual feels the same about the treatment of other people as they do about their own treatment, then their PTO response should be that they are indifferent between 10 people receiving T1 and 10 people receiving T2. In this case, the ratio of the number of people receiving T2 that is considered equivalent to the number of people receiving T1 would be 1. A ratio >1 implies that, in the treatment of other people, the respondent prefers T1 to T2, whilst a ratio <1 implies that the respondent prefers T2 to T1.

The PTO responses and resulting ratios for each respondent are shown in Table 5.3.4. It can be seen that only 3 respondents consider that 10 people receiving T1 yields the same social value as 10 people receiving T2. Of the remaining 24 respondents who answered the PTO question (the one non-response is discussed below), 7 prefer T1 to T2 whilst 17 prefer T2 to T1. The proportion preferring T2 is statistically significantly different from the proportion preferring T1 ($p<0.01$). In aggregate, the preference for T2 is reflected by a geometric mean of 0.7 which suggests that, on average, 10 people receiving T1 yields the same social value as 7 people receiving T2.

The analysis of whether PTO responses are affected by the background characteristics of the respondent is clearly limited by the relatively small number of respondents in the study. However, preferences over T1 and T2 do not appear to be systematically related to the age or sex of the respondent, nor, importantly, do they appear to be related to the choice of the interval state.

Qualitative responses from the PTO exercise

Table 5.3.5 gives an interpretation of the comments made by respondents in the PTO task. Of the 7 respondents who preferred T1, 5 commented on the fact that they

would prefer to help those who were worse-off; for example, No. 24 said "*The group which needs treatment 1 is the group suffering the most*". The other 2 respondents who preferred T1 stated that they felt T1 gave a greater improvement in health. Of the 17 respondents who preferred T2, 12 commented that T1 did not offer a great deal of benefit to those receiving it; for example, No. 15 commented "*I'm thinking about quality of life .. I would say that treatment 2 was a definite improvement .. in treatment 1 the difference isn't so great*".

In thinking about their PTO response, 3 respondents (all of whom preferred T2) mentioned that they were considering the implications of their choice for society as a whole (for example, the costs of keeping people in a particular health state). When faced with a decision involving an unequal number of people being treated, 16 (out of 24) respondents referred to the numbers of people involved, with 8 seemingly using the numbers involved as a decision variable. It can be seen from Table 5.3.5 that there was one non-response to the PTO question (No. 13). This respondent refused to take part in the exercise on the grounds that decisions about which patients should be given priority was one that the public should not be asked to make. Instead, those with more knowledge and experience in making such decisions should decide.

Discussion

One of the main aims of this study was to test whether two treatments that are considered equivalent by an individual are also considered equivalent when the same individual has to make choices about the treatment of other people. If respondents' preferences change when they are asked to imagine themselves in the role of social, rather than individual, decision-maker, then, it is questionable whether social preferences can be accurately represented by the unweighted summation of individual utilities.

When asked to compare two treatments using the PTO method, which Nord recommends as one way of measuring social value, all except 3 respondents strictly prefer one treatment to the other, despite the fact that the two treatments generate the same utility (according to the comparison of treatments exercise) for respondents themselves. This suggests that social choices are indeed considered differently to individual ones. Thus, the null hypothesis that two treatments that have the same individual utility will also have the same social value is rejected by the results from this study. This indicates that unadjusted individual utilities cannot be used to represent social values.

However, the data does not appear to support the alternative hypothesis either; that the treatment with the greatest initial severity of illness will yield the greatest social value. Table 5.3.4 shows that more than twice as many respondents prefer T2 to T1 than prefer T1 to T2. Since T2 involves the move between two states that might be regarded as higher up the utility scale and T1 is the treatment with the greatest initial severity, these results do not suggest that people are more concerned about doing something for those in severe states than they are about doing the same (in terms of health gain) for those in less severe states. Clearly, initial severity does matter to some respondents (for example, 5 of the 7 who chose T1 mentioned this as an important consideration) but the results presented here suggest that, for the majority of respondents, it is better to give a benefit of the same magnitude to someone who is in a better health state to start off with.

This discussion has been premised on the notion that indifference between the two treatments, established very early in the interview, is maintained throughout. However, it is possible, as respondents proceed through the interview and become more familiar with the health states, that their preferences may become more refined. Thus, a comparison between the comparison of treatments stage near the beginning of the interview and PTO responses near the end of the interview is not unproblematic. Indeed, the process by which an interval state is selected remains a crucial

consideration for any future study which attempts to address the issues raised in this chapter. In this study, the interval state was determined before valuations were elicited so that respondent burden in subsequent valuation tasks could be minimised. Selecting the interval state after valuations had been elicited might have reduced the likelihood that a 'wrong' choice was made through inexperience but might also have increased the likelihood that a 'wrong' choice was made through fatigue since many more valuations would need to be elicited.

That 14 respondents (2 who preferred T1 and 12 who preferred T2) indicated that their choice was based on the fact that one treatment yielded more (or less) benefit than the other, might indicate that preferences changed as the interview progressed. Whilst this possibility cannot be discounted, it is expected that 'preference refinement' of this kind would result in T1 being preferred to T2 roughly as often as T2 is preferred to T1, which is clearly not the case. Thus, it is puzzling why so many more respondents should explicitly state (in the PTO exercise) that T1 does not confer as much benefit as T2 rather than vice versa, particularly as the move in T2 for 8 of these 12 respondents is to a state (21222 or 22112) that might be considered to be very close to the initial state (22222).

If perceptions of T1 relative to T2 change as respondents progress through the interview because their own *individual* preferences over T1 and T2 change, it is difficult to make inferences about the acceptability of either the null hypothesis or the alternative hypothesis since both require that the benefits from T1 and T2 are considered to be equivalent. However, if perceptions of T1 relative to T2 change because the context of the choice that the respondent faces changes (i.e. they are initially faced with a choice about their own treatment and then with a choice about the treatment of other people), then the null hypothesis can *de facto* be rejected but the alternative hypothesis cannot.

Whilst, then, the alternative hypothesis cannot be rejected given the results from the qualitative data, these data do give some more general indications about the degree to which subjects emphasise benefit of treatment relative to initial severity of illness. Moreover, T1 does confer some benefit on those receiving it (they can move from 22323 to 12223). Therefore, it would appear that more respondents in this study focused on the benefits from the two treatments (whether they were considered to yield the same benefit or not) than on the severity of the pre-treatment health state, which all respondents considered to be worse in T1 than in T2.

That T2 was preferred by more respondents than T1 in the PTO exercise might suggest that the health state that people are in after treatment is also an important consideration. Even after a beneficial treatment, people who receive T1 are still left in a relatively severe health state; they are still extremely anxious or depressed, for example. It may be that unless somebody in a severe state can benefit more substantially from treatment, respondents would rather give a benefit of a similar magnitude to someone who is in a less severe pre-treatment state. It is difficult to tell from the qualitative data the extent to which such considerations were taken into account by respondents but it is another way (in addition to considerations about health gain) in which comments to the effect of "T1 doesn't do much good" could be interpreted.

If considerations of this kind are important, then the relationship between health gain and severity may be a rather more complex one than hitherto suggested. For example, consider the (highly-stylised) scenario where a given health gain can only be afforded to one of three patients, x, y or z, such that $h_x > h_y > h_z$ where h_i is the initial health state of patient i. Rather than being indifferent about who should receive treatment (as would be the case if maximising health gain were the sole objective) and rather than having a preference ordering such that $z \succ y \succ x$ where '$\succ$' indicates 'preferred to' (as would be the case if those with a greater initial severity were given priority), it might be that social preferences would imply an ordering such that $y \succ z \succ x$. In other words,

there might be a non-monotonic relationship between efficiency and equity in the HRSWF. Results rather similar to this were reported in the previous chapter, where it was noted that there was apparently greater aversion to inequality when both individuals end up in health states that have values closer to full health than to dead. The possibility, then, that equity might be considered a "luxury" good appears to be a strong one.

CHAPTER 6: AN AGENDA FOR FUTURE RESEARCH

Introduction

In this chapter, three important themes which have emerged from this thesis will be discussed in the context of how they may provide a general strategy for future research into the measurement and valuation of health. Specific suggestions for future research, relating to particular topic areas, can be found in the discussion sections of the preceding chapters.

The nature of individual preferences

Since the majority of this thesis has been centred around a number of empirical studies which have attempted in various ways to elicit individual preferences, it is worth considering the nature of these preferences. The received wisdom amongst economists is that individuals have clear, well-defined preference functions which can be 'tapped into' by appropriate questions. This viewpoint is referred to by Fischhoff [1991] as the philosophy of articulated values and is summed up by the notion that "if we've got questions, then they've got answers" (p835). An implication of this viewpoint is that if a particular respondent gives different answers to two questions, then implicitly the questions must have been different. Proponents of this paradigm focus on ensuring that questions are formulated and understood as intended, arguing that any 'slip' could invoke a precise, thoughtful answer to a "wrong" question.

However, this paradigm has been called into question by many studies which have shown that seemingly subtle changes in problem structure, question format, or other aspects of the assessment process, can sometimes dramatically change the stated preferences of respondents (for good examples of this, see Kahneman and Tversky [1981] and Fischhoff

and Furby [1988]). This thesis contains a number of examples of possible framing effects of this kind. Perhaps the most obvious can be found in the study reported in Chapter 2 where, contrary to expectations, greater discrepancies in valuations were found *across* different variants of the SG and TTO rather than between the *methods* themselves. And the results of the study reported in Chapter 4.2 suggested that the way in which the TTO questions were framed resulted in one month in a particular dysfunctional state appearing as worse than ten years in that same state.

Such findings can be accounted for by an alternative paradigm - referred to by Fischhoff [1991] as the philosophy of basic values - which asserts that people cannot be expected to have articulated opinions on more than a small set of issues (of which health is unlikely to be one) with which they are very familiar. In complex or unfamiliar decision problems, Slovic [1995] argues that "preferences are not simply read off some master list but are constructed on the spot by an adaptive decision maker" (p369). Thus, if responses are affected by superficial changes in question formulation, then respondents must not have 'true' underlying preferences; rather, the elicitation procedures are major forces in shaping stated preferences. But arguably there are examples in this thesis where this paradigm is also questionable. For example, in the study reported in Chapter 2, valuations appeared to be unaffected by the order of presentation of the tasks and throughout the thesis there is evidence that the ordinal rankings of health states are robust to changes in value elicitation procedure.

A philosophy of partial perspectives lies somewhere between the extremes of articulated and basic values. This viewpoint holds that, whilst preferences, particularly regarding health, do not come as fully fledged and instantly accessible as economists typically believe (or at least believed), people in very general terms do have what Fischhoff refers to as "stable values of moderate complexity" (p836). Such a viewpoint would suggest that elicitation procedures can help to shape preferences but also that, after deliberation and

reflection, respondents are able to give answers to questions that enable us to infer something about their 'true' preferences.

The prevailing philosophical framework upon which the studies reported in this thesis have been based is this partial perspective. As with the philosophy of articulated values, it is crucial when adopting a partial perspective that respondents interpret questions in the way that they are intended to be interpreted and that the context of the question is plausible to respondents. In addition, respondents must be given time and opportunity to think about what is being asked of them. The sections in Chapters 2.1 and 3.1 which described the designs of the pilot and main studies, respectively, provide only a snapshot of the considerable time, effort and resources that went into formulating questions that were likely to be considered meaningful by respondents yet also isolated from factors (such as financial considerations) that were considered irrelevant and extraneous.

In the context of the Main Study, an important objective was that respondents should become as familiar as possible with the EuroQol health states before being asked to value them using the TTO (subject to the constraint that each interview should not last more than about one hour). This familiarisation was crucial since respondents were unlikely to have previously thought about health in the way they were being asked to in the interview. For these reasons, all respondents were presented with ranking and rating tasks, which in many cases took over half of the time for the whole interview, before TTO valuations were elicited. All the indications from the quantitative data (for example, in terms of logical consistency in the TTO responses) and from interviewer feedback were that respondents benefited from these 'warm-up' tasks.

Of course, the spectrum that the philosophy of partial perspectives covers is wide given the difference between the philosophies of articulated and basic values. The paradigm adopted in the work of the MVH Group was possibly closer on the continuum to the philosophy of articulated values than to the philosophy of basic values.. For example, in

the studies reported in Chapters 2-4, respondents were engaged in one (one-hour) interview during which time they were required to assimilate all the information provided to them and asked to give valuations to a number of health states, in many cases using a number of different valuation methods. Whilst the studies using convenience samples that were reported in Chapter 5 were principally aimed at testing the feasibility of different approaches to measuring social preferences, they too were essentially of the same format i.e. some introductory information and 'warm-up' exercises followed immediately by value elicitation procedures. But as Loomes [1994] suggests "many respondents cannot attach values to the quality of life entailed by states of health unfamiliar to them".

So perhaps future studies should begin by adopting a partial perspective that is closer to the philosophy of basic rather than articulated values, on the basis that people's preferences over states of health and illness are not very well developed but could be better constructed if they were provided with even greater opportunity to consider their responses. This might involve presenting respondents with a summary of all of their responses at the end of an interview and allowing them to revise any of their answers in the light of this 'overview', or even confronting them with any apparent inconsistencies and again giving them the opportunity to revise their responses. It might also be that respondents are better able to articulate something approximating a 'true' preference if they are given much more time for the deliberation and reflection alluded to above. This could involve more than one interview, possibly including a pre-interview focus group meeting in which respondents discuss issues relating to health and illness and a post-interview feedback meeting in which they review their responses.

Whatever their precise protocols, such studies would, of course, be much more resource intensive per respondent than the Main Study was but perhaps we have now reached the stage where more in-depth studies are necessary. Before the Main Study, most health status measurement studies had been conducted with small convenience samples of respondents; about half of the studies using the VAS, SG or TTO have used less than 100

subjects, typically patients or students (see Froberg and Kane [1989]). This led many to question their generalisability, particularly economists brought up in a tradition of quantitative data (and, of course, a philosophy of articulated values). Therefore, the Main Study, with a representative sample of nearly 3500 members of the UK population, was a necessary response to this. Now that such a large-scale study has been conducted, we might be more willing to trade-off quantitative data for the more detailed qualitative data that intensive questioning would generate.

This qualitative data should provide insights into the cognitive processes that respondents use in order to arrive at their responses, thus enabling researchers to get a better understanding of *why* valuations differ in addition to *how* they differ. Many of the empirical chapters in this thesis have been written in a way typical of an economist; namely to postulate a null hypothesis, to then collect quantitative data that tests the hypothesis, and finally to engage in considerable 'post-hoc' theorising when the results, as invariably happens, do not conform with the null hypothesis. Rather than 'second guessing' respondents, the collection of qualitative data "straight from the horse's mouth" appears a more appropriate strategy in this context. Such data might then help us get a clearer picture of why valuations appear to be as much a function of the way in which a method is administered as they are of the method itself, why some respondents are unwilling to trade-off any time in order to avoid poor states of health, why valuations differ according to the background characteristics of the respondent, and so on.

In addition, it is now widely recognised that the specific context of a particular choice can have an significant effect on a respondent's stated preference. For example, in Chapters 2.1 and 5.2, particular patterns of responses have been explained in terms of perceived reference points. Of course, such an effect has no place in standard economic theory which is built around an EU framework. The only considerations that yield utility in such a framework are the outcomes resulting from the particular choice, thus ruling out any utility associated with the process of that choice. However, a number of alternatives to EU have

been proposed by economists which allow for the context of the choice to play a role in determining the overall level of utility. Qualitative data can help shed light on which choice contexts are, at a descriptive level, considered relevant and which are not. And, as noted by Froberg and Kane [1989], "to predict new context effects we need to better understand the psychological processes inherent in decision making".

A more 'in-depth' approach to preference elicitation should also enhance our understanding of the extent to which observed disparities (for example, between different population sub-groups or between different ways questions have been framed) are "real" or "artefactual". For example, the results from the Main Study suggest that TTO valuations are primarily affected by the age and sex of the respondent (see Chapter 3.1). In an attempt to gain a better understanding of the causes for such differences, a more qualitative follow-up study was conducted in which 45 respondents who had taken part in the Main Study were re-interviewed using a protocol similar to that used in the Main Study but with fewer health states (see Robinson *et al* [1996]). Respondents were asked to 'think aloud' as they completed the interview, and to explain why they made certain decisions during the TTO exercise; for example, why they decided that a particular health state was better or worse than dead and, if applicable, why they were unwilling to sacrifice any life expectancy to avoid a particular state.

No compelling explanation was forthcoming as to why female respondents assign lower valuations than do male respondents, suggesting that differences in valuations according to gender are "real" differences. However, it emerged that older respondents were less likely than younger respondents to find the worse than dead scenario plausible. Whilst none of the fifteen respondent in the 18-39 age group questioned whether they would return to full health after a number of years in poor health, seven of the fourteen respondents in the 60 or over age group said they thought this was impossible. These seven respondents gave lower valuations for the states rated as worse than dead than the other seven respondents in their age group did, Thus, there is some evidence that the lower values elicited from

older respondents might be attributable to this particular artefact. Without this qualitative follow-up study, this possibility would not have been brought to our attention.

Although the collection of qualitative data was not a primary objective of the study designed to elicit attitudes towards inequality using the Atkinson method, when respondents were invited to discuss the results that they had collectively generated, it became apparent that for many loss aversion had played a significant role in their responses (see Chapter 5.2). Although we may well have posited this as an explanation for the surprising lack of inequality aversion amongst respondents, to have the conjecture supported by evidence obviously enhances its credibility. In the study using the person trade-off method, respondents were encouraged to 'think aloud' (see Chapter 5.3). Although the results were by no means conclusive, the qualitative data did give some general indications about the degree to which subjects emphasised benefit of treatment relative to initial severity of illness. Thus, although neither study involved the degree of intensive interviewing described above, useful additional data was generated.

The emphasis in this discussion that has been placed on "getting behind the numbers" does not conform to the standard welfare economics framework in which preferences are relegated to the status of a (given) 'black-box'. But then economics is not the only discipline that contributes to the study of preference elicitation, be it in health or elsewhere. For a great many years, psychologists, sociologists and philosophers have also contributed greatly to the area. In the context of preference elicitation, this author is inclined to agree with Etzioni [1986] who argues that the complaint that preferences are not the economist's concern is "but one reason for a paradigm change, to combine economics with psychology and sociology, to develop a socioeconomics". And, as Fischhoff [1991] points out, understanding the source of one's own and other disciplinary prejudices is essential both for paradigms to evolve and for this multi-disciplinary collaboration to work.

The measurement of social welfare

The Main Study reported on in Chapter 3 was principally designed to allow a social tariff (or tariffs) of values for all 245 EuroQol health states to be generated, which can then be used to inform policy decisions. The methodology adopted to do this was in line with standard economic practice: that is, to represent the views of a given group, we first elicit the preferences of the individual members of that group and then we aggregate them. The first issue has been addressed by eliciting TTO-based valuations from a representative sample of the UK population. Questions can be (and of course have been) raised about the extent to which the TTO (or, indeed, any of the other methods) yields valid representations of individual preferences, but in principle the method provides answers to the questions that standard economic theory requires to be asked.

On the issue of aggregation, many economists would argue that it is important to take account of the strength of preference of each individual and thus advocate aggregating individual valuations by taking the mean. However, in the realm of public policy, others have suggested that we should take the preferences of the 'median voter' as our 'representative' value. Therefore, from the same set of TTO valuations, it has been necessary to generate both a means-based tariff (see Chapter 3.2) and a medians-based tariff (see Chapter 3.4). The important point is that, whichever measure of central tendency is adopted, the starting point is the same; namely, individual valuations.

Since the implications for health care priorities may depend on how individual preferences are aggregated, we may not be able to make an unambiguous choice between different policy options (see Chapter 3.4 for an example). However, given that the aggregated health state utilities are, for the purposes of public policy, interpreted as measures of social value, then the appropriateness of the different measures of central tendency can be tested by asking respondents whether they agree with the implied priorities for health care that

result from their use. This is referred to by Rawls [1971] as a test of reflective equilibrium.

That some people may disagree with the policy implications that result when individual utilities are aggregated is not surprising. After all, no-one is suggesting that the mean or median view is an accurate representation of the preferences of each and every individual member of the group from whom individual valuations were elicited. Thus, there are likely to be people within the group who also disagree with the resultant policy implications. But each individual counts for only one. Therefore, since an individual may disagree with the implications of mean or median values simply because their own individual valuations differ from the mean or median ones, the issue is whether the implications of the use of mean or median values are agreed with by the 'mean' or 'median' person.

Nord [1995] suggests that this issue should be addressed by presenting respondents with a series of PTO questions (as described in Chapter 5.3) and then looking at whether either of the TTO-based tariffs provides a good approximation of the values implied from the PTO responses. However, this is predicated on the assumption that, in determining the social value attached to different states of health, the PTO represents the "gold standard" by which other methods are to be judged. But the interpretation of PTO responses is problematic. This is because such responses will contain the relative weights a respondent attaches to a number of different attributes. These include the severity of the pre-intervention health state, the severity of the post-intervention health state, the health gain as a result of intervening, and the number of persons treated. Some respondents might also think about the resource implications of people being in particular health states (as indeed some did in the study reported in Chapter 5.2), even if they are explicitly told not to.

Therefore, it is virtually impossible to disentangle the relative weights attached to each of these considerations and hence it is difficult to generalise from results generated by the PTO method. Whilst all the attributes are likely to be important, different weights attached

to each may have quite different implications for the nature of the trade-off (if any) between efficiency and equity and thus for resource allocation decisions. In addition, the fact that respondents are asked to weigh up a number of quite diverse things when thinking about their answers to PTO questions, increases the likelihood of cognitive overload and thus may reduce the validity of responses. There is no doubt that future research is needed to address these issues and the study reported in Chapter 5.3 shows one way in which such research might proceed. Certainly much more work is needed before we are in a position (if ever one could be reached) to argue, as Nord does, that preferences over the treatment of other people should be used as the "gold standard" by which to judge the aggregation of individual preferences.

Moreover, those who advocate the PTO do so on the grounds that it includes factors relevant to social decision-making (for example, attitudes towards inequality) which measures of individual utility, like the TTO, ignore. But it has already been shown that taking account of equity and distributive considerations is not inconsistent with the measurement of individual utility (see Chapter 5.1). This, together with the uncertainty about what valuations elicited by the PTO are based upon (unlike those by the TTO which are based on the elicitation of individual preferences), suggests that future research effort should also be directed towards estimating the shape of the health-related social welfare function.

This approach, unlike the PTO methodology, allows us to address two key issues: 1) how individuals value one health state compared with another, and 2) how society values those same health states. It is important that these two questions are kept separate since each may need to be answered in a different way in different contexts. For example, in conducting a clinical trial we may want to concentrate only on the first issue whilst when comparing the results of different clinical trials, we will have to compare "health gains" from one activity with those from another. It is at this second stage level that we may wish to import notions of distributive justice, but we may not wish to use the same rules in all

situations. Keeping efficiency and equity arguments separate should enable us to look explicitly the types of equity that are considered important in different contexts and at the extent (if any) of the trade-offs between efficiency and equity that result.

The framework developed in Chapter 5.1 should also enable us to consider the extent to which the value attached to a health benefit depends on who is to receive it. The general assumption is that it is the nature of the change in the recipient's HRQoL that is the focus of interest, not who the recipient is. As well as being convenient for research purposes, this assumption has a strong ethical justification. However, as has been shown in previous research as well as in the results from Chapter 5.2, many people think that priority should be given to the young over the old and to people who have cared for their own health over those who have not (see Williams [1988]). Whether policy-makers would ultimately wish to take such considerations into account is not discussed here, but the framework is in place should they wish to do so.

An important issue relates to the way in which we measure attitudes towards equity. In the studies reported in Chapter 5, respondents were asked to make choices about the treatment of other people. They were effectively asked to imagine themselves in the role of a social decision maker who was personally unaffected by the choices they made. This is also the way in which PTO questions have typically been framed. The processes by which respondents arrive (or should arrive) at an answer might be unclear; for example, do (or should) they base their responses on what they think other members of society would prefer or do (or should) they really play the role of a social decision-maker, in which case their answers might be 'contaminated' by the 'political baggage' that real-world policy making comes with (e.g. considerations about the likelihood of public or media outcry about one decision compared with another)?. But the outcomes are clear; the source of value for making these judgements about equity is detached from an individuals vested interests - in the words of Culyer [1980] it is extrinsic to preferences.

There is, however, an alternative approach which requires respondents to think about the impact that their resource allocation choices will have on their own well-being. For example, respondents could be asked to base their answers on what they themselves would prefer were they to face the (known or unknown) probability of being the individuals they are asked to choose between (or in the groups they are asked to choose between in the case of the PTO). In other words, respondents would be required to make their choices as if behind a 'veil of ignorance' which insures that these choices are impartial in certain ways (see Rawls [1971]). This approach means that factors intrinsic to the individual, such as their attitude towards risk. will be important parameters that influence responses. The approach has recently be used by Johannesson and Gerdtham [1996] who asked respondents, from behind a veil of ignorance, to choose between two societies that differ with respect to the number of QALYs received by two groups. The results suggested that, on average, respondents are willing to give up 1 QALY in the group with more QALYs to gain 0.45 QALYs in the group with fewer QALYs.

It is not immediately obvious which of these two broad approaches is the most appropriate for addressing issues related to distributive justice and one will depend on the philosophical perspective taken. In Rawls' world nobody is sick but it is possible to simply include health amongst the primary social goods (although Daniels [1985] agues that a 'thinner veil' would be needed when selecting principles to govern health care resource allocation decisions since we must know something about the society; for example, its resource limitations). However, Dworkin [1977] questions whether Rawls' contract would be binding; "the fact that a particular choice is in my interest at a particular time, under conditions of great uncertainty, is not a good argument for the fairness of enforcing that choice against me later under conditions of much greater knowledge" (p153) ... "His [Rawls'] contract is hypothetical and hypothetical contracts do not supply an independent argument for the fairness of enforcing their terms. A hypothetical contract is not simply a pale form of an actual contract, it is no contract at all" (p151).

These are philosophical questions that cannot be dealt with in the context of this thesis. However, it is an empirical question whether the 'social decision-making' and 'veil of ignorance' approaches yield substantively different results. Therefore, future research effort should be directed towards asking the same question (in any one of the forms outlined in Chapter 5) but from each of these two perspectives. If the results do not differ greatly from one another, then, so far as the implications for resource allocation decisions are concerned, the philosophical issues will be of little practical significance.

Implications for the use of the tariffs

The discussion in this chapter has raised a number of important methodological questions and provided important challenges for future research. It also raises the question of whether the empirical results presented in this thesis, and particularly the tariffs generated in Chapter 3, are currently of any practical use, or whether the methodological issues need to be addressed before the tariffs can be used as an aid to health care decision-making in any meaningful way.

In attempting to answer a question of this kind, it is important to be clear about the context in which the discussion should take place. Whilst it is possible to judge the appropriateness (or otherwise) of the different EuroQol tariffs against the yardstick of the properties that the 'ideal' or 'perfect' tariff might contain, this author feels that this is not a particularly useful framework within which to operate. If this were the background against which all new technologies were judged, then almost all would fall a long way short of the ideal, and hence current practice would prevail. But, of course, much of what is currently practised also falls a long way short of being ideal. Thus, a much more useful starting point for a discussion of a new technology is to compare the situation without the technology with that which results from its introduction.

If the benefits associated with different health care programmes or policies a) do not contain information on the effects associated with changes in HRQoL and b) cannot be expressed in monetary terms (and Chapter 1.1 suggested reasons why the latter might not be possible), then an evaluation of the alternatives must rely on cost-effectiveness analysis (CEA). Since benefits in a CEA are measured in terms of a single outcome expressed in natural units, this type of evaluation is of limited use at the level of allocating resources across different programmes and policies (again see Chapter 1.1).

In principle, then, there is little doubt that the tariffs presented in Chapter 3 represent a better and more useful way of informing resource allocation decisions than the use of, say, life years gained. In practice, of course, there a number of problems associated with using the tariffs to inform such decisions. In addition to the issues raised earlier in this chapter, perhaps the most fundamental issue concerns the generalisability of the valuations contained within the tariffs. All tariffs presented in Chapter 3 relate to valuations elicited for states of health which last for 10 years. Therefore, it seems reasonable to use these valuations in the context of interventions which, in very general terms, are directed at chronic conditions. But given the impact that the time spent in a state can have on its subsequent valuation (an issue which is discussed at some length in Chapters 4.2 and 4.3), it is questionable whether the tariffs can be so readily applied to acute conditions.

Of course, that health state valuations might differ according to a particular effect is not a problem in itself. If the nature and extent of the effect can be estimated then the appropriate adjustments can be made to the valuations in order to account for the effect. For example, in addition to a duration effect, this thesis has also shown that the benefit derived from moving between different health states will differ according to: i) which method is used (see Chapters 2.1 and 2.2), ii) how valuations are aggregated (see Chapters 3.2 and 3.4), and iii) which population sub-group is used (see Chapter 3.3). Although the precise magnitude of these differences will depend on the initial and final

health states, it has been possible to some extent to draw general conclusions about the differential effect that using one set of TTO valuations will have compared to another (see Chapters 3.3 and 3.4 for illustrative examples of the differences that might result from using a tariff based on the values of those aged 60 or over compared to one based on the values of the whole population, and from using a tariff that approximates mean values compared to one based on median values, respectively).

But as much of the discussion in Chapters 4.2 and 4.3 suggests, it is very difficult to estimate the magnitude of the duration effect. Therefore, in the absence of further empirical evidence on the different levels of HRQoL associated with different lengths of time spent in particular health states, this author would be cautious about attaching values derived for states of health lasting 10 years to health states that last for much shorter periods of time, such as a few weeks or months. But without information on the extent to which the use of the tariffs presented in Chapter 3 might misrepresent the HRQoL effects associated with improvements in acute conditions, it is difficult to make a judgement about precisely how cautious to be.

This suggests that in a practical policy sense, a more important question than "what is the magnitude of a particular effect on the tariff values?" is "what difference does using one tariff as opposed to another actually make when applied to real choices?" The short answer to this question is that at the moment we do not really know. This is primarily because such tariffs have only been in the public domain for a very short period of time. In addition, as has been noted elsewhere, there is no real consensus regarding what size difference between valuations is (or should be) considered meaningful.

Perhaps the reason for this lack of consensus is that remarkably few of the previous valuation studies have been subjected to 'real-world' sensitivity analysis. Exceptions are Rosser and Kind [1978], who suggest that many decisions at the community level may not be sensitive to variations in values and Read *et al* [1984] who, contrariwise, show that,

when VAS values are substituted for SG ones in a coronary artery disease decision, the recommended course of action changed for 60% of patients. And perhaps part of the apparent reluctance to test the sensitivity of valuations in a real-world environment results from many of the earlier studies being seen as essentially methodological research, aimed at enhancing our understanding of health status measurement rather than at contributing to resource allocation decisions.

However, whilst the studies which culminated in the generation of the TTO-based tariffs have addressed (and certainly raised) many methodological questions (see Chapters 2 and 3), an important objective of the studies was to generate a tariff (or set of tariffs) that could be used to inform real-world decision-making. Indeed, many of the decisions taken by the MVH Group were taken against this background; for example, considerable time and effort was devoted towards ensuring that the sample in the Main Study was as representative of the UK general population as possible and it was important that the sample was large enough to detect important differences between population sub-groups. Achieving both of these objectives in the Main Study, and the subsequent modelling of valuations for all EuroQol health states (arguably the most significant contribution of this thesis), suggests that potential users of the tariffs might be more willing to consider their application than they were with similar technologies in the past.

Moreover, now that we have good information on the extent to which these tariff values differ according to the measure of central tendency that is chosen and according to certain respondent characteristics, it should be possible in future cost-utility analyses that use EuroQol tariff values to consider the extent to which the cost-utility ratios are affected by the choice of tariff and, crucially, to consider the implications for resource allocation decisions.

When carrying out this sensitivity analysis, it is important to remember that the validity of the tariff values do not rest on there being a precise answer to the question of how many

QALYs a particular programme generates. In many cases, it is likely that the use of different tariff values will make no difference to the ordinal conclusions reached about what programme generates more QALYs than what. As Lockwood [1988] has argued, "only a very radical scepticism, according to which one could not even, with any confidence, set numerical limits in such comparisons, would have the effect of rendering the QALY approach wholly useless." This author agrees with Lockwood's assessment that "such wholesale scepticism would ... be very difficult convincingly to sustain".

Moreover, real-life resource allocation decisions depend upon a great many considerations and the results from a CUA are only likely to play a small part. For example, choices will be based upon historical decisions and considerations about the political 'fallout' of particular choices. Also, decision-makers must currently make their own judgments about the distributional consequences of their choices. Indeed, the identity of the recipient group (for example, the poor or the elderly) may be the motivation for a particular programme in the first place. Against this background, it may be shown that once the implications of using whichever EuroQol tariff have been 'watered down' by these other considerations, for practical policy purposes choosing one tariff compared to another actually makes very little difference.

If such a conclusion is reached, although the philosophical questions about the appropriateness of different measures of central tendency and about whose values should count will remain, much of the heat will be taken out of the debate. On the other hand, if real implications for public policy are a function of the choice of tariff, as in some contexts they might be (even in the face of these other considerations), then the philosophical questions become very real economic ones.

Table 2.1.1 Sample characteristics

	TEST (n=335)		RETEST (n=110)		GENERAL POPULATION
	%	n	%	n	(GHS 1989) %
Female	58.5	196	54.5	60	52.0
Age: 16-20	2.4	8	2.7	3	7.7*
21-60	69.5	233	70.9	78	69.2
61+	27.8	93	26.4	29	23.1
(missing)	(0.3)	(1)	0	-	
Children living with them	33.1	111	32.7	36	47.0
Main Activity					
paid work	43.9	147	42.7	47	59.5
looking after home	25.1	84	30.0	33	-
other	31.1	104	27.3	30	-
Education					
left school at min. age	48.4	162	46.4	51	-
training since school	29.6	99	29.1	32	-
degree or profession	22.1	74	24.5	27	8.0
Cigarette smoker	34.6	116	30.9	34	30.0
Health Status					
Problems with Mobility	21.2	71	18.2	20	
Self Care	3.6	12	1.8	2	
Usual Act	15.8	53	14.5	16	
Pain	34.1	114	27.3	30	
Mood	23.0	77	16.3	18	
Experience of Illness					
Job looking after ill	14.6	49	20.9	23	
Serious illness in self	27.2	91	31.8	35	
in family	36.7	123	48.2	53	
in others	32.8	110	34.5	38	
Experience of any states used in survey	64.5	216	70.9	78	

* from GHS 1990, figures are for ranges 16-19, 20-59 and 60+

Table 2.1.2 Experimental groups
(numbers in brackets refer to incomplete interviews)

Group	1st method	2nd method	Test n	Retest n
1	TTO NP	SG NP	44 (2)	16
2	SG NP	TTO NP	46 (3)	16
3	TTO NP	SG P	49 (1)	17
4	SG P	TTO NP	44 (2)	12
5	TTO P	SG NP	44 (1)	19
6	SG NP	TTO P	29 (0)	4
7	TTO P	SG P	46 (3)	14
8	SG P	TTO P	33 (2)	12
	TOTAL		335 (14)	110
	SG - props		172	55
	SG - no props		163	55
	TTO - props		152	49
	TTO - no props		183	61

Table 2.1.3 Variables used in the regression analysis

Variable	Definition
YOUNG	A dummy variable taking a value of 1 if the respondent is aged 18-29 years, and 0 otherwise.
OLD	A dummy variable taking a value of 1 if the respondent is aged 60 years or over, and 0 otherwise.
SEX	A dummy variable taking a value of 1 if the respondent is female, and 0 otherwise.
SELF	The respondent's rating of their own health state on the VAS.
YOUNGVAS	The product of the variable YOUNG and the VAS score for state x given by this respondent.
OLDVAS	The product of the variable OLD and the VAS score for state x given by this respondent.
SEXVAS	The product of the variable SEX and the VAS score for state x given by this respondent.
SELFVAS	The product of the variable CROWN and the VAS score for state x given by this respondent.

Table 2.1.4 Results of the regression models

Variable	TTO With props	TTO No props	SG With props	SG No props
V_{ix}	-2.86	0.91	-1.51	1.17
	(-3.81)	(3.59)	(-2.29)	(3.51)
$(V_{ix})^2$	7.72	-	4.35	0.28
	(6.02)	-	(4.25)	(1.74)
$(V_{ix})^3$	-4.30	-	-2.50	-
	(-5.58)		(-4.46)	
YOUNG	-0.18	-0.05	0.15	-0.24
	(-1.52)	(-0.64)	(-2.14)	(-2.87)
YOUNGVAS	0.17	0.14	0.27	0.32
	(1.01)	(1.15)	(2.76)	(2.80)
OLD	-0.10	0.39	0.13	0.10
	(-0.85)	(6.10)	(1.74)	(1.08)
OLDVAS	0.01	-0.51	-0.18	-0.17
	(0.04)	(-5.95)	(-1.85)	(-1.48)
SEX	-1.12	0.04	0.07	-0.07
	(-1.27)	(0.68)	(1.02)	(-0.94)
SEXVAS	0.18	0.02	-0.03	0.14
	(1.40)	(0.29)	(-0.31)	(1.56)
SELF	0.15	0.10	-0.06	0.83
	(0.50)	(0.49)	(-0.32)	(3.82)
SELFVAS	-0.19	-0.04	0.05	-1.16
	(-0.47)	(-0.01)	(0.18)	(-4.00)
CONSTANT	0.46	0.01	0.49	-0.23
	(1.56)	(0.03)	(2.20)	(-1.11)
HET				
V_{IX}	0.18	-	-	-
	(0.97)			
$(V_{IX})^2$	-	0.06	0.18	-
		(0.83)	(1.55)	-
$(¶_{IX})^2$	-	-	-	-
Functional form	(0.82)	(-1.38)	(-0.17)	(-1.54)
Het. Disturb	(-2.55)	(-2.11)	(0.06)	(-3.22)
Likelihood Ratio	0.18	0.18	0.28	0.37
Sample	855	982	836	869

Table 2.1.5 Average values of variables used in mapping functions

Variable	TTO With props	TTO No props	SG With props	SG No props
YOUNG	0.29	0.19	0.24	0.20
OLD	0.20	0.24	0.25	0.25
SEX	0.59	0.55	0.56	0.60
SELF	81.05	83.64	84.70	81.06

The values of the variables YVASA, OVASA, SVASA and OWNVASA depend on the VAS score. The values for these variables at which the predicted value functions are evaluated are given by the product of the values given above and the particular value of the index function for VAS.

Table 2.2.1 Completion rates for each method

METHOD	n	STATES UNVALUED			
		TEST		RETEST	
		%	n	%	n
SGP	55	5.3	52	2.4	8
SGNP	54	4.4	41	6.2	20
TTOP	49	0.8	7	0	0
TTONP	61	4.2	44	3.0	11

Table 2.2.2 Consistency rates for each method
Medians (and interquartile ranges)

METHOD	TEST		RE-TEST	
	n	Consistency	n	Consistency
SGP	136	83.3 (66.7-91.7)	45	83.3 (66.7-91.7)
SGNP	145	87.5 (62.5-95.8)	47	83.3 (58.3-100)
TTOP	145	91.7 (75.0-91.7)	48	91.7 (77.1-91.7)
TTONP	163	91.7 (66.7-100)	58	91.7 (66.7-100)

Table 2.2.3 Valuations for each state at test
Medians (and interquartile ranges)

State	SGP	SGNP	TTOP	TTONP
21111	0.85 (0.60-0.95)	0.90 (0.75-0.95)	0.95 (0.75-0.95)	0.95 (0.85-0.95)
11122	0.70 (0.45-0.90)	0.85 (0.50-0.90)	0.90 (0.70-0.95)	0.90 (0.65-0.95)
21221	0.60 (0.25-0.75)	0.75 (0.50-0.90)	0.80 (0.60-0.90)	0.85 (0.65-0.90)
21232	0.30 (0.15-0.55)	0.55 (0.30-0.80)	0.45 (0.05-0.75)	0.55 (0.30-0.70)
22323	0.35 (0.10-0.55)	0.50 (0.30-0.80)	0.40 (0-0.70)	0.55 (0.30-0.70)
33333	0.00 (-0.05-0.10)	0.10 (-0.10-0.40)	-0.30 (-2-0.05)	0.10 (-1.5-0.45)

Table 2.2.4 Valuations for each state at retest
Medians (and interquartile ranges)

State	SGP	SGNP	TTOP	TTONP
21111	0.85 (0.55-0.95)	0.90 (0.70-0.95)	0.95 (0.75-0.95)	0.95 (0.80-0.95)
11122	0.70 (0.50-0.85)	0.80 (0.35-0.90)	0.90 (0.55-0.95)	0.80 (0.60-0.90)
21221	0.70 (0.40-0.85)	0.65 (0.45-0.85)	0.80 (0.60-0.90)	0.80 (0.60-0.90)
21232	0.35 (0.10-0.60)	0.50 (0.30-0.70)	0.40 (0.05-0.75)	0.60 (0.30-0.80)
22323	0.30 (0.05-0.50)	0.50 (0.25-0.80)	0.30 (0-0.65)	0.55 (0.30-0.70)
33333	0.00 (-0.05-0.10)	0.05 (-2-0.40)	-0.60 (-3-0.05)	0.05 (-4-0.40)

Table 2.2.5 Within-respondent comparison of valuations

State	TTONP v SGNP	TTONP v SGP	TTOP v SGNP	TTOP v SGP
21111		T R		T R
11122		T		T R
21221		T R		T R
21232		T	T	
22323		T	T	
33333			T	X

T = TTO valuation is higher than SG one at test ($p < 0.05$)
R = TTO valuation is higher than SG one at re-test ($p < 0.05$)
X = SG valuation is higher than TTO one at test ($p < 0.05$)

Table 2.2.6 Effect of own health state on valuations
(Figures are median scores)

Method	State	Own Health State Dysfunctional[1]							
		Mobility		Usual Activities		Pain/discomfort		Anxiety/Depression	
		No	Yes	No	Yes	No	Yes	No	Yes
SGP	22323 33333	0.35	0.45			-0.05	0.05		
SGNP	21111 11122 21221 21232 22323	0.70 0.50 0.45	0.85 0.70 0.63	0.85 0.75 0.50 0.45	0.90 0.85 0.70 0.60			0.90 0.85 0.45	0.95 0.90 0.55
TTOP	21221 21232	0.75	0.85	0.75	0.85	0.75 0.40	0.85 0.55		
TTONP	33333	0.05	0.3						

[1] Self care omitted due to small numbers with any problems at all.
All differences shown are significant at $p < 0.05$ (Mann-Whitney U tests)

Table 2.2.7 Correlation between test and retest scores

METHOD	Without Important Event n	With Important Event n	SPEARMAN		PEARSON	
			With out	With	With out	With
SGP	25	11	0.63*	0.750	0.63*	0.691
SGNP	31	8	0.71	0.529	0.74	0.498
TTOP	37	11	0.81	0.727	0.83	0.763
TTONP	25	14	0.54*	0.643	0.55*	0.622

* significantly lower than TTO Props ($p < 0.05$)

Table 2.2.8 Correlation for different time intervals
(For respondents without an important event)

Method	n		SPEARMAN		PEARSON	
	<73 days	>73 days	<73 days	>73 days	<73 days	>73 days
SGP	11	14	0.60	0.64	0.57	0.67
SGNP	19	12	0.79	0.56	0.83	0.59
TTOP	18	19	0.81	0.81	0.83	0.83
TTONP	13	12	0.54	0.48	0.64	0.42

Table 3.1.1 Sample characteristics
(Figures are percentages)

Characteristic		Full Sample (n=3395)	After Exclusion (n=3337)	GHS
Sex:	Male	43	43	47
	Female	57	57	53
Age:	18-34	31	32	31
	35-49	25	25	27
	50-59	14	14	15
	60+	31	30	28
Education:	Degree	9	9	8
	Higher	11	11	10
	A/O levels	40	41	45
	None	37	37	35
	Foreign/Other	3	3	3
Social Class:	I, II	29	30	30
	III Non-manual	24	24	22
	III Manual	20	21	21
	IV, V	25	25	21
	Other	1	1	3
Marital status:	single	17	17	21
	married	60	60	64
	widowed	13	12	9
	divorced	10	11	6
Those reporting problems on:				
Mobility		18.4	18.1	-
Self-care		4.2	4.2	-
Usual activities		16.3	16.2	-
Pain/discomfort		32.9	32.8	-
Anxiety/depression		20.9	20.8	-

Table 3.1.2 TTO health state valuations

State	N	Mean (SD)		Median (IQR)	
21111	1306	0.87	(0.24)	0.95	(0.83 - 1.00)
11211	1335	0.87	(0.23)	0.95	(0.83 - 1.00)
11121	1310	0.85	(0.25)	0.93	(0.80 - 1.00)
12111	1310	0.83	(0.30)	0.93	(0.80 - 1.00)
11112	1309	0.82	(0.29)	0.93	(0.75 - 1.00)
12211	828	0.76	(0.33)	0.90	(0.63 - 1.00)
12121	828	0.74	(0.32)	0.85	(0.60 - 1.00)
11122	816	0.72	(0.37)	0.83	(0.63 - 1.00)
22121	830	0.64	(0.42)	0.78	(0.50 - 0.93)
22112	840	0.66	(0.38)	0.74	(0.50 - 0.95)
11312	824	0.55	(0.47)	0.68	(0.40 - 0.93)
21222	823	0.55	(0.46)	0.65	(0.40 - 0.91)
12222	830	0.54	(0.47)	0.65	(0.38 - 0.93)
21312	811	0.51	(0.49)	0.65	(0.33 - 0.93)
22122	809	0.53	(0.47)	0.63	(0.39 - 0.93)
22222	834	0.50	(0.49)	0.63	(0.35 - 0.88)
11113	823	0.39	(0.56)	0.50	(0.00 - 0.88)
13212	820	0.38	(0.54)	0.50	(0.04 - 0.78)
13311	810	0.33	(0.56)	0.50	(0.00 - 0.75)
11131	812	0.20	(0.60)	0.38	(-0.33 - 0.72)
12223	828	0.21	(0.56)	0.35	(-0.28 - 0.63)
21323	819	0.15	(0.59)	0.30	(-0.38 - 0.60)
23321	821	0.14	(0.61)	0.30	(-0.41 - 0.63)
32211	833	0.14	(0.60)	0.25	(-0.38 - 0.63)
21232	826	0.06	(0.61)	0.13	(-0.48 - 0.55)
22323	812	0.04	(0.59)	0.03	(-0.48 - 0.53)
33212	829	-0.02	(0.60)	0.00	(-0.50 - 0.48)
23313	830	-0.07	(0.58)	0.00	(-0.55 - 0.40)
22331	814	-0.01	(0.60)	0.00	(-0.53 - 0.50)
11133	829	-0.05	(0.61)	0.00	(-0.58 - 0.48)
21133	826	-0.07	(0.59)	-0.03	(-0.60 - 0.45)
23232	827	-0.10	(0.59)	-0.08	(-0.63 - 0.43)
33321	828	-0.14	(0.57)	-0.23	(-0.63 - 0.38)
32313	832	-0.16	(0.57)	-0.23	(-0.63 - 0.30)
22233	829	-0.15	(0.57)	-0.28	(-0.63 - 0.34)
32223	825	-0.19	(0.56)	-0.28	(-0.68 - 0.23)
13332	812	-0.23	(0.55)	-0.38	(-0.70 - 0.18)
32232	818	-0.23	(0.57)	-0.38	(-0.73 - 0.20)
32331	826	-0.27	(0.55)	-0.38	(-0.78 - 0.03)
Uncon	3294	-0.41	(0.39)	-0.38	(-0.83 - -0.03)
33232	824	-0.33	(0.51)	-0.43	(-0.75 - 0.00)
33323	833	-0.39	(0.49)	-0.48	(-0.83 - -0.03)
33333	3289	-0.54	(0.41)	-0.65	(-0.93 - -0.28)

Table 3.1.3 Variables used in the regression analysis

Variable	Definition
SEX	A dummy taking the value of 1 if respondent is female, and 0 otherwise
AGE	A continuous variable for the respondent's age
AGE2	Age-squared
EDU1	A dummy taking the value of 1 if the respondent has intermediate qualifications, and 0 otherwise
EDU2	A dummy value of 1 if respondent has no qualifications, and 0 otherwise
SOC1	A dummy value of 1 if respondent is in social class III, and 0 otherwise
SOC2	A dummy value of 1 if respondent is in social class IV or V, and 0 otherwise
MAR1	A dummy taking the value of 1 if the respondent is separated, divorced or widowed, and 0 otherwise
MAR2	A dummy taking the value of 1 if the respondent is single, and 0 otherwise
MOB	1 if respondent reports problems on mobiltity, 0 otherwise
SELF	1 if respondent reports problems on self-care, 0 otherwise
UACT	1 if respondent reports problems on usual act, 0 otherwise
PAIN	1 if the respondent reports pain/discomfort, 0 otherwise
MOOD	1 if the respondent reports anx/depression, 0 otherwise

Table 3.1.4 Results of the regression analysis

Variable	Mild States	Moderate States	Severe States
Adjusted R^2	0.01	0.02	0.03
Constant	.511 (14.495)	.118 (2.129)	-.262 (-6.808)
Sex	-.029 (-3.655)	-0.063 (-5.008)	-.070 (-7.987)
Age	.008 (5.839)	.010 (4.815)	.008 (5.400)
Age 2	-.00010 (-7.114)	-0.00013 (-6.458)	-0.00012 (-8.100)
Edu1	.003 (1.804)	-0.001 (-0.377)	-0.004 (-2.432)
Edu2	.0004 (0.370)	-0.001 (-0.400)	-0.001 (-1.777)
Soc1	-.003 (2.146)	0.002 (1.037)	-0.0003 (-0.242)
Soc2	.0003 (0.305)	-0.001 (-0.544)	-0.003 (-2.992)
Mar1	-0.006 (-4.347)	-0.007 (-3.855)	-0.003 (-2.320)
Mar2	-0.005 (-4.324)	-0.004 (-2.627)	-0.003 (-2.114)
Mob	0.037 (3.095)	0.064 (3.401)	0.039 (3.099)
Self	0.001 (0.132)	0.008 (0.469)	0.010 (0.880)
Uact	0.007 (0.675)	-0.004 (-0.251)	0.013 (1.079)
Pain	-0.0003 (-0.037)	-0.044 (-3.166)	-0.049 (-5.068)
Mood	-0.011 (-1.142)	0.011 (0.739)	0.007 (0.719)

Table 3.2.1 Parameter estimates for the whole sample
(t-statistics are in parentheses)

Variable	Whole Sample	Internal Sample
a	.081 (10.35)	.075 (8.64)
MO	.069 (13.44)	.071 (10.21)
SC	.104 (19.23)	.105 (17.45)
UA	.036 (5.85)	.036 (4.64)
PD	.123 (23.92)	.121 (18.26)
AD	.071 (13.42)	.071 (11.76)
M2	.176 (19.40)	.177 (16.03)
S2	.006 (0.68)	.008 (0.66)
U2	.022 (2.33)	.023 (1.76)
P2	.140 (14.55)	.141 (12.97)
A2	.094 (9.78)	.091 (7.18)
N3	.269 (38.12)	.272 (31.19)
R_2	.046	.046
LM Test	p<0.0001	p<0.0001

As an example of how the tariff is generated, consider the state 11223 estimated for the whole sample:

Full health	= 1.000
Constant term (for any dysfunctional state)	- 0.081
Mobility: level 1	- 0
Self-care: level 1	- 0
Usual activities: level 2 (1xUA)	- 0.036
Pain or discomfort: level 2 (1xPD)	- 0.123
Anxiety or depression: level 3 (2xAD + 1xA2)	- 0.236
N3 (level 3 occurs within at least one dimension)	- 0.269
Therefore, the estimated value for 11223	= 0.255

Table 3.2.2 Comparison of estimated and actual values

State	Actual mean	Estimated	Mean - Estimated
2 1 1 1 1	0.878	0.850	0.028
1 1 2 1 1	0.869	0.883	-0.014
1 2 1 1 1	0.834	0.815	0.019
1 1 1 2 1	0.850	0.796	0.054
1 1 1 1 2	0.829	0.848	-0.019
1 2 2 1 1	0.767	0.779	-0.012
1 2 1 2 1	0.742	0.692	0.050
1 1 1 2 2	0.722	0.725	-0.003
2 2 1 2 1	0.645	0.623	0.022
2 2 1 1 2	0.662	0.675	-0.013
1 1 3 1 2	0.552	0.485	0.067
2 2 1 2 2	0.540	0.552	-0.012
2 1 3 1 2	0.536	0.416	0.120
2 1 2 2 2	0.553	0.620	-0.067
1 2 2 2 2	0.551	0.585	-0.034
2 2 2 2 2	0.500	0.516	-0.016
1 3 2 1 2	0.389	0.329	0.060
1 3 3 1 1	0.346	0.342	0.004
1 1 1 1 3	0.392	0.414	-0.022
1 1 1 3 1	0.200	0.264	-0.064
1 2 2 2 3	0.216	0.151	0.065
2 1 3 2 3	0.160	0.128	0.032
2 3 3 2 1	0.147	0.150	-0.003
3 2 2 1 1	0.152	0.196	-0.044
2 1 2 3 2	0.064	0.088	-0.024
2 2 3 2 3	0.042	0.024	0.018
1 1 1 3 3	-0.049	0.028	-0.077
2 2 3 3 1	-0.011	-0.003	-0.008
2 3 3 1 3	-0.070	0.037	-0.107
3 3 2 1 2	-0.022	0.015	-0.037
2 3 2 3 2	-0.084	-0.126	0.042
2 1 1 3 3	-0.063	-0.041	-0.022
3 3 3 2 1	-0.120	-0.095	-0.025
3 2 3 1 3	-0.152	-0.098	-0.054
2 2 2 3 3	-0.142	-0.181	0.039
3 2 2 2 3	-0.174	-0.163	-0.011
3 2 2 3 2	-0.223	-0.261	0.038
1 3 3 3 2	-0.228	-0.115	-0.113
3 2 3 3 1	-0.276	-0.248	-0.028
3 3 2 3 2	-0.332	-0.371	0.039
3 3 3 2 3	-0.386	-0.331	-0.055
3 3 3 3 3	-0.543	-0.594	0.051
Mean absolute difference			0.039

Table 3.2.3 Predicting the values of an external sample

State	Mean of external sample	Estimated from internal sample	Mean - Estimated
2 1 1 1 1	0.878	0.854	0.024
1 1 2 1 1	0.860	0.889	-0.029
1 2 1 1 1	0.821	0.820	0.001
1 1 1 2 1	0.850	0.804	0.046
1 1 1 1 2	0.805	0.854	-0.049
1 2 2 1 1	0.739	0.784	-0.045
1 2 1 2 1	0.736	0.699	0.037
1 1 1 2 2	0.717	0.733	-0.016
2 2 1 2 1	0.654	0.628	0.026
2 2 1 1 2	0.650	0.678	-0.028
1 1 3 1 2	0.527	0.487	0.040
2 2 1 2 2	0.501	0.557	-0.056
2 1 3 1 2	0.523	0.416	0.107
2 1 2 2 2	0.545	0.626	-0.081
1 2 2 2 2	0.528	0.592	-0.064
2 2 2 2 2	0.523	0.521	-0.002
1 3 2 1 2	0.412	0.328	0.084
1 3 3 1 1	0.404	0.340	0.064
1 1 1 1 3	0.383	0.420	-0.037
1 1 1 3 1	0.169	0.270	-0.101
1 2 2 2 3	0.204	0.158	0.046
2 1 3 2 3	0.189	0.133	0.056
2 3 3 2 1	0.133	0.148	-0.015
3 2 2 1 1	0.135	0.193	-0.058
2 1 2 3 2	0.086	0.092	-0.006
2 2 3 2 3	0.073	0.028	0.045
1 1 1 3 3	-0.106	0.037	-0.143
2 2 3 3 1	0.010	-0.001	-0.011
2 3 3 1 3	-0.038	0.036	-0.074
3 3 2 1 2	-0.005	0.009	-0.014
2 3 2 3 2	-0.085	-0.126	0.041
2 1 1 3 3	-0.047	-0.034	-0.013
3 3 3 2 1	-0.099	-0.100	-0.001
3 2 3 1 3	-0.149	-0.099	-0.050
2 2 2 3 3	-0.185	-0.175	-0.010
3 2 2 2 3	-0.164	-0.161	-0.003
3 2 2 3 2	-0.129	-0.261	0.132
1 3 3 3 2	-0.219	-0.114	-0.105
3 2 3 3 1	-0.235	-0.249	0.014
3 3 2 3 2	-0.322	-0.374	0.052
3 3 3 2 3	-0.375	-0.333	-0.042
3 3 3 3 3	-0.520	-0.595	0.075
Mean absolute difference			0.046

Table 3.3.1 Parameter estimates
(t-statistics in parentheses)

Variable	18-59 year-olds	>60 year-olds
MO	.071 (4.93)	.074 (4.62)
SC	.075 (4.77)	.126 (7.75)
UA	.022 (1.54)	.054 (2.79)
PD	.099 (6.68)	.123 (7.75)
AD	.076 (2.94)	.044 (2.64)
M2	.178 (6.98)	.228 (8.09)
S2	.045 (1.51)	-.016 (-0.55)
U2	.062 (2.94)	-.048 (-1.59)
P2	.158 (6.24)	.127 (4.37)
A2	.086 (3.13)	.098 (3.27)
MOgend	-.009 (-0.41)	.003 (0.13)
SCgend	.015 (0.58)	.048 (2.18)
UAgend	.017 (0.61)	-.011 (-0.43)
PDgend	.036 (1.05)	.016 (0.79)
ADgend	-.008 (-0.36)	.045 (2.08)
M2gend	-.015 (-0.35)	-.064 (-1.72)
S2gend	-.022 (-0.74)	-.073 (-1.92)
U2gend	-.038 (-0.96)	.039 (1.00)
P2gend	-.002 (-0.06)	-.064 (-1.64)
A2gend	.032 (0.87)	-.044 (-1.12)
N3	.219 (10.7)	.328 (14.8)
N3gend	.034 (1.31)	.036 (1.25)
a	.081 (4.66)	.077 (3.15)
gend	-.010 (-0.35)	.025 (0.77)

For men, the value for a particular health state is calculated using the variables without the suffix 'gend', and for women is calculated using the variables with the suffix 'gend'.

Table 3.3.2 Differences between age groups
Figures are t-ratios

Variable	>60 values - men	>60 values - women
MO	-0.66	-1.48
SC	-4.04	-5.23
UA	-3.91	-1.16
PD	-1.99	-0.56
AD	2.53	-1.76
M2	-2.67	-0.09
S2	1.86	2.41
U2	5.14	1.35
P2	1.99	2.76
A2	-1.12	1.53
N3	-7.54	-8.07

Figures are t-ratios, calculated as $(B_2 - B_1) /$ standard error of B_2

Positive t-ratios indicate that the 18-59 age group has a higher coefficient than the 60 or over age group i.e. that they attach a greater disutility to that dimension, and vice versa.

Table 3.3.3 Coefficients on each dimension for each age group

Dimension	18-59 Men	18-59 Women	>60 Men	>60 Women
Constant	.081	.071	.077	.102
Mobility				
level 2	.071	.062	.074	.076
level 3	.320	.287	.374	.316
Self-care				
level 2	.075	.090	.126	.174
level 3	.195	.203	.236	.259
Usual activities				
level 2	.022	.039	.054	.043
level 3	.106	.102	.060	.077
Pain/discomfort				
level 2	.099	.135	.123	.139
level 3	.356	.426	.374	.342
Anxiety/depression				
level 2	.076	.068	.044	.088
level 3	.238	.254	.186	.230
Any level 3	.219	.253	.328	.364

Note: Higher numbers correspond to a greater decrement in health state value.

The value for a particular health state is calculated as 1 (the value for full health) minus the constant term minus the relevant coefficients on the dysfunctional dimensions minus the 'Any level 3' variable, if applicable.

Table 3.3.4 Differences between actual and estimated values

Difference in value mean - estimated	Males 18-59	Females 18-59	Males >60	Females >60
> -0.10	1	2	2	3
-0.06 - -0.10	7	5	6	9
-0.01 - -0.05	8	14	11	11
0	3	2	5	3
0.01 - 0.05	17	13	12	6
0.06 - 0.10	5	5	5	5
> 0.10	1	1	1	5

Table 3.4.1 Coefficients on variables using medians

Variable	Coefficient
a	-.038 (-1.03)
MO	.054 (1.73)
SC	.129 (3.79)
UA	.044 (1.14)
PD	.114 (3.60)
AD	.100 (3.02)
M2	.350 (6.22)
S2	-.004 (-0.06)
U2	.071 (1.23)
P2	.294 (4.78)
A2	.123 (1.98)
N3	.163 (3.34)
Adjusted R^2	.98

As an example of how the tariff is generated, consider the state 11223 estimated for the whole sample:

Full health	= 1.000
Constant term (for any dysfunctional state)	- 0.038
Mobility: level 1	- 0
Self-care: level 1	- 0
Usual activities: level 2 (1xUA)	- 0.088
Pain or discomfort: level 2 (1xPD)	- 0.228
Anxiety or depression: level 3 (2xAD + 1xA2)	- 0.323
N3 (level 3 occurs within at least one dimension)	- 0.163
Therefore, the estimated value for 11223	= 0.160

Table 3.4.2 Actual and estimated values using medians

State	Median Value	Estimated Value	Median-Estimated
2 1 1 1 1	0.950	0.984	-0.034
1 1 2 1 1	0.950	0.994	-0.044
1 2 1 1 1	0.925	0.909	0.016
1 1 1 2 1	0.925	0.924	0.001
1 1 1 1 2	0.925	0.938	-0.013
1 2 2 1 1	0.900	0.865	0.035
1 2 1 2 1	0.850	0.795	0.055
1 1 1 2 2	0.825	0.824	0.001
2 1 2 1 2	0.775	0.741	0.034
2 2 1 1 2	0.750	0.756	-0.006
1 1 3 1 2	0.675	0.616	0.059
2 2 1 2 2	0.650	0.642	0.008
2 1 3 1 2	0.650	0.563	0.087
2 1 2 2 2	0.650	0.727	-0.077
1 2 2 2 2	0.650	0.652	-0.002
2 2 2 2 2	0.625	0.598	0.027
1 3 2 1 2	0.500	0.478	0.022
1 3 3 1 1	0.500	0.462	0.038
1 1 1 1 3	0.500	0.552	-0.052
1 1 3 1 1	0.375	0.353	0.022
1 2 2 2 3	0.375	0.266	0.109
2 1 3 2 3	0.325	0.225	0.100
2 3 3 2 1	0.300	0.295	0.005
3 2 2 1 1	0.275	0.245	0.030
2 1 2 3 2	0.138	0.156	-0.018
2 2 3 2 3	0.025	0.096	-0.071
1 1 1 3 3	0.000	0.030	-0.030
2 2 3 3 1	0.000	0.012	-0.012
2 3 3 1 3	0.000	0.085	-0.085
3 3 2 1 2	0.000	0.020	-0.020
2 3 2 3 2	-0.025	-0.097	0.072
2 1 1 3 3	-0.025	-0.024	-0.001
3 3 3 2 1	-0.175	-0.110	-0.065
3 2 3 1 3	-0.225	-0.194	-0.031
2 2 2 3 3	-0.225	-0.196	-0.029
3 2 2 2 3	-0.275	-0.192	-0.083
3 2 2 3 2	-0.375	-0.377	0.002
1 3 3 3 2	-0.375	-0.159	-0.216
3 2 3 3 1	-0.375	-0.393	0.018
3 3 2 3 2	-0.425	-0.501	0.076
3 3 3 2 3	-0.475	-0.433	-0.042
3 3 3 3 3	-0.625	-0.841	0.216

mean absolute difference 0.047

Table 3.4.3 Example of differences between mean and median tariffs

Definition of groups

Group 1: 5 states with only one dimension at level 2
Group 2: 26 states with any combination of levels 1 and 2
Group 3: 80 states with one dimension at level 3
Group 4: 80 states with two dimensions at level 3
Group 5: 51 states with three or more dimensions at level 3

Mean health state values for patients in each group

Group	Individual-based tariff	Medians-based tariff
1	.838	.950
2	.695	.793
3	.240	.355
4	.032	.056
5	-.226	-.313

Table 4.1.1 Adjusted TTO scores for various discount rates

τ	$\rho = 1$	$\rho = 0.95$	$\rho = 0.91$	$\rho = 0.87$
1	0.9	0.88	0.85	0.83
2	0.8	0.76	0.72	0.68
3	0.7	0.65	0.59	0.54
4	0.6	0.54	0.48	0.43
5	0.5	0.44	0.38	0.33
6	0.4	0.34	0.29	0.25
7	0.3	0.25	0.21	0.17
8	0.2	0.16	0.13	0.11
9	0.1	0.08	0.06	0.05

Table 4.2.1 Sample characteristics

		n (=39)	%
Age:	18-39	15	40
	40-59	14	37
	60+	9	24
Gender:	Male	12	31
	Female	27	69
Marital Status:	Married	24	62
	Separated/divorced widowed	6	15
	Single	9	23
Education:	Degree	7	18
	Intermediate	21	54
	None	11	28
Smoker:	Yes	13	33
	No	26	67

Table 4.2.2 Transformed VAS scores

STATE	MEDIAN (IQR)
11111	1.00
11121	0.81 (0.73-0.89)
11122	0.57 (0.44-0.67)
21232	0.38 (0.20-0.50)
22233	0.19 (0.06-0.28)
33333	-0.03 (-0.11-0.10)
Death	0.00

Table 4.2.3 Preferences over the timing of one year of poor health
Figures are numbers of respondents

State	Poor health later preferred to poor health now (ie r >0)	Poor health now equivalent to poor health later (ie r = 0)	Poor health now preferred to poor health later (ie r<0)
11121	5	10	11
11122	8	12	10
21232	7	9	16
22233	11	9	13
33333	8	15	10
Total	39 (25%)	55 (36%)	60 (39%)

Table 4.2.4 Discount rates for health states (%)

State	Mean (S.D.)	Median (IQR)
11121	-2.94 (5.34)	0.00 (-4.75-0.00)
11122	-2.58 (8.09)	0.00 (-5.30-0.00)
21232	-3.50 (10.73)	0.00 (-13.33-0.00)
22233	-0.46 (15.52)	0.00 (-7.60-7.20)
33333	+1.35 (13.37)	0.00 (0.80-10.18)

Table 4.2.5 Discount rates for respondents

Discount Rate	Number of Respondents
+ only	2
- only	2
0 only	8
+ or 0	4
- or 0	8
+ or -	6
+ or - or 0	6
missing	3
TOTAL	39

Table 4.2.6 Implied median TTO scores

State	n	1 Month	1 Year Now	1 Year Later	10 Years
11121	37	0.083	0.250 b	0.250 b	0.83 a
11122	35	0.083	0.250 b	0.125	0.73 a
21232	35	-.083 c	-.100	-.100	0.03 a
22232	34	-.067 c	-.200	-.175	-0.10 a
33333	33	-.067 c	-.363	-.275	-0.33 a

a significantly higher than all other durations
b significantly higher than one month valuations
c significantly higher than one year valuations

Table 4.3.1 Variables used in the regression analysis

Variable	Definition
Month duration	Valuations for states lasting one month
Year duration	Valuations for states lasting one year
Ten year duration	Valuations for states lasting ten years
Health state group 1	'Very mild' health states (11112, 11121, 11211, 12111, 21111) - forms baseline category for other health state group dummy variables
Health state group 2	'Mild' health states. Any state containing only 1s and 2s (except those in group 1)
Health state group 3	'Moderate' health states. Any state contining one 3 plus those containing two 3s if the other three dimensions are all 1s
Health state group 4	'Severe' health states. Any state containing two 3s (except those in group 3)
Health state group 5	'Very severe' health states. Any state containing three or more 3s plus 'unconscioous'
Age	Age in years
Education group 2	Intermediate qualifications
Education group 3	no qualifications
TYD * group 2 TYD * group 3 TYD * group 4 TYD * group 5	Interaction terms between ten year health state durations and the respective health state groups.

Table 4.3.2 Sample characteristicse
(Figures are percentages)

		Sample (n=236)	GHS (1992)
Sex:	Male	45	47
	Female	55	53
Age:	18-34	28	31
	35-49	30	27
	50-59	13	15
	60+	29	28
Education:	Degree	10	8
	Higher	12	10
	A/O levels	42	45
	None	35	35
	Foreign/other	1	3
Social Class:	I and II	31	30
	III non-manual	21	22
	III manual	23	21
	IVand V	23	21
	Other	2	3
Marital Status	Single	15	21
	Married	66	64
	Widowed	7	9
	Divorced	13	6

Table 4.3.3 Mean health state score across durations

Health state	Month	Year	Ten year
11111	100	100	100
32211	46.10	39.87	24.79
23321	37.78	36.12	28.39
11211	85.46	83.81	80.37
11112	85.84	84.88	83.43
12121	73.70	71.90	70.11
21232	40.50	36.16	32.10
13332	23.81	15.31	9.84
12222	62.66	58.62	53.09
12211	73.37	72.15	67.93
33212	37.46	35.03	27.92
21312	52.41	48.18	47.87
33323	18.81	10.14	3.54
33232	20.15	15.33	9.00
11113	57.23	53.94	48.36
22323	36.20	29.69	26.05
32223	29.31	20.74	13.37
32331	23.76	22.42	16.08
12223	46.13	35.14	31.95
11131	51.19	45.07	38.77
32232	26.31	24.37	17.92
11133	42.64	36.52	28.68
33321	26.80	16.47	13.06
22331	31.30	32.28	28.32
13311	54.41	43.77	35.01
21323	41.92	33.28	28.40
12111	85.40	85.03	82.56
13212	57.74	50.90	43.62
22233	28.99	21.29	17.78
32313	30.65	23.57	17.76
22222	59.60	51.60	44.85
22112	70.71	66.93	65.63
11121	85.11	85.39	81.19
22121	65.86	64.49	59.59
22122	63.63	57.38	57.16
21111	87.32	85.17	83.17
21222	62.17	60.79	54.34
23313	31.05	29.28	18.25
33333	5.72	-1.08	-5.75
21133	36.95	27.93	29.87
23232	29.77	25.33	18.17
11312	59.13	56.03	53.46
11122	74.21	72.20	69.26

Table 4.3.4 Number of worse than dead VAS valuations

	Total number of valuations	Month	Year	Ten year
Very mild states (Health state group 1)	344	0	0	4
Mild states (Health state group 2	507	2	3	8
Moderate states (Health state group3)	569	2	13	28
Severe states (Health state group 4)	568	7	22	48
Very severe states (Health state group 5)	830	55	100	204

Table 4.3.5 Parameter estimates for different durations

Variable	One month	One Year	10 Years
a	.107	.113	.144
MO	.055	.052	.050
SC	.064	.073	.078
UA	.041	.045	.067
PD	.079	.096	.096
AD	.056	.063	.047
M2	.045	.047	.059
S2	-.006	-.008	.001
U2	.020	.005	-.044
P2	.036	-.005	-.021
A2	.003	.014	.031
N3	.147	.183	.211
Adjusted R^2	.63	.62	.55

Table 4.3.6 Regression of one month on one year scores

	Males[†] sample = 1200 respondents = 100	Females sample = 1488 respondents = 124
Response: 1 month duration		
Fixed:	0.213 (0.218)	0.301 (0.167)
Constant	-0.096 (0.014)	-0.074 (0.012)
Health state group 2	-0.174 (0.016)	-0.145 (0.014)
Health state group 3	-0.281 (0.019)	-0.226 (0.016)
Health state group 4	-0.364 (0.022)	-0.307 (0.018)
Health state group 5	0.016 (0.014)	0.009 (0.011)
Age	-0.0003 (0.0003)	-0.0002 (0.0002)
Age squared	$1.37e^{-6}$ ($1.94e^{-6}$)	$1.45e^{-6}$ ($1.44e^{-6}$)
Age cubed	-0.044 (0.022)	-0.022 (0.025)
Education group 2	-0.071 (0.024)	-0.022 (0.027)
Education group 3	0.436 (0.024)	0.510 (0.020)
Year duration		
Random:		
σ_e^2 : within individuals	0.019 (0.001)	0.018 (0.001)
σ_u^2 : between individuals	0.006 (0.001)	0.007 (0.001)
RESET test, t-ratio (p value)	0.66 (p > 0.05)	1.60 (p > 0.05)
Prediction[††]:	0.171	0.151

[†] 12 male respondents had missing age data and were omitted from the analysis.

[††] Based on re-estimation of model on 70% of the data, and comparing predicted to actual values of remaining 30% of the data..

Table 4.3.7 **Regression of one year on ten year scores**

	Males[†] sample = 1200 respondents = 100	Females sample = 1488 respondents = 124
Response: 1 year duration		
Fixed:		
Constant	0.267 (0.323)	0.591 (0.185)
Health state group 2	-0.164 (0.046)	-0.318 (0.053)
Health state group 3	-0.3714 (0.043)	-0.337 (0.052)
Health state group 4	-0.450 (0.043)	-0.435 (0.051)
Health state group 5	-0.577 (0.043)	-0.516 (0.051)
Age	0.026 (0.021)	0.0006 (0.012)
Age squared	-0.0006 (0.0004)	$-1.76e^{-7}$ (0.0002)
Age cubed	$3.77e^{-6}$ ($2.84e^{-6}$)	$-1.35e^{-7}$ ($1.55e^{-6}$)
Education group 2	-0.015 (0.032)	0.004 (0.027)
Education group 3	-0.017 (0.035)	0.005 (0.030)
Ten Year duration (TYD)	0.253 (0.051)	0.288 (0.060)
TYD * Health state group 2	0.050 (0.060)	0.277 (0.067)
TYD * Health state group 3	0.198 (0.056)	0.145 (0.066)
TYD * Health state group 4	0.122 (0.057)	0.170 (0.067)
TYD * Health state group 5	0.322 (0.062)	0.277 (0.067)
Random:		
σ_e^2: within individuals	0.020 (0.001)	0.022 (0.001)
σ_u^2: between individuals	0.015 (0.002)	0.008 (0.001)
RESET test; t-ratio (p value)	1.38 ($p > 0.05$)	1.86 ($p > 0.05$)
Prediction[††]:	0.184	0.183

[†] 12 male respondents had missing age data and were omitted from the analysis.

[††] Based on re-estimation of model on 70% of the data, and comparing predicted to actual values of remaining 30% of the data.

Table 5.2.1 Questions in Section A

	Current health state X and Y	Mean gain M	Distribution of gains X:Y	Potential outcomes under current allocation X	Y
QA1	2	60	1:4	26	98
QA2	2	15	1:4	8	26
QA3	2	60	1:2	42	82
QA4	2	15	1:2	12	22
QA5	40	15	1:4	46	64
QA6	76	15	1:4	82	100

Table 5.2.2 Results of Section A

	Equally distributed gain ξ			Atkinson Index I	
	Mean gain M	mean	median	mean	median
QA1	60	60.32	60	-0.005	0
QA2	15	16.24	15	-0.083	0
QA3	60	60.54	60	-0.009	0
QA4	15	16.03	15	-0.050	0
QA5	15	15.76	15	-0.050	0
QA6	15	14.70	15	0.020	0

The Atkinson Index is calculated as 1 - ξ /M.

Table 5.2.3 Attitudes to inequality

	Averse ($I>0$)	Neutral ($I = 0$)	Seeking ($I < 0$)
QA1	9	14	14
QA2	8	11	18
QA3	9	17	11
QA4	6	13	18
QA5	8	13	16
QA6	18	8	11

Where I = the value of the Atkinson Index.

Table 5.2.4 Results from Sections B and C

	Equally distributed gain ξ			Atkinson Index I	
	Mean gain M	mean	median	mean	median
QB1	15	18.46	19	-.231	-.267
QB2	15	18.49	19	-.232	-.267
QB3	15	16.19	16	-079	-.067
QC1	15	18.87	20	-.258	-.333
QC2	15	17.67	18	-.175	-.200

The Atkinson Index is calculated as 1 - ξ /M.

Table 5.2.5 Section A results after weighing losses and gains

	Equally distributed gain ξ			Atkinson Index I	
	Mean gain M	mean	median	mean	median
QA1	60	48.71	48.00	0.188	0.200
QA2	15	13.48	12.00	0.101	0.200
QA3	60	54.25	53.33	0.096	0.111
QA4	15	14.29	13.33	0.047	0.111
QA5	15	12.92	12.00	0.139	0.200
QA6	15	12.06	12.00	0.196	0.200

The figures in this Table have been calculated on the assumption that potential losses are weighted twice as heavily as potential gains.

Table 5.2.6 Inequality attitudes after weighing losses and gains

	Averse (I>0)	Neutral (I = 0)	Seeking (I < 0)
QA1	35	1	1
QA2	26	8	3
QA3	31	4	2
QA4	27	3	7
QA5	30	1	6
QA6	29	4	4

Where I = the value of the Atkinson Index calculated on the assumption that potential losses are weighted twice as heavily as potential gains.

Table 5.3.1 Sample characteristics (n=28)

Sex:	Male	15	(53.5%)
	Female	13	(46.5%)
Age:	18-24 yrs	3	(10.7%)
	25-34 yrs	17	(60.7%)
	35-49 yrs	5	(17.9%)
	50-59 yrs	3	(10.7%)
Education:	Degree	17	(60.7%)
	Intermediate	10	(35.7%)
	None	1	(3.6%)
Marital Status:	Married	10	(35.7%)
	Divorced	1	(3.6%)
	Single	17	(60.7%)
Interview time:	mean: 42 min (range: 30-55 mins)		

Table 5.3.2 **States used and their mean ranks**

11111	1.04
11121	2.39
11112	3.00
11122	4.39
22112	5.54
21222	6.46
22122	6.54
11131	7.96
22222	8.39
12223	9.14
21323	10.64
33212	11.64
22323	11.86
Death	14.04
33333	14.21

Table 5.3.3 States identified as the outcome from treatment 2

Resp. No.	Interval state	BIS	WIS	Used in PTO
2		11121	22112	22112
3		11122	11131	11131
4		11121	11122	11122
5	11122			11122
6	11121			11121
7	11122			11122
8	22112			22112
9	11131			11131
10		21222	11122	11122
11	11122			11122
12	21333			21222
13	11121			non-response
14		11121	11122	11122
15	11122			11122
16		21222	12223	21222
17	21222			21222
18	21222			21222
19				22122
20				11121
21	22112			22112
22	21222			21222
23	22112			22112
24	11112			11112
25			11111	11121
26		21222		21222
27		21222		21222
28		21222		21222
29		21222		21222

Where BIS is the state that makes treatment 2 just preferred to treatment 1 and WIS is the state that makes treatment 1 just preferred to treatment 2.

Table 5.3.4 PTO responses (Geometric mean = 0.694)

Respondent	Treatment 1	Treatment 2	T2/T1
7	10	10	1.00
17	10	10	1.00
18	10	10	1.00
6	3.5	10	2.86
14	4	10	2.50
22	4	10	2.50
24	4	10	2.50
28	5	10	2.00
3	7	10	1.43
29	9.5	10	1.05
9	10	0.5	0.05
12	10	0.5	0.05
2	10	3	0.30
19	10	3.5	0.35
5	10	4	0.40
8	10	4.5	0.45
20	10	4.5	0.45
26	10	4.5	0.45
10	10	5	0.50
25	10	5	0.50
4	10	6	0.60
21	10	7	0.70
23	10	7.5	0.75
27	10	7.5	0.75
15	10	8	0.80
11	10	9	0.90
16	10	9.5	0.95
13	non-response	non-response	non-response

Table 5.3.5 Interpretation of PTO qualitative data

Responses to "Could you tell me why you made this choice?"

No	Pr	A	B	C	D	E
7	Ind					
17	Ind					
18	Ind					
6	1			*		*
14	1			*		*
22	1		*			*
24	1			*		*
28	1		*			
3	1			*		
29	1			*		
9	2	*				
12	2	*				
2	2	*			*	
19	2				*	*
5	2					*
8	2	*				*
20	2	*				*
26	2	*				*
10	2	*				*
25	2					
4	2				*	*
21	2	*				*
23	2	*				*
27	2	*				*
15	2	*				*
11	2		*			
16	2	*				*
13	NR					

Key:

Pref	Preference for either T1 or T2, or indifference (ind)
A	Respondent commented that T1 was not a very good treatment
B	Respondent stated preferred treatment gave greater improvement
C	Respondent commented on wanting to help the worse off
D	Respondent mentioned either economic or societal costs
E	Respondent made some reference to the number of people treated

Figure 1.2.1 The EuroQol descriptive 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

Figure 2.1.1: A hypothetical value function

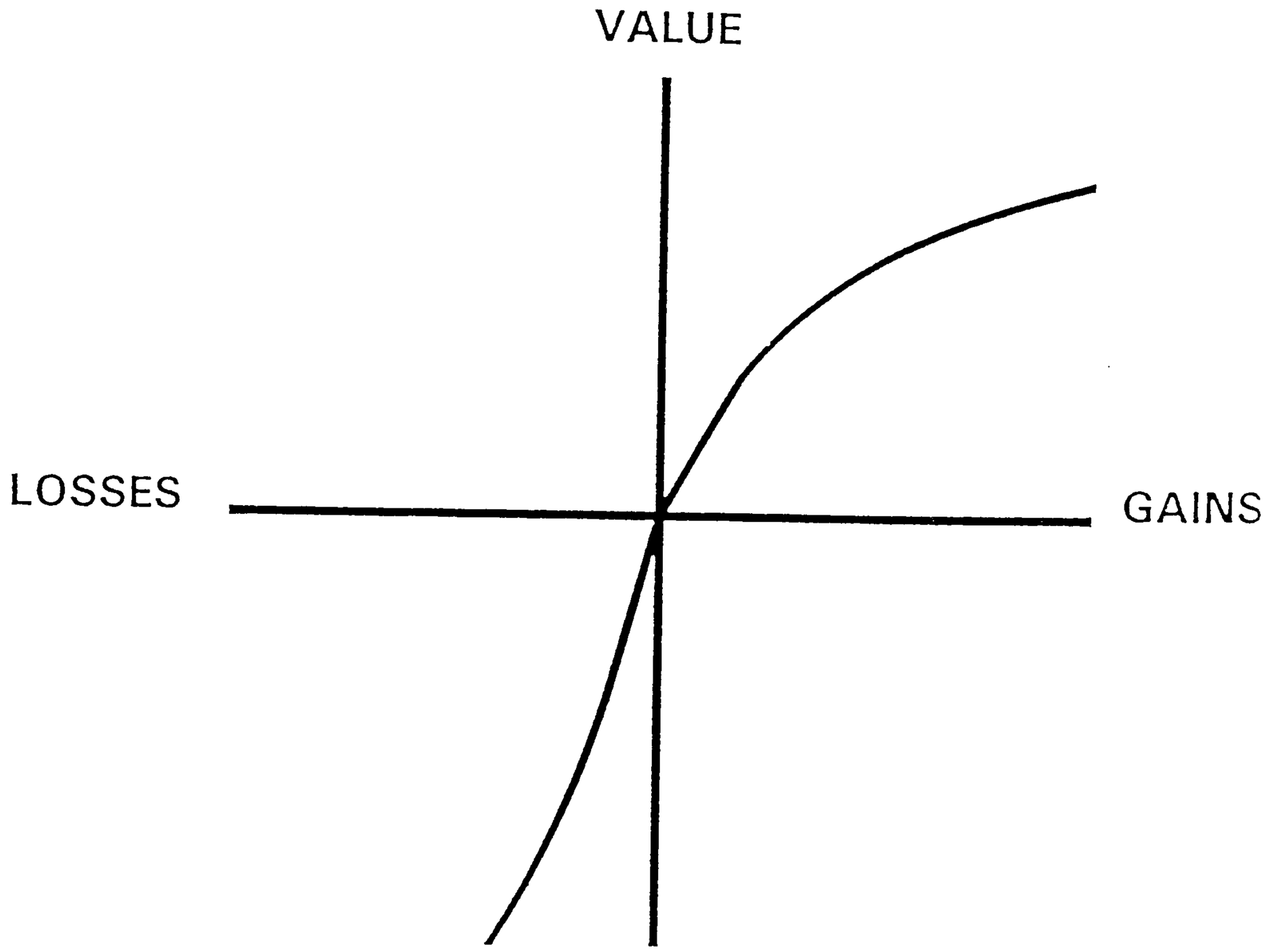


Figure 2.1.2: Mapping Function for TTO Props

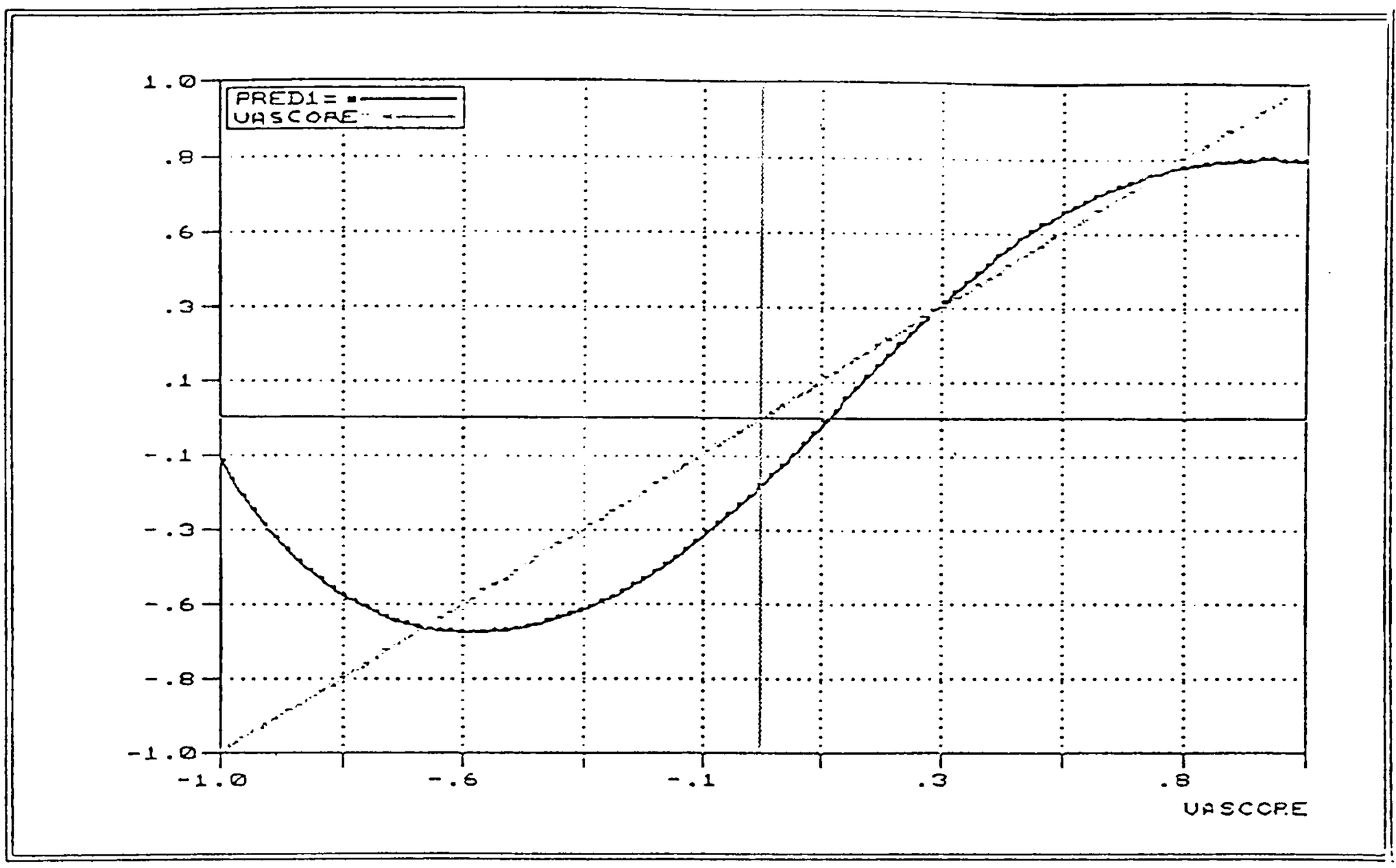


Figure 2.1.3: Mapping Function for TTO No Props

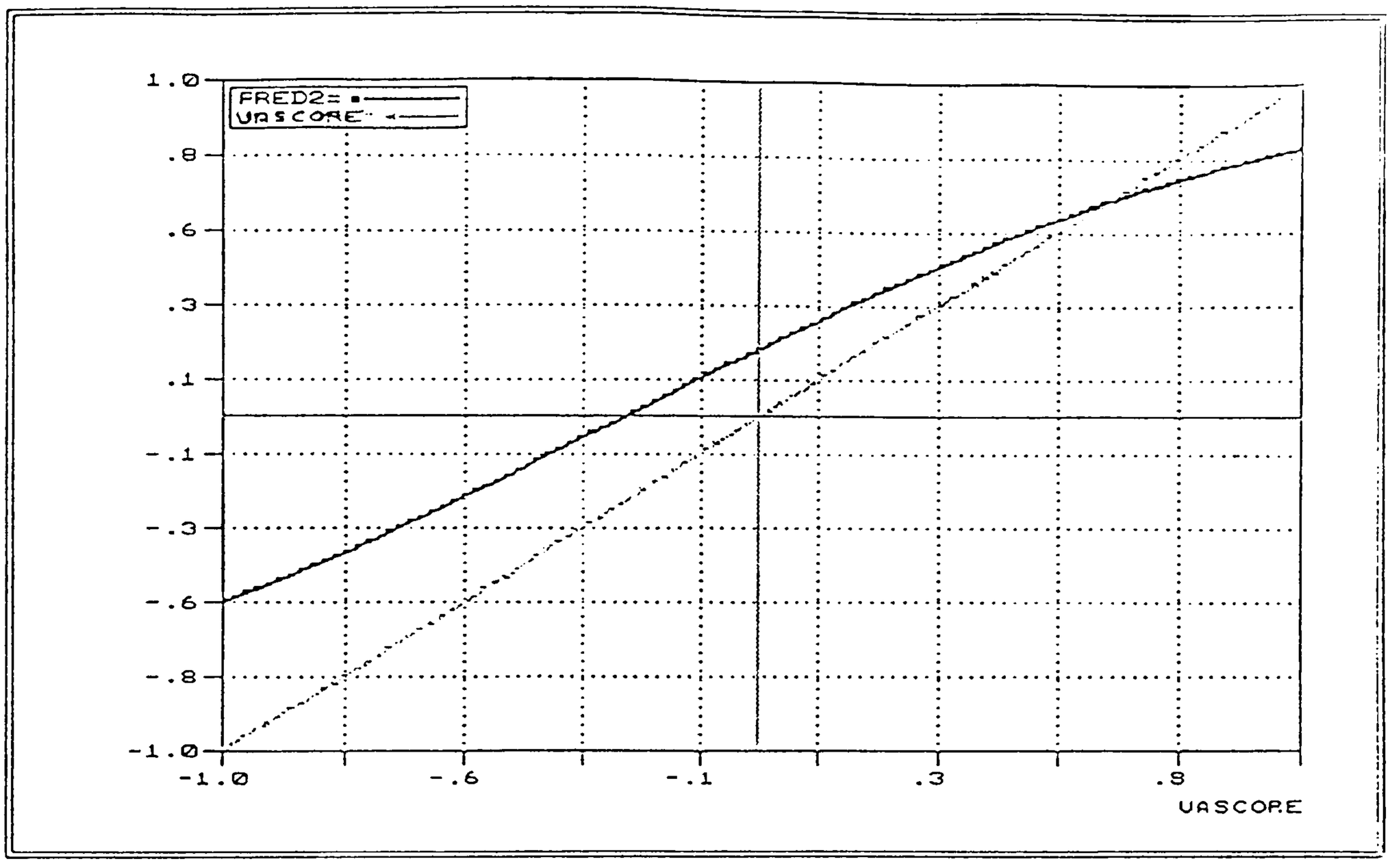


Figure 2.1.4: Mapping Function for SG Props

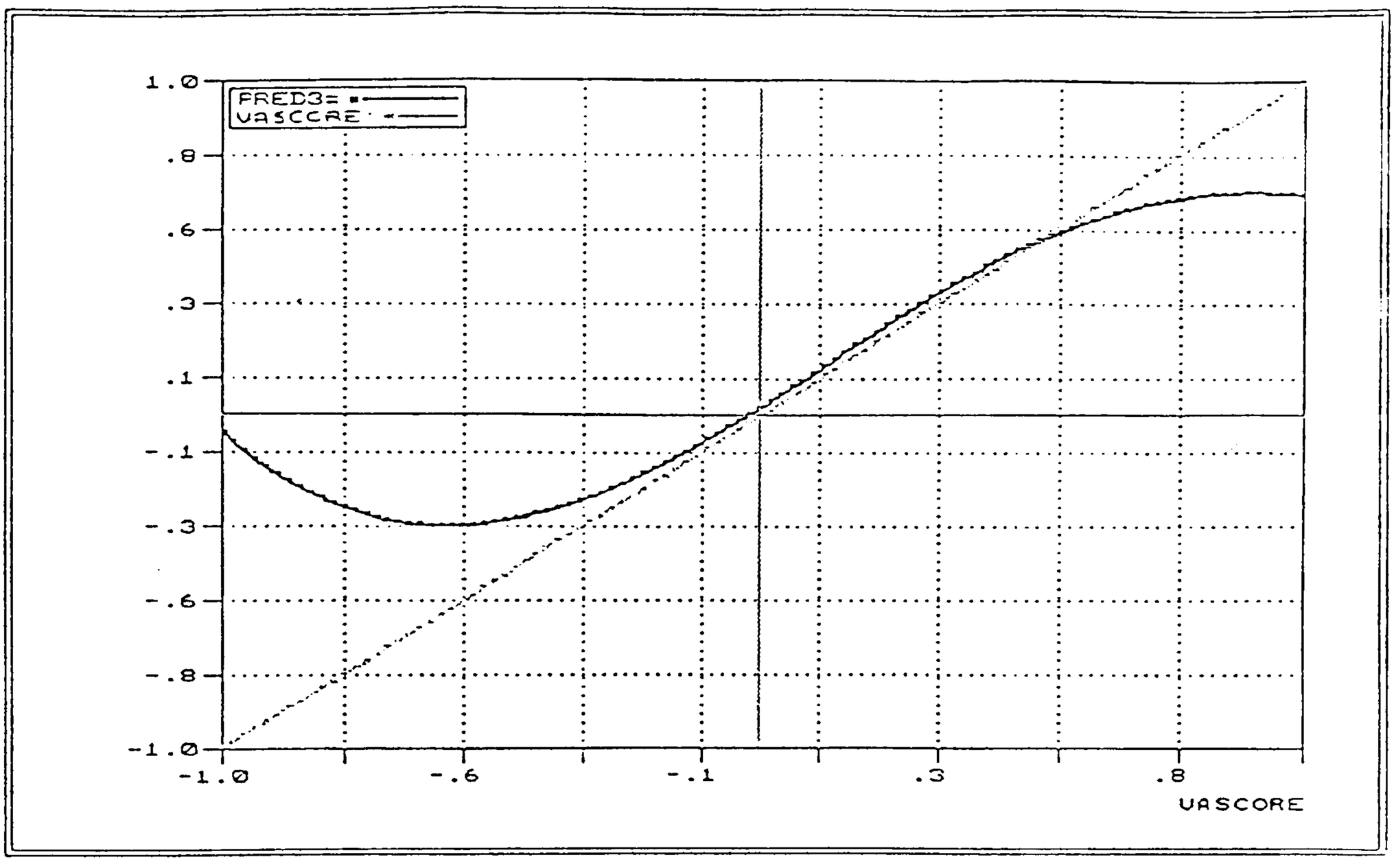


Figure 2.1.5: Mapping Function for SG No Props

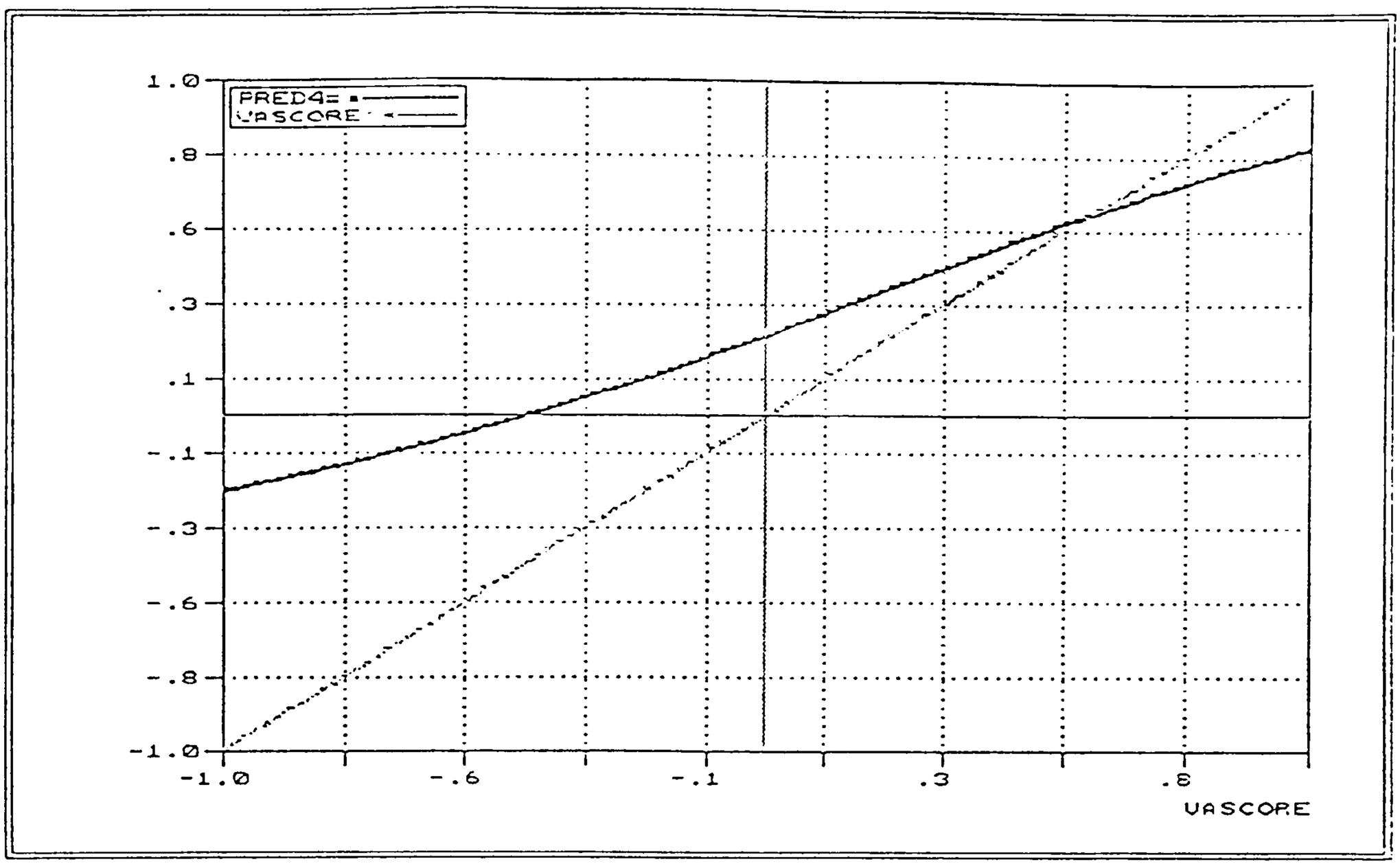


Figure 3.1.1 States valued in the MVH main study

Each respondent valued 11111, Immediate Death, 33333 and unconscious

plus

2 from 5 "very mild" states:

11112 11121 11211 12111 21111

plus

3 from 12 "mild" states:

11122 11131 11113 21133 21222 21312 12211 11133 22121
12121 22112 11312

plus

3 from 12 "moderate" states:

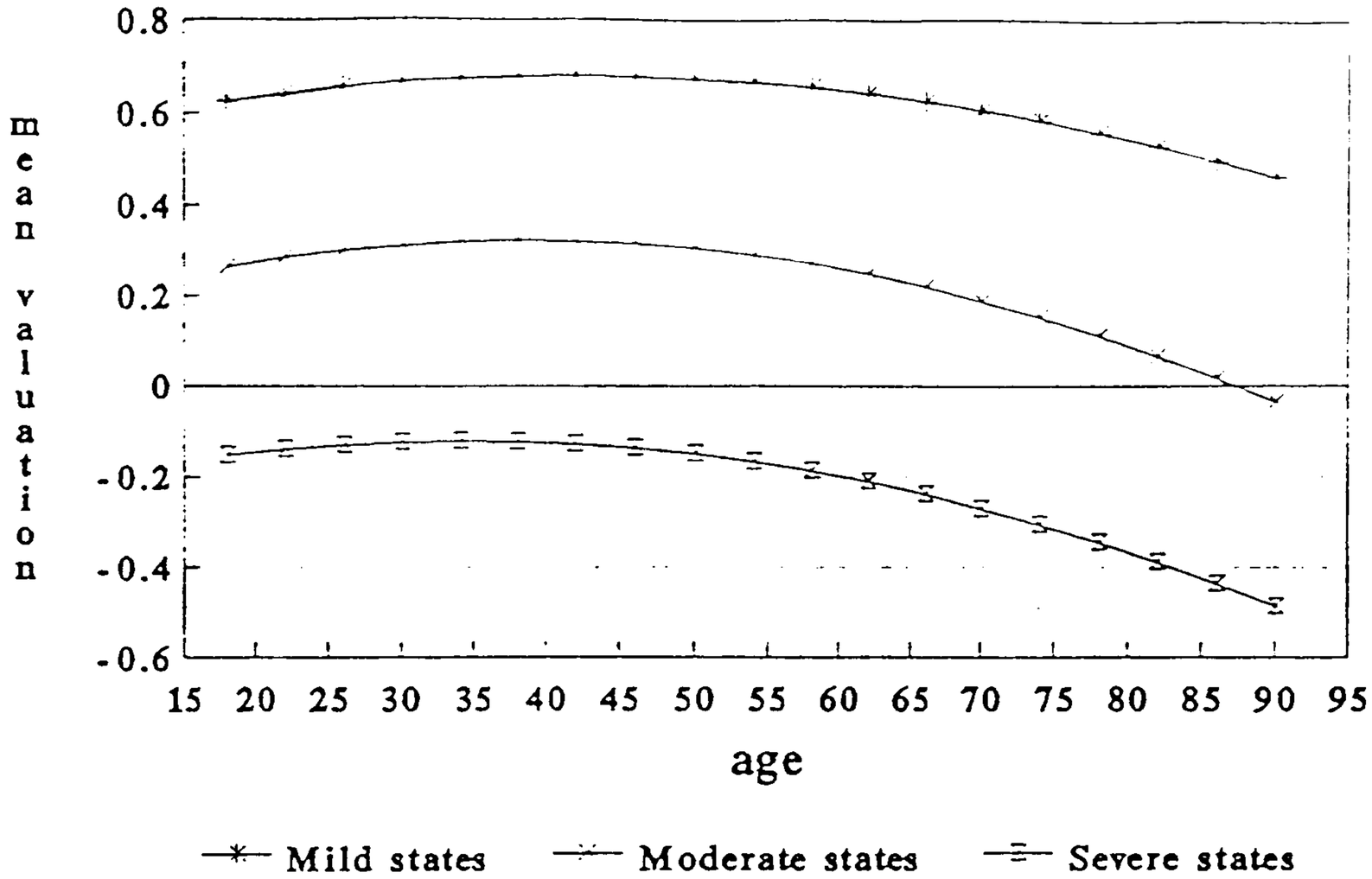
13212 32331 13311 22122 12222 21323 32211 12223 22331
21232 32313 22222

plus

3 from 12 "severe" states:

33232 23232 23321 13332 22233 22323 32223 32232 33321
33323 23313 33212

Figure 3.1.2: The effect of age on TTO valuations



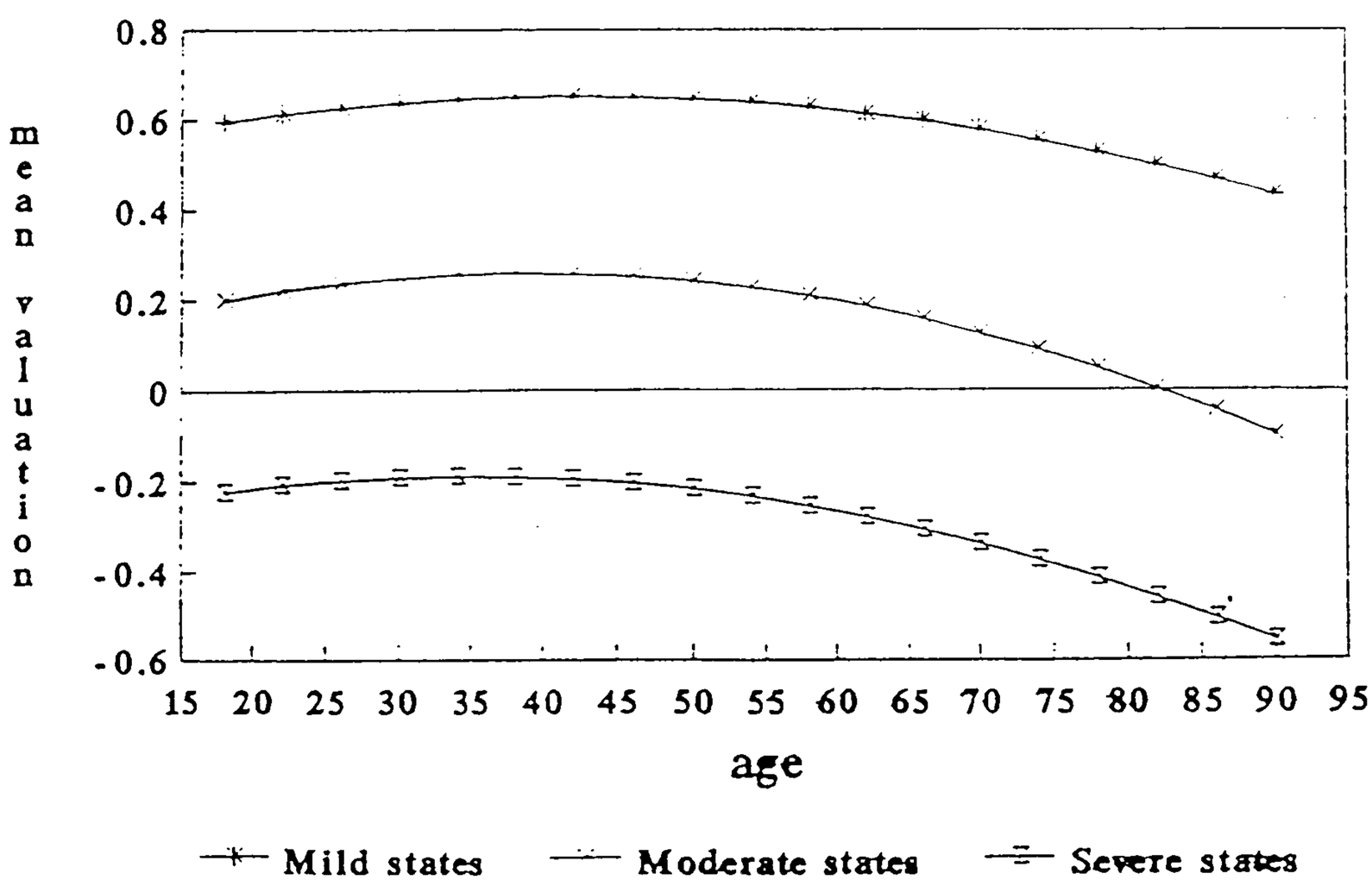


Figure 3.1.3: Difference between test and retest

Values are test median minus retest median

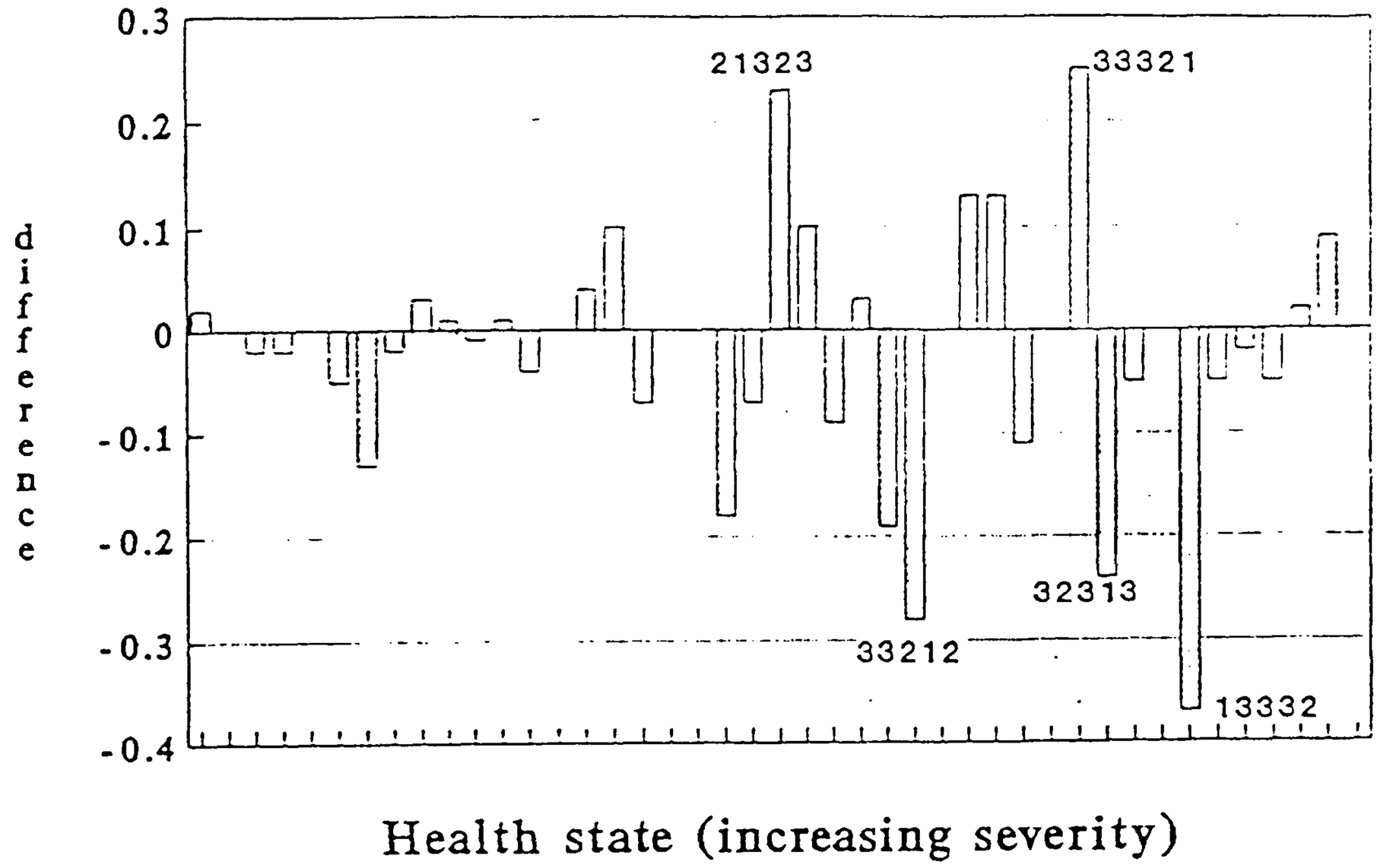


Figure 3.1.4

Distribution of ICCs

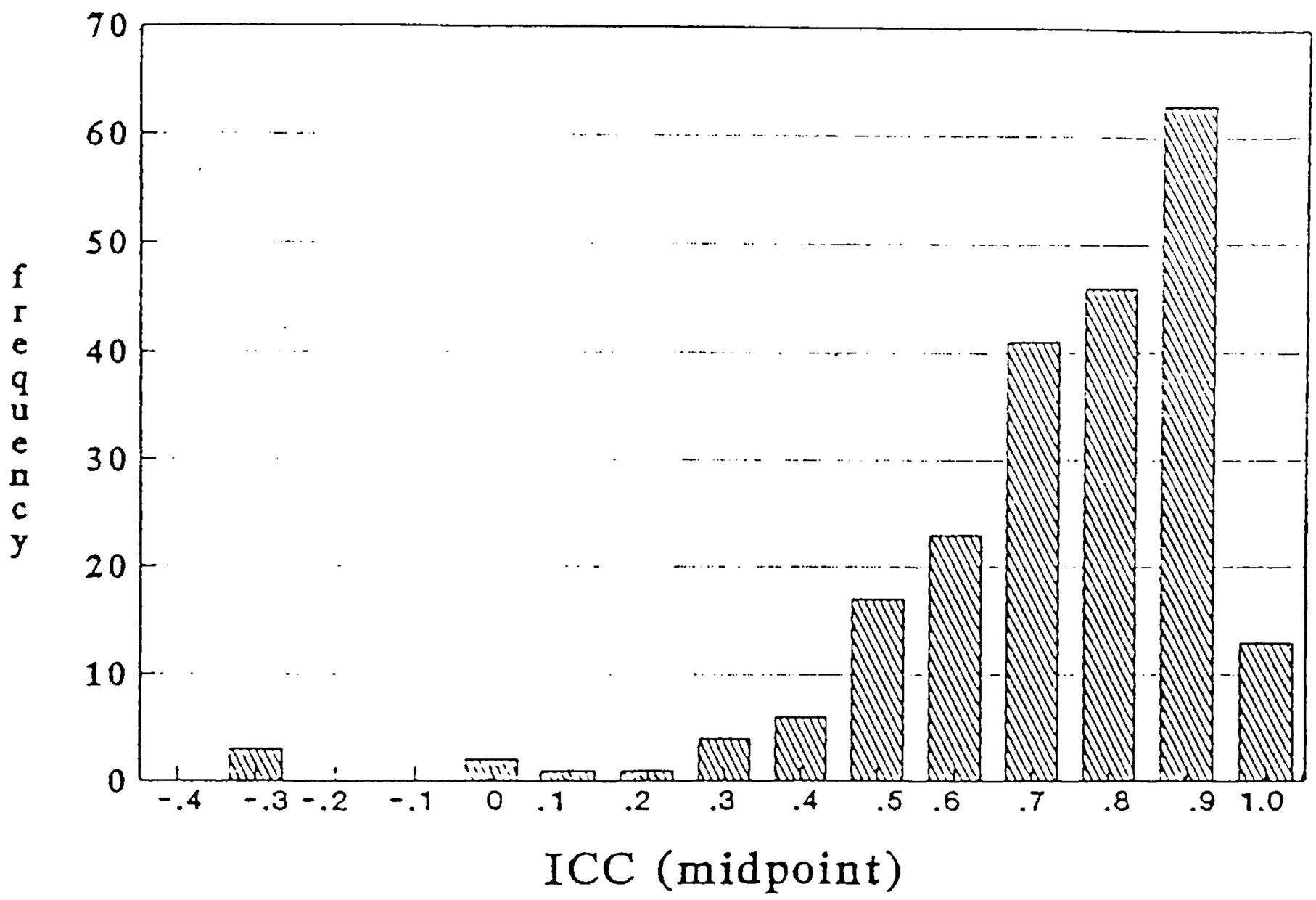


Figure 3.2.1 Variables used in the modelling

Variable	Definition
a	Constant: associated with any move away from full health
MO	1 if mobility is level 2; 2 if it is level 3; 0 otherwise
SC	1 if self-care is level 2; 2 if it is level 3; 0 otherwise
UA	1 if usual activities is level 2; 2 if it is level 3; 0 otherwise
PD	1 if pain/discomfort is level 2; 2 if it is level 3; 0 otherwise
AD	1 if anxiety/depression is level 2; 2 if it is level 3; 0 otherwise
M2	1 if mobility is level 3; 0 otherwise
S2	1 if self-care is level 3; 0 otherwise
U2	1 if usual activities is level 3; 0 otherwise
P2	1 if pain/discomfort is level 3; 0 otherwise
A2	1 if anxiety/depression is level 3; 0 otherwise
MOSC	The product of MO and SC
MOUA	The product of MO and UA
MOPD	The product of MO and PD
MOAD	The product of MO and AD
SCUA	The product of SC and UA
SCPD	The oroduct of SC and PD
SCAD	The Product of SC and AD
UAPD	The product of UA and PD
UAAD	The product of UA and AD
PDAD	The product of PD and AD
F11	1 if the health state contains 1 dimension at level 1; 0 otherwise
F21	1 if the health state contains 2 dimensions at level 1; 0 otherwise
F31	1 if the health state contains 3 dimensions at level 1; 0 otherwise
F41	1 if the health state contains 4 dimensions at level 1; 0 otherwise
F13	1 if the health state contains 1 dimension at level 3; 0 otherwise
F23	1 if the health state contains 2 dimensions at level 3; 0 otherwise
F33	1 if the health state contains 3 dimensions at level 3; 0 otherwise
F43	1 if the health state contains 4 dimensions at level 3; 0 otherwise
F53	1 if the health state contains 5 dimensions at level 3; 0 otherwise
N3	1 if any dimension is level 3; 0 otherwise

Figure 3.4.1: Comparison of mean and median values

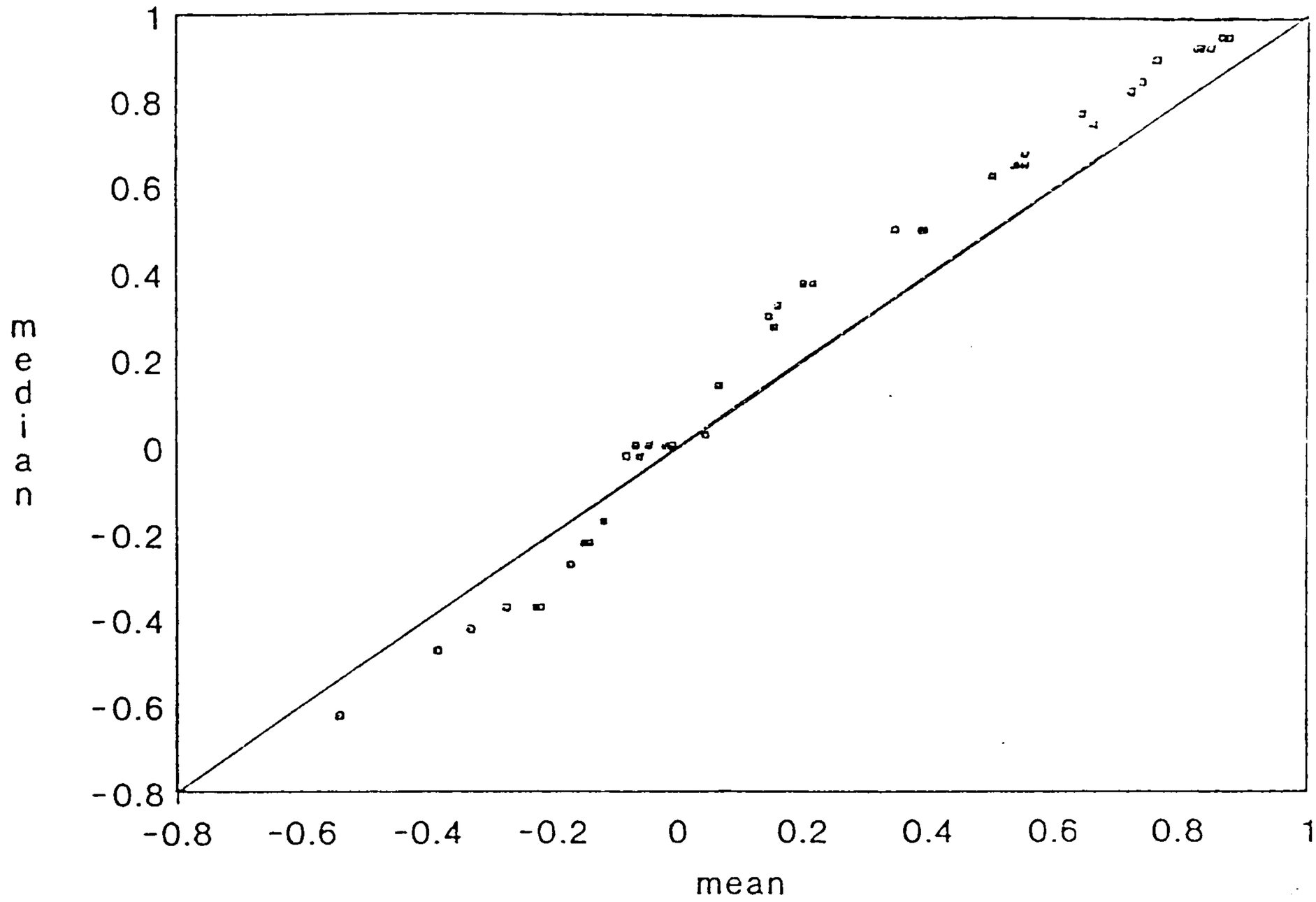


Figure 3.4.2: Comparison of mean and median tariffs

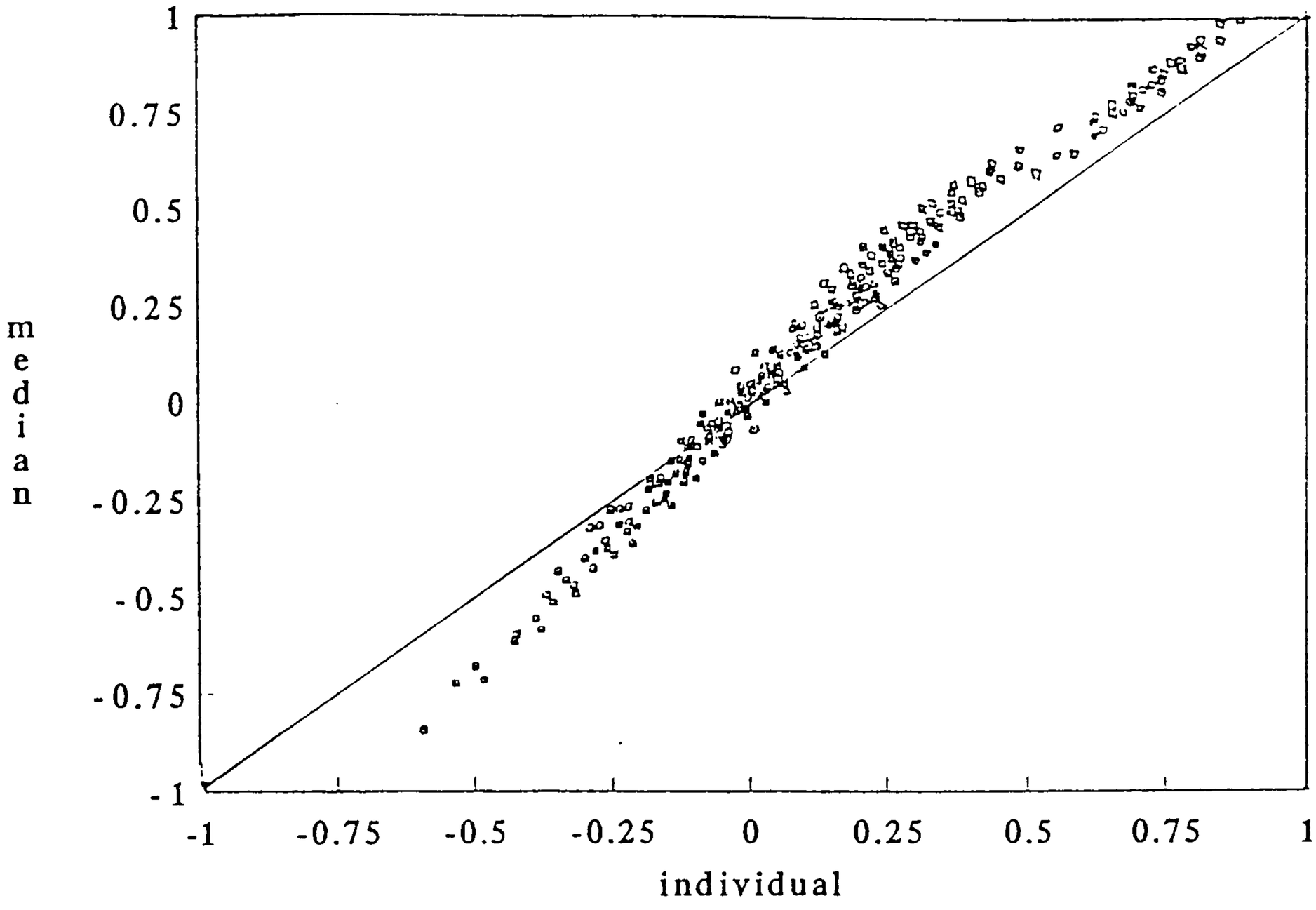


Figure 4.2.1: Distribution of discount rates in time preference study

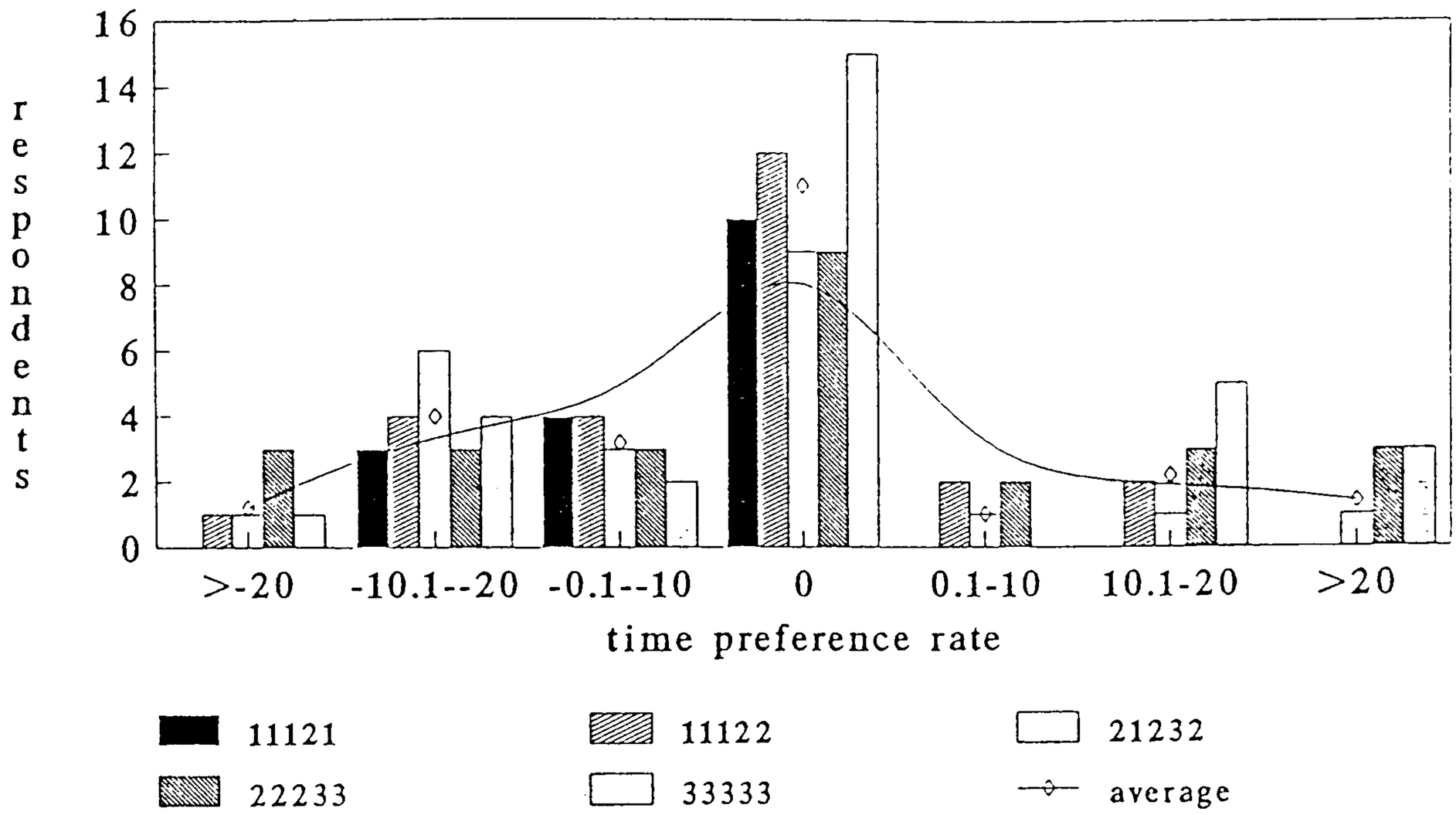


Figure 4.2.2: Median and profile values from time preference study

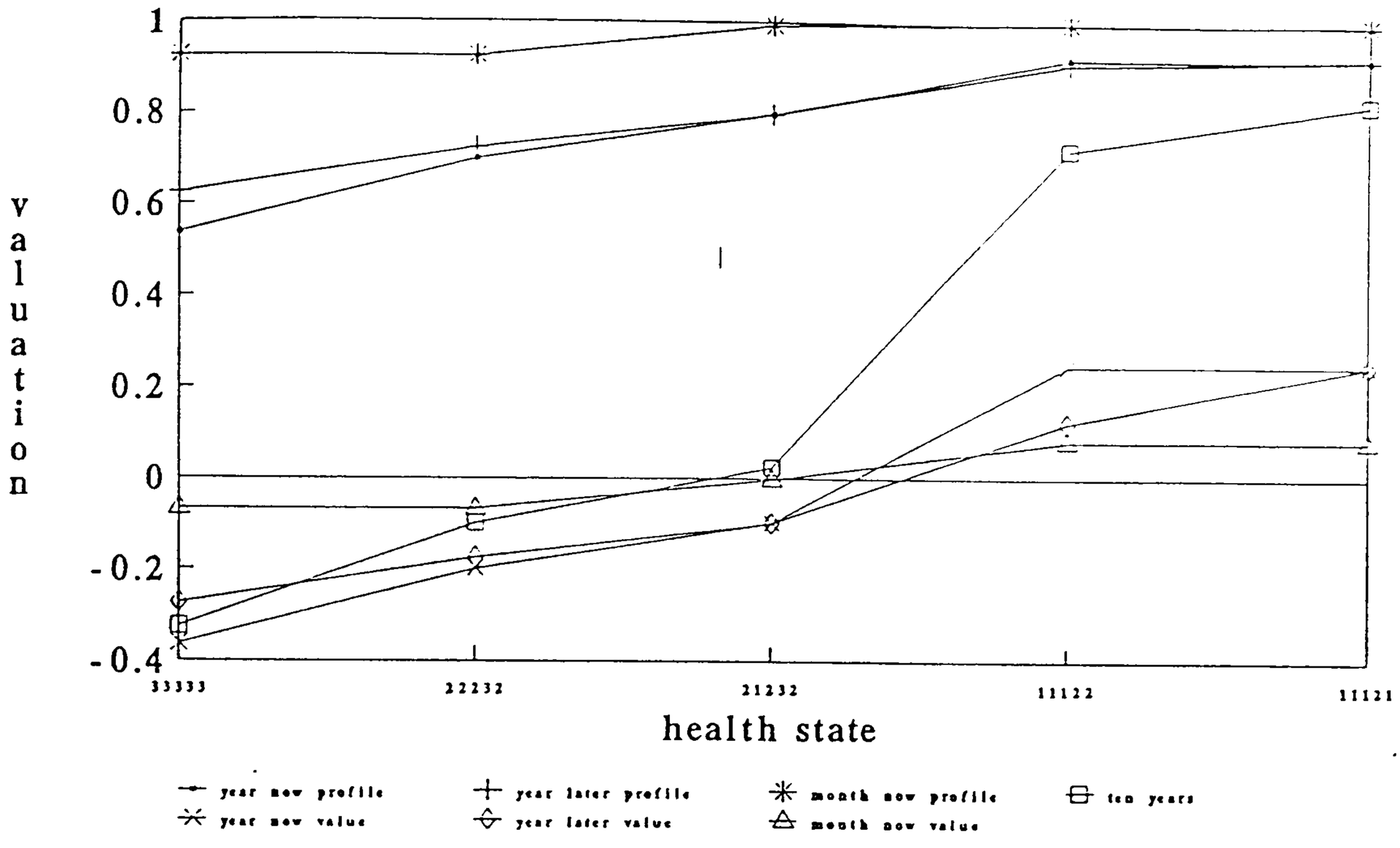
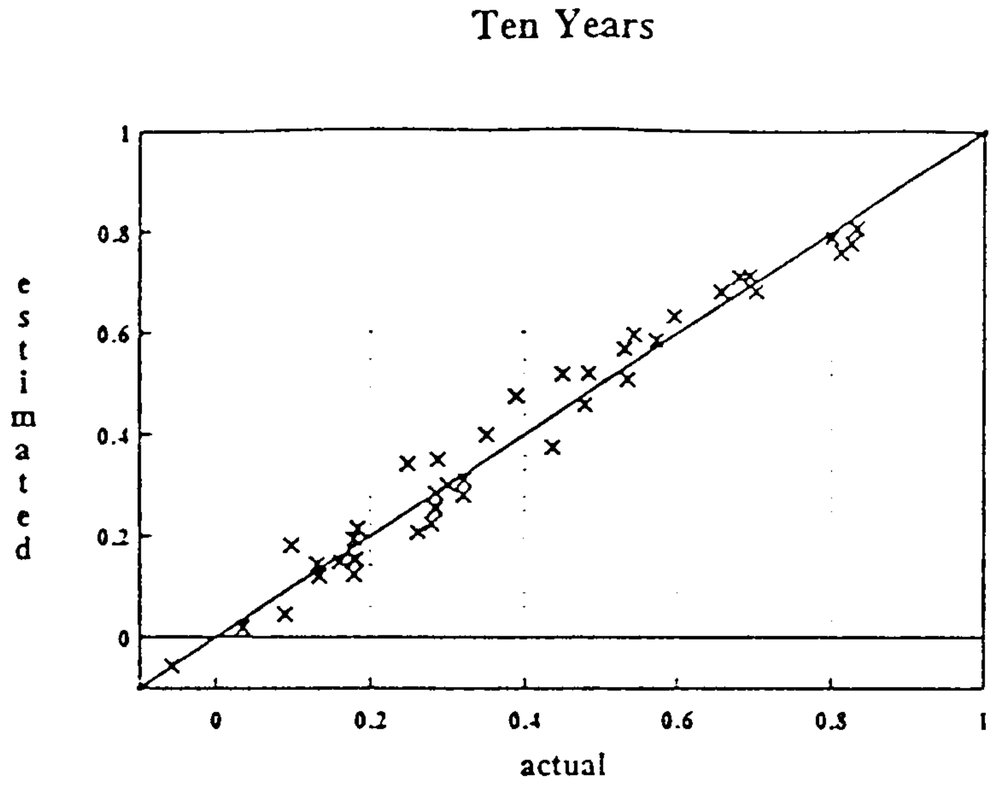
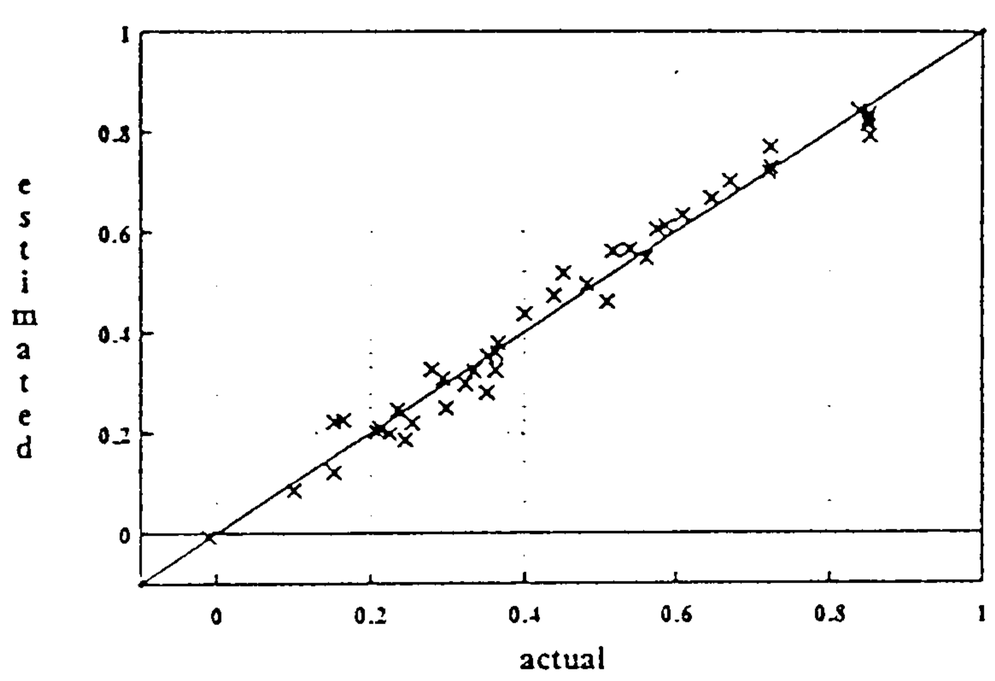


Figure 4.3.1: Actual and estimated values from the duration study





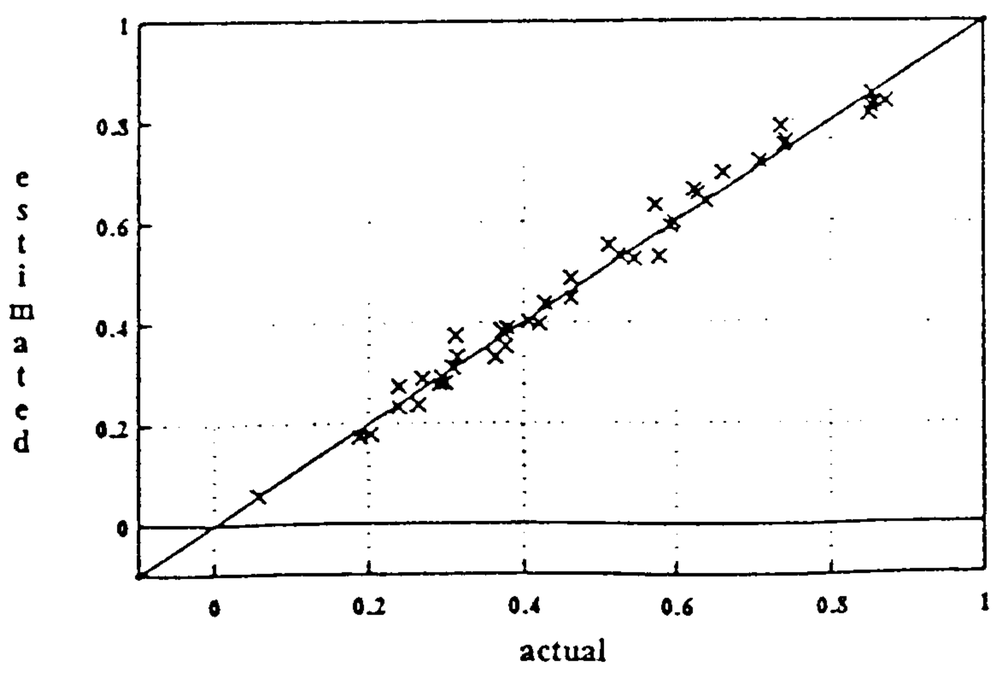


Figure 4.3.2: Estimated values for the three durations

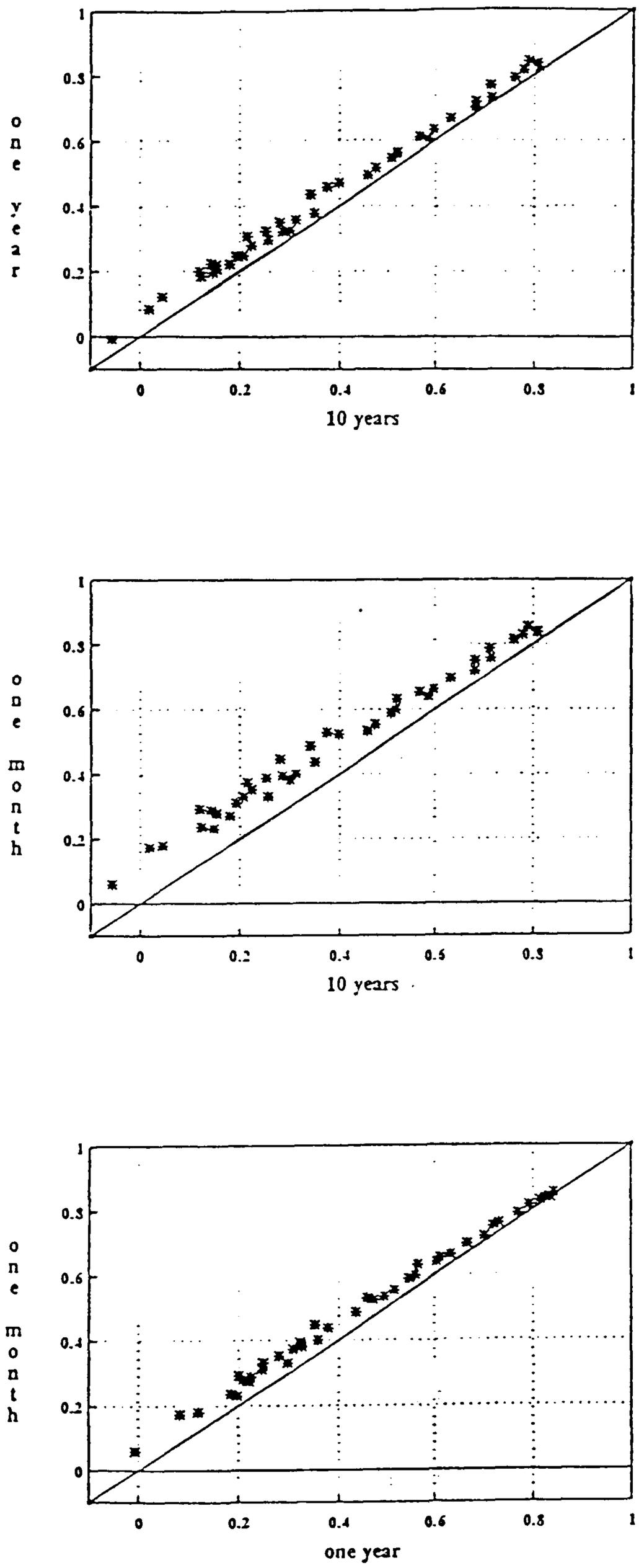
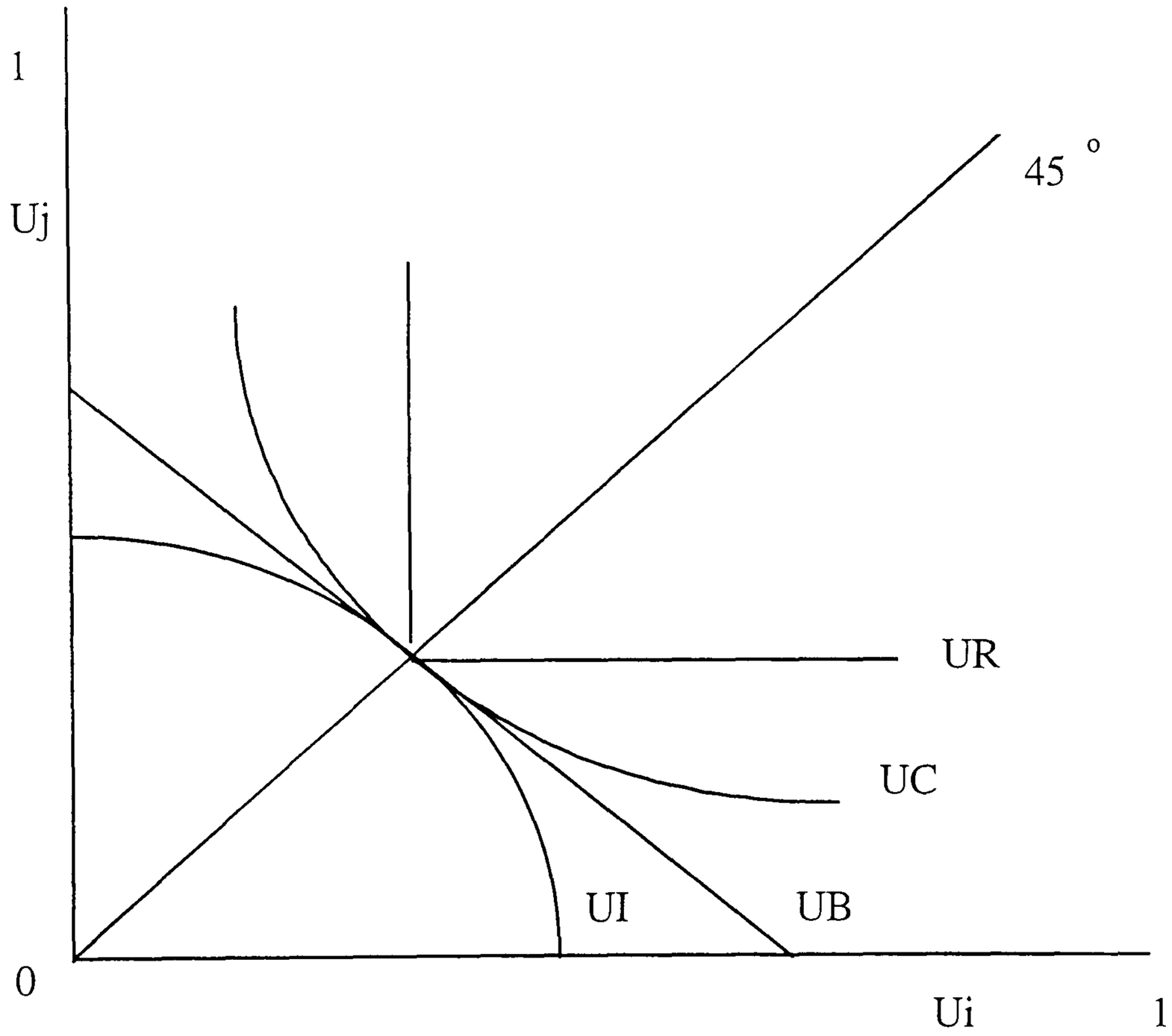


FIGURE 5.1.1: Different formulatives of the social welfare function (SWF)



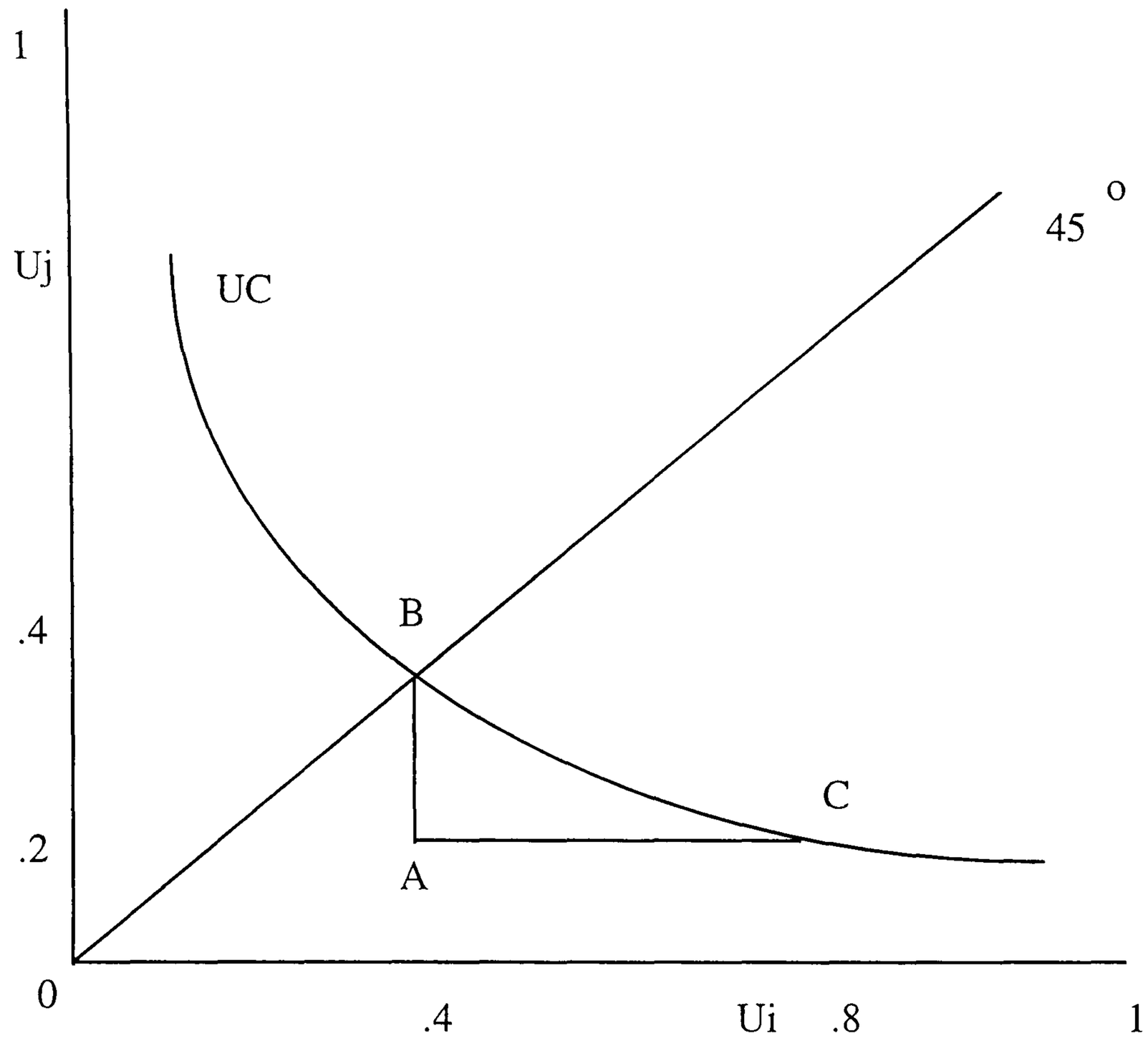
UR: Rawlsian SWF

UC: Convex SWF

UB: Utilitarian/Benthamite SWF

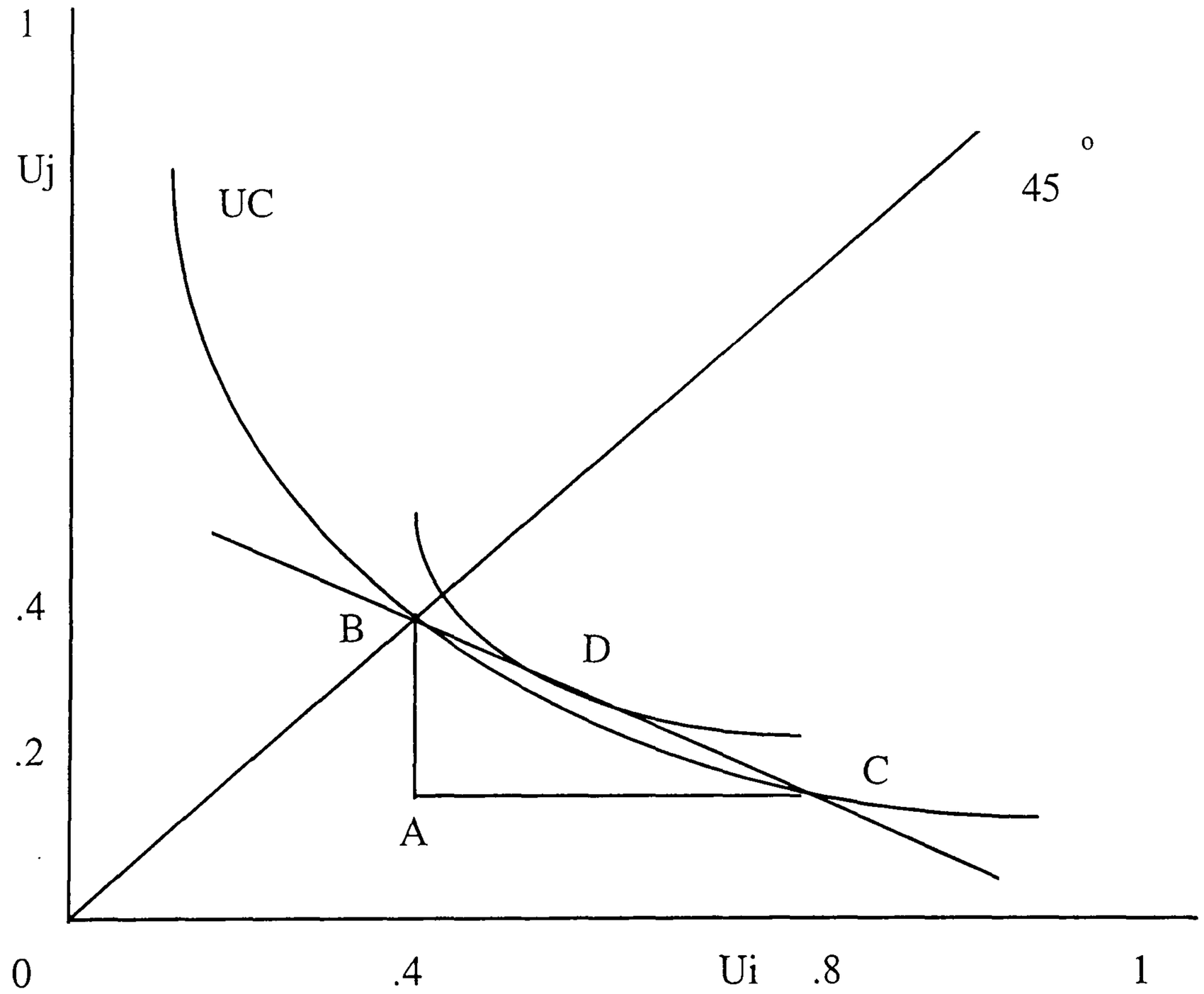
UI: Inequality-prone SWF

FIGURE 5.1.2: Example of the Cobb-Douglas SWF



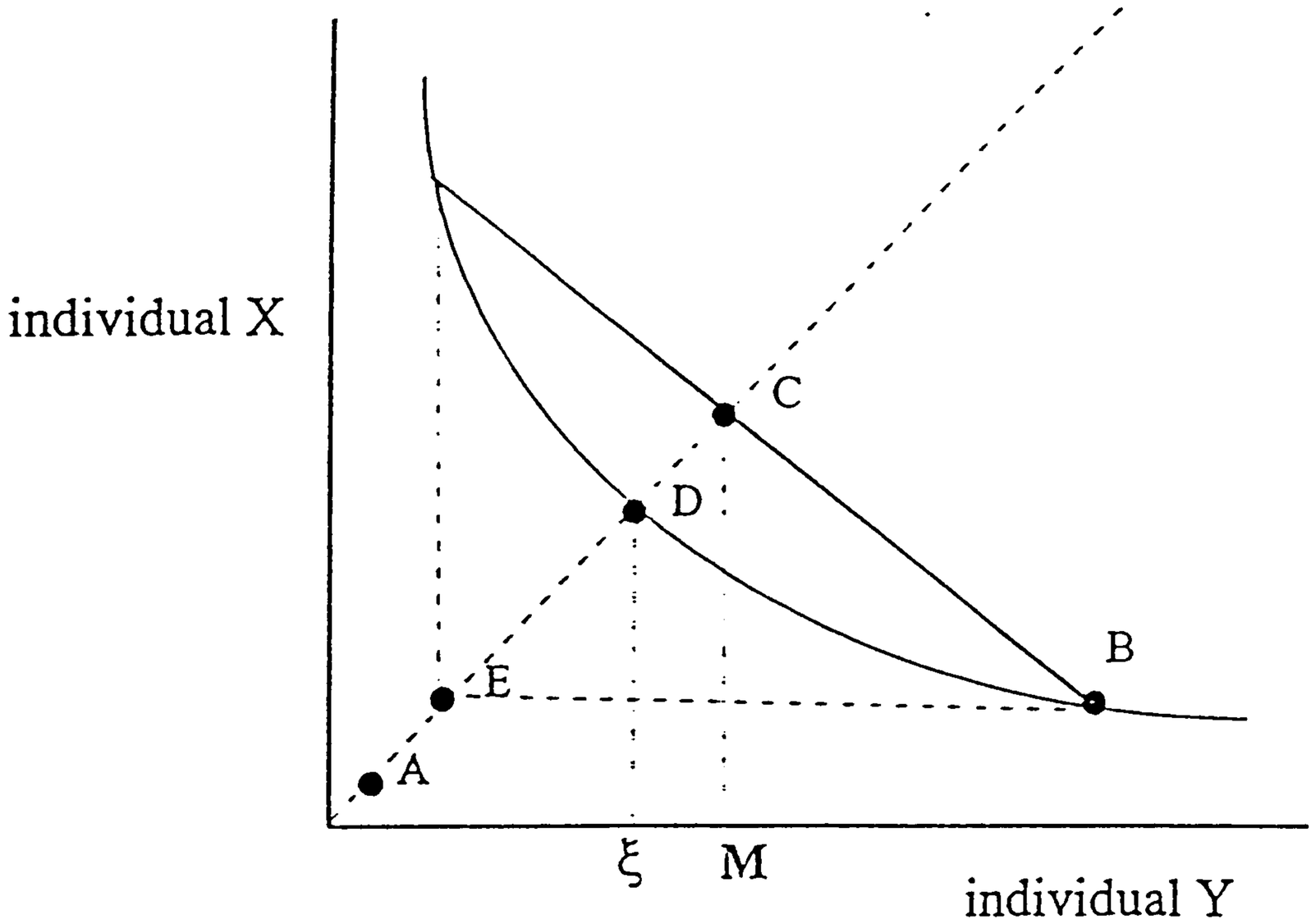
The curvature of UC shows that a move from 0.4 to 0.8 for person i yields the same social welfare as a move from 0.2 to 0.4 for person j.

FIGURE 5.1.3: Example of the Cobb-Douglas SWF with budget line



With this budget constraint, social welfare is maximised when person i's health improves from 0.2 to 0.3 and person j's health improves from 0.4 to 0.6.

Figure 5.2.1: Atkinson's social welfare function format



Moving from A to C yields the same health benefit as moving from A to B. But, if people are averse to inequality, then moving from A to D may yield the same social value as moving from A to B. The Atkinson Index in this case is measured as 1-ξ/M.

APPENDIX - STANDARD GAMBLE BOARD FOR A STATE RATED AS BETTER THAN DEAD

No problems in walking about

No problems with self-care

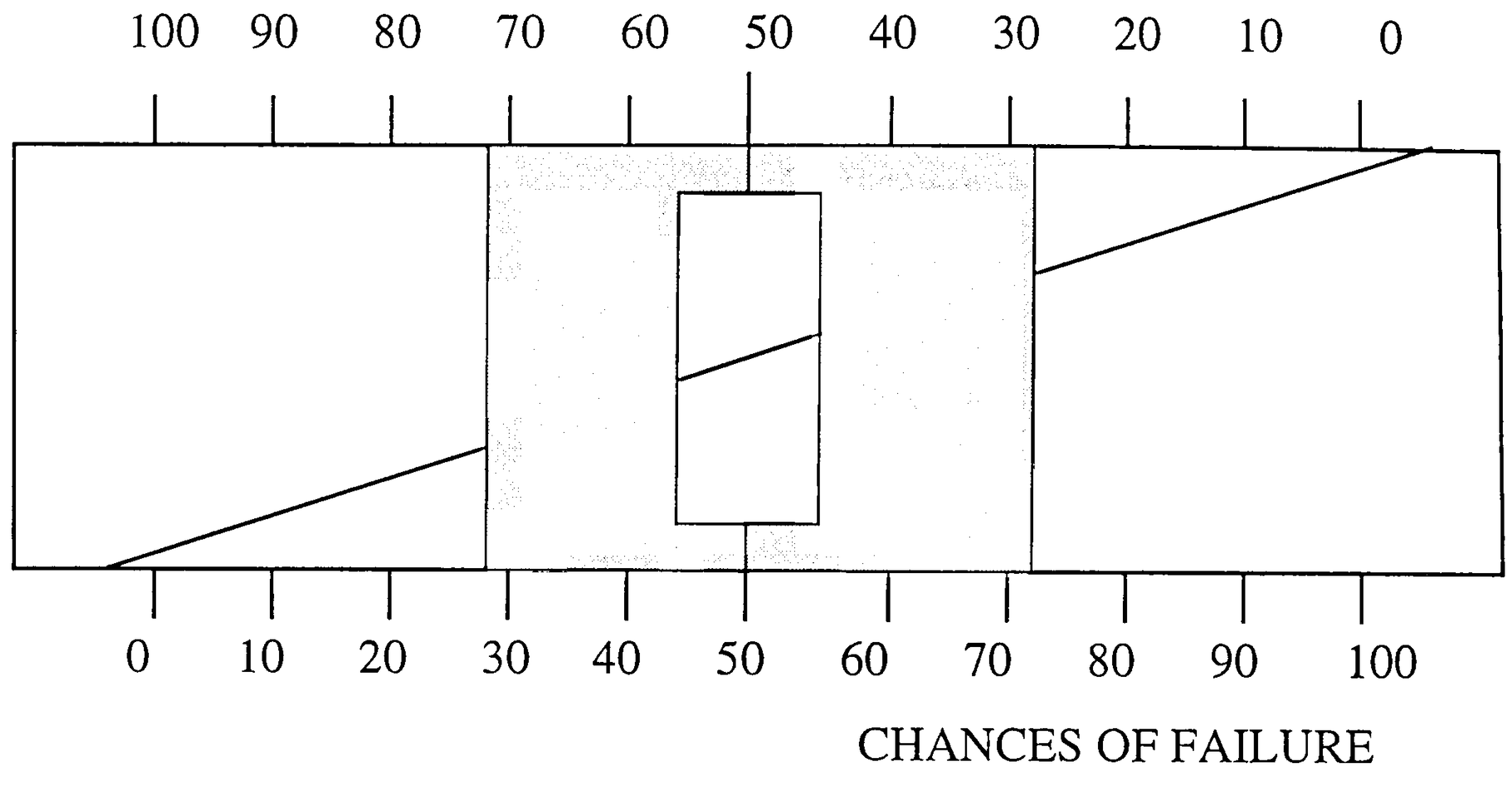
No problems with performing usual activities

No pain or discomfort

Not anxious or depressed

IMMEDIATE DEATH

CHOICE A



CHOICE B

100% CHANCE

STATE TO BE VALUED

260

APPENDIX - STANDARD GAMBLE BOARD FOR A STATE RATED AS WORSE THAN DEAD

No problems in walking about

No problems with self-care

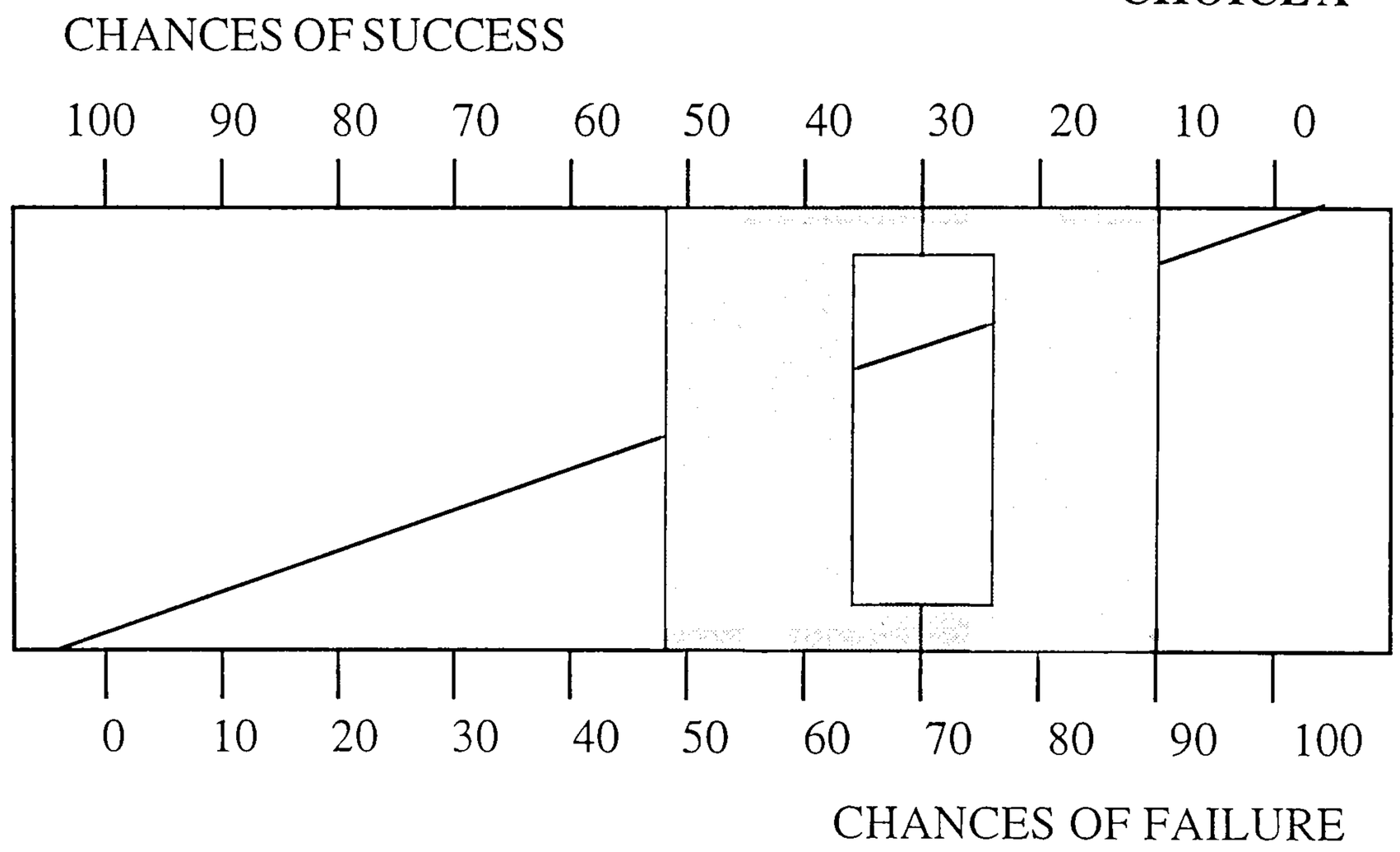
No problems with performing usual activities

No pain or discomfort

Not anxious or depressed

STATE TO BE VALUED

CHOICE A



100% CHANCE

IMMEDIATE DEATH

CHOICE B

APPENDIX - EXAMPLE SHOWCARD FOR SG NO PROPS

CHOICE "A"

EXAMPLE CARD

NO PROBLEMS IN WALKING ABOUT

NO PROBLEMS WITH SELF-CARE

NO PROBLEMS WITH PERFORMING USUAL ACTIVITIES

NO PAIN OR DISCOMFORT

NOT ANXIOUS OR DEPRESSED

EXAMPLE CARD

IMMEDIATE DEATH

CHOICE B

100% CHANCE

EXAMPLE CARD

SOME PROBLEMS IN WALKING ABOUT

SOME PROBLEMS IN WASHING OR DRESSING SELF

SOME PROBLEMS WITH PERFORMING USUAL ACTIVITIES (e.g. work, study, housework, family or leisure activities)

MODERATE PAIN OR DISCOMFORT

MODERATELY ANXIOUS OR DEPRESSED

APPENDIX - EXAMPLE ANSWER SHEET FOR SG NO PROPS

THE CHANCES IN CHOICE A:

Chances of Success	Chances of Failure		CHOICE B:
100 in 100	0 in 100	√	100 in 100
95 in 100	5 in 100	√	100 in 100
90 in 100	10 in 100	=	100 in 100
85 in 100	15 in 100	x	100 in 100
80 in 100	20 in 100	x	100 in 100
75 in 100	25 in 100	x	100 in 100
70 in 100	30 in 100	x	100 in 100
65 in 100	35 in 100	x	100 in 100
60 in 100	40 in 100	x	100 in 100
55 in 100	45 in 100	x	100 in 100
50 in 100	50 in 100	x	100 in 100
45 in 100	55 in 100	x	100 in 100
40 in 100	60 in 100	x	100 in 100
35 in 100	65 in 100	x	100 in 100
30 in 100	70 in 100	x	100 in 100
25 in 100	75 in 100	x	100 in 100
20 in 100	80 in 100	x	100 in 100
15 in 100	85 in 100	x	100 in 100
10 in 100	90 in 100	x	100 in 100
5 in 100	95 in 100	x	100 in 100
0 in 100	100 in 100	x	100 in 100

Please put a √ against all cases where you are CONFIDENT that you would choose the risky treatment in Choice A.

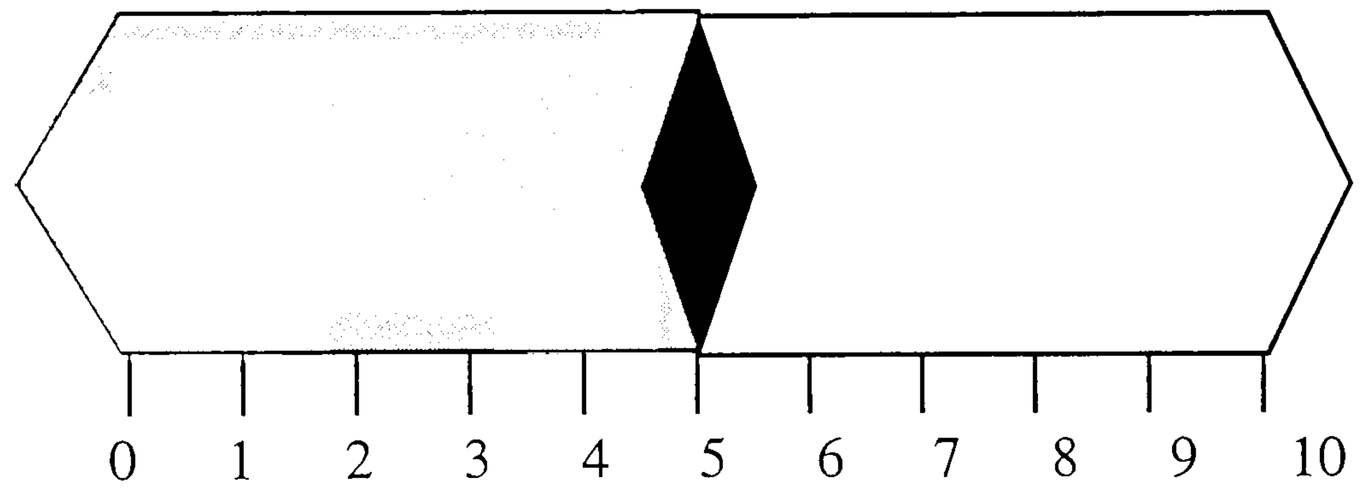
Please put an x against all cases where you are CONFIDENT that you would REJECT the treatment and accept the health state in Choice B.

Please put a = against the case where you think it would be most difficult to choose between having the risky treatment (Choice A) and not having the treatment (Choice B).

APPENDIX - TIME TRADE-OFF BOARD FOR A STATE RATED AS BETTER THAN DEAD

LIFE A

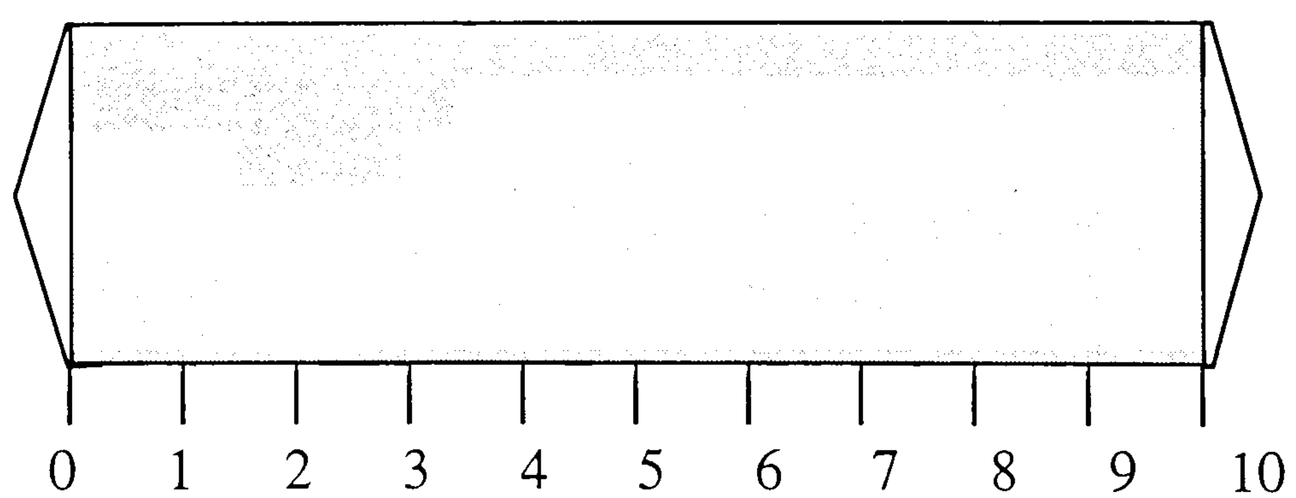
- ☐ No problems in walking about
- ☐ No problems with self-care
- ☐ No problems with performing usual activities
- ☐ No pain or discomfort
- ☐ Not anxious or depressed



Number of Years

STATE TO BE VALUED

LIFE B

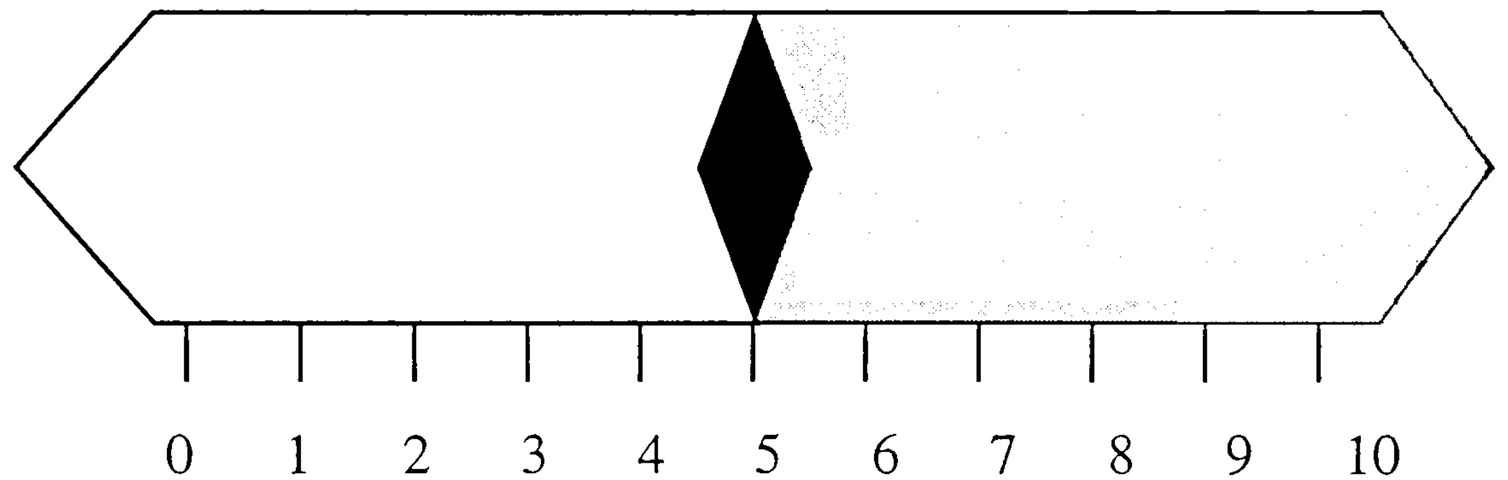


Number of Years

APPENDIX - TIME TRADE-OFF BOARD FOR A STATE RATED AS WORSE THAN DEAD

LIFE A

STATE TO BE VALUED



- ☐ No problems in walking about
- ☐ No problems with self-care
- ☐ No problems with performing usual activities
- ☐ No pain or discomfort
- ☐ Not anxious or depressed

LIFE B

IMMEDIATE DEATH

APPENDIX - EXAMPLE OF SHOWCARD FOR TTO NO PROPS

LIFE "A"

EXAMPLE CARD

NO PROBLEMS IN WALKING ABOUT

NO PROBLEMS WITH SELF-CARE

NO PROBLEMS WITH PERFORMING USUAL ACTIVITIES

NO PAIN OR DISCOMFORT

NOT ANXIOUS OR DEPRESSED

LIFE "B"

EXAMPLE CARD

SOME PROBLEMS IN WALKING ABOUT

SOME PROBLEMS WITH WASHING OR DRESSING SELF

SOME PROBLEMS WITH PERFORMING USUAL ACTIVITIES (e.g. work, study, housework, family or leisure activities)

MODERATE PAIN OR DISCOMFORT

MODERATELY ANXIOUS OR DEPRESSED

APPENDIX - EXAMPLE ANSWER SHEET FOR TTO NO PROPS

10 YEARS	√	10 YEARS
9 YEARS 6 MONTHS	√	10 YEARS
9 YEARS	=	10 YEARS
8 YEARS 6 MONTHS	x	10 YEARS
8 YEARS	x	10 YEARS
7 YEARS 6 MONTHS	x	10 YEARS
7 YEARS	x	10 YEARS
6 YEARS 6 MONTHS	x	10 YEARS
6 YEARS	x	10 YEARS
5 YEARS 6 MONTHS	x	10 YEARS
5 YEARS	x	10 YEARS
4 YEARS 6 MONTHS	x	10 YEARS
4 YEARS	x	10 YEARS
3 YEARS 6 MONTHS	x	10 YEARS
3 YEARS	x	10 YEARS
2 YEARS 6 MONTHS	x	10 YEARS
2 YEARS	x	10 YEARS
1 YEARS 6 MONTHS	x	10 YEARS
1 YEAR	x	10 YEARS
0 YEARS 6 MONTHS	x	10 YEARS
0 YEARS	x	10 YEARS

Place a '√' if you prefer Life "A"

Place a 'x' if you prefer Life "B"

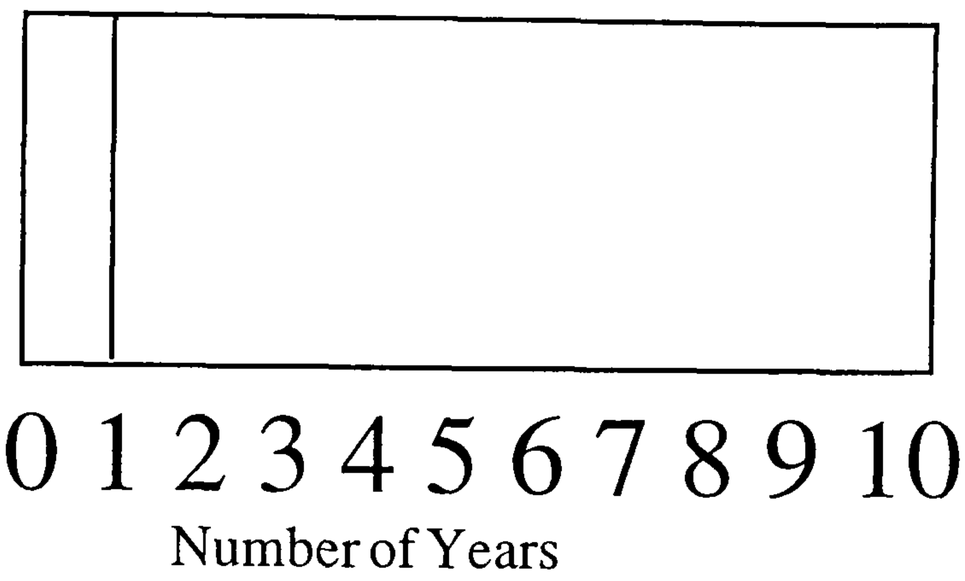
Place a '=' if you cannot choose between Life "A" and Life "B"

APPENDIX - PRESENTATION OF ALTERNATIVES IN 'TIME PREFERENCE' STUDY

1 YEAR IN THIS STATE: **SP1**

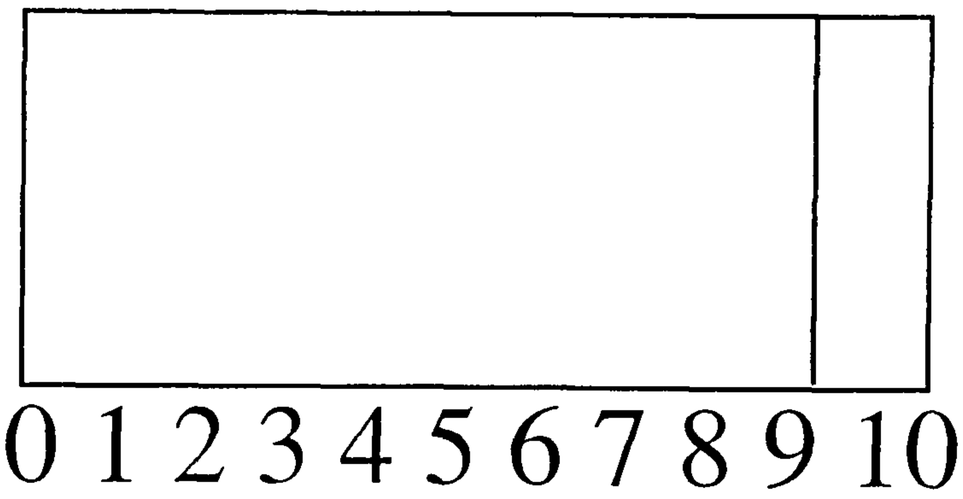
- ☐ No problems in walking about
- ☐ No problems with self-care
- ☐ No problems with performing usual activities
- ☐ Moderate pain or discomfort
- ☐ Not anxious or depressed

FOLLOWED BY 9 YEARS IN STATE 'AP'



9 YEARS IN STATE 'AP' FOLLOWED BY 1 YEAR IN THIS STATE: **SP2**

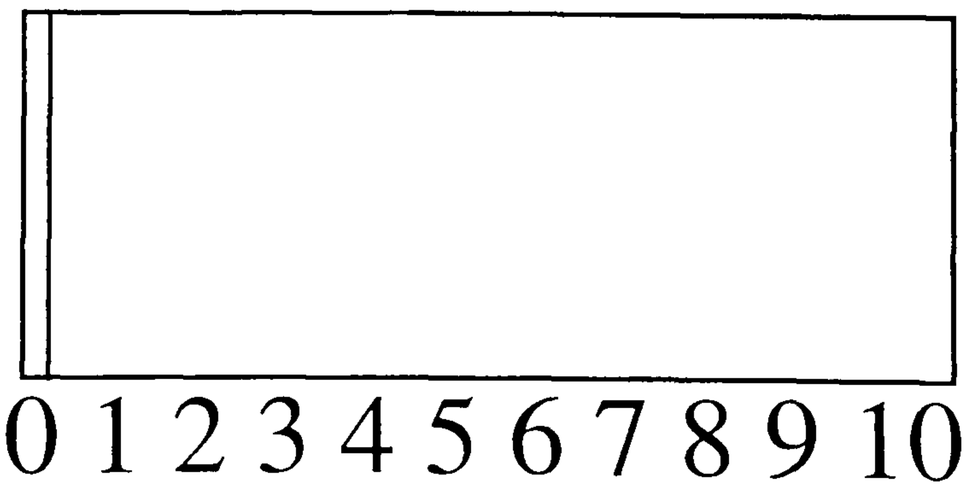
- ☐ No problems in walking about
- ☐ No problems with self-care
- ☐ No problems with performing usual activities
- ☐ Moderate pain or discomfort
- ☐ Not anxious or depressed



1 MONTH IN THIS STATE: **SP3**

- ☐ No problems in walking about
- ☐ No problems with self-care
- ☐ No problems with performing usual activities
- ☐ Moderate pain or discomfort
- ☐ Not anxious or depressed

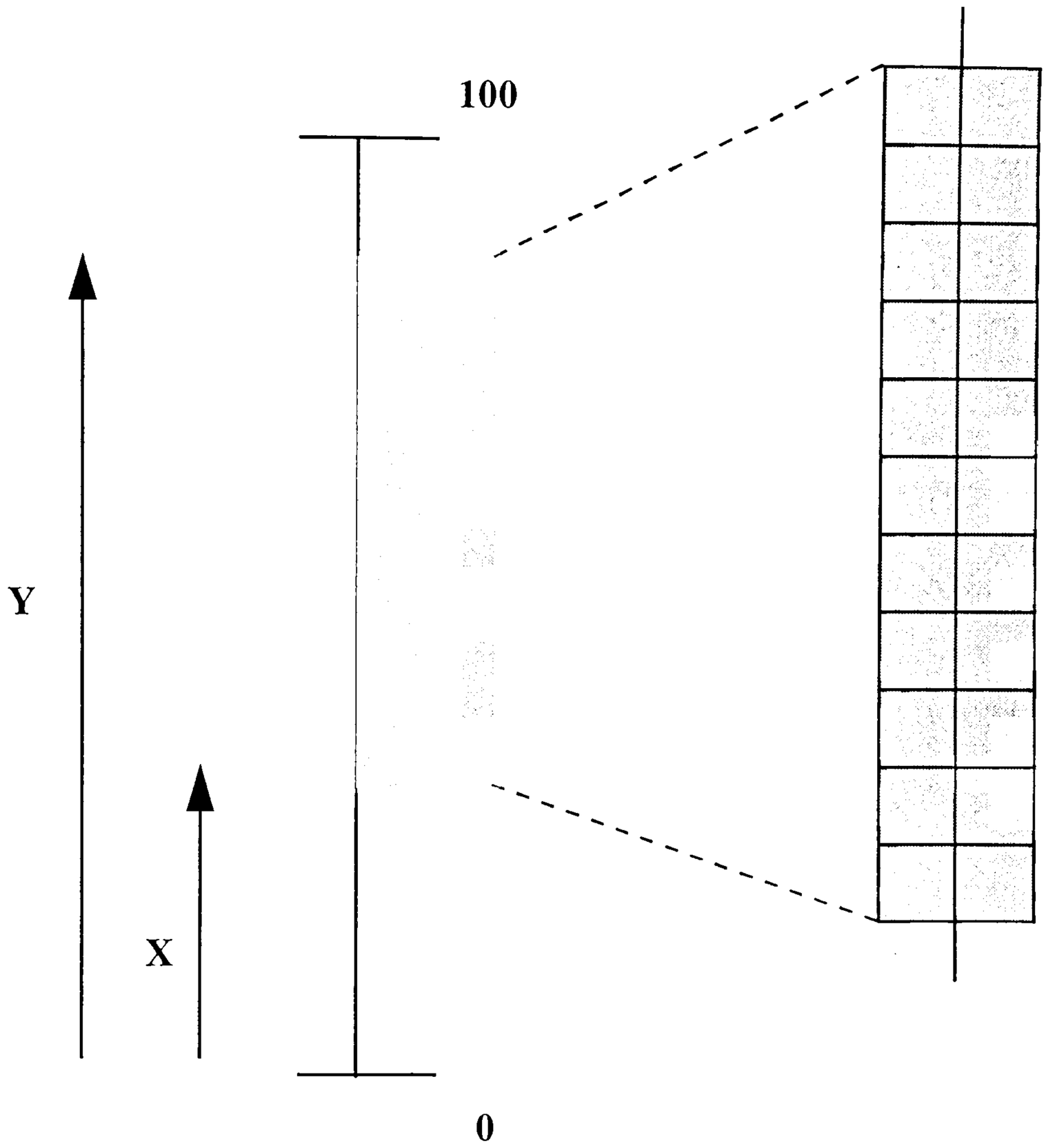
FOLLOWED BY 9 YEARS AND 11 MONTHS IN STATE 'AP'



Where state 'AP' refers to EuroQol state 11111.

Each of these alternatives was presented in 'Life B' on the TTO board shown on page 265 and was compared to 'Life A' (i.e.. years in full health) in the standard way. This was repeated for each of the five health states.

APPENDIX - EXAMPLE FORMAT OF QUESTIONS IN THE ATKINSON INDEX STUDY



Respondents were asked to place a tick on the scale on the right to indicate indifference between this common outcome for X and Y and the different outcomes shown on the left.

APPENDIX - BOARD USED IN THE 'PERSON TRADE-OFF' STUDY

NUMBER OF PEOPLE	TREATMENT 1
	A ➡ B

NUMBER OF PEOPLE	TREATMENT 2
	C ➡ D

REFERENCES

Arrow KJ (1951) Social choice and individual values, Wiley, New York.

Arrow K, Solow R, Portney PR, Leamer EE, Radner R and Schuman H (1993) Report of the NOAA panel on contingent valuation, Resources for the Future, Washington DC.

Atkinson AB (1970) On the measurement of inequality, Journal of Economic Theory 2, 244-263.

Bennet K, Torrance GW, and Tugwell P (1991) Methodologic challenges in the development of utility measures of health related quality of life in rheumatoid arthritis?, Controlled Clinical Trials 12.

Bergner M, Bobbitt RA, Carter W and Gilson BS (1981) The sickness impact profile: development and final revision of a health status measure, Medical Care 36, 428-434.

Bergson A (1938) A reformulation of certain aspects of welfare economics, Quarterly Journal of Economics 52, 310-34.

Berzon R and Shumaker S (1993) A critical review of cross national health-related quality of life instruments, Quality of Life Newsletter 5, 1-2.

Bland JM and Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement, Lancet i, 307-310.

Bombardier C, Wolfson AD, Sinclair AJ and McGreer A (1982) comparison of three preference measurement methodologies in the evaluation of a functional status index, in Choices on Health Care: Decision-Making and Evaluation of Effectiveness, Deber RB and Thompson CG.

Boyd NF, Sutherland HJ, Ciampi A, Tibshirani R, Till JE, Harwood A (1982) A comparison of methods of assessing voice quality in laryngeal cancer, in Choices in health care: decision-making and evaluation of effectiveness, Department of Health Administration, University of Toronto.

Brazier J, Jones N and Kind P (1993) Testing the validity of the Euroqol and comparing it with the SF36 health survey questionnaire, Quality of Life Research 2, 169-180.

Brooks RG (1991) Health status and quality of life measurement: issues and developments, Swedish Institute for Health Economics, IHE, Lund.

Broome J (1991) Weghing Goods, Basil Blackwell, Oxford.

Broome J (1993), Qalys, Journal of Public Economics, 50, 149-167.

Buckingham K and Drummond M (1993) A theoretical and empirical classification of health valuation techniques, HESG Conference, Strathclyde.

Bullinger M (1991) Quality of life: Definition, conceptualization and implications - a methodologist's view, Theoretical Surgery 6, 143-148.

Burrows C and Brown K (1992) Time perception: some implications for the development of scale values in measuring health status and quality of life, National Centre for Health Program Evaluation Working Paper No. 15, Melbourne, Australia.

Cairns J (1992) Future discounting: health, wealth and time preference, Project Appraisal 7(1), 31-40.

Cairns J, Johnston K and McKenzie L (1991) Developing QALYs from condition-specific outcome measures, HERU Working Paper 14, University of Aberdeen.

Carson RT, Mitchell RC, Hanemann WM, Kopp RJ, Presser S and Ruud PA (1992) A contingent valuation study of lost passive use values resulting from the Exxon Valdez oil spill, Report to the Attorney General of the State of Alaska, Westat Inc, Rockville, MD.

Carter WB, Bobbitt RA, Bergner M, Gibson BS (1976) Validation of interval scaling; the sickness impact profile, Health Services Research 516-528.

Cassileth BR, Lusk EJ, Strouse TB, Miller DS, Brown LL, Cross PA and Tenaglia AN (1984) Psychosocial status in chronic illness: a comparative analysis of six diagnostic groups, New England Journal of Medicine 311, 506-511.

Charny MC, Lewis PA and Farrow SC (1989) Choosing who shall not be treated in the NHS, Social Science and Medicine 28 (12), 1331-8.

Christensen-Szalanski JJ (1984) Discount functions and the measurement of patients' values: women's decisions during childbirth, Medical Decision Making 4, 45-58.

Churchill DN, Morgan J and Torrance GW (1984), Quality of life in end-stage renal disease, Peritoneal Dialysis Bulletin 4, 20-23.

Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizu A, Smith EKM (1987) Measurement of quality-of-life in end-stage renal disease: the time trade-off approach, Clinical and Investigative Medicine 10, 439-471.

Copas JB (1983) Regression, prediction and shrinkage, Journal of the Royal Statistical Society B, 45, 311-54.

Culyer AJ (1976) Need and the national health service, Martin Robertson, Oxford.

Culyer AJ (1980) The politcal economy of social choice, Martin Robertson, Oxford.

Culyer AJ (1995) Equality of what in health policy? conflicts between the contenders, CHE Discussion Paper, 142, University of York.

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

Daniels N (1985) Just Health Care, Cambridge University Press.

Dworkin R (1977) Taking rights seriously, Cambridge University Press.

Dyer JS and Sarin RK (1982) Relative risk aversion, Management Science 28 (8), 857-86.

Erens R (1994) Health-related quality of life, SCPR Technical Report, London.

Essink-Bot ML (1990) Valuations of health states by the general public: Feasibility of a standardised measurement procedure, Social Science and Medicine 31, 1201-1206.

Etzioni A (1986) The case for a multiple utility conception, Economics and Philosophy 2 (159-83).

Euroqol Group (1990) Euroqol: A new facility for the measurement of health related quality of life, Health Policy 16, 199-208.

Feeny D, Furlong W, Boyle M and Torrance G. Multi-attribute health status classification systems, PharmacoEconomics,7(6), 490-502.

Festinger L (1957) A theory of cognitive dissonance, Stanford University Press.

Fischoff B (1991) Value elicitation: is there anything there?, American Psychologist 46, 835-847.

Fischhoff B and Furby L (1988) Measuring values: a conceptual framework for interpreting transactions with special reference to contingent valuation of visibility, Journal of Risk and Uncertainty 1, 147-184.

Fishburn PC and Kochenberger GA (1979) Two piece Von-Neumann Morgenstern utility functions, Decision Sciences 10, 503-518.

Fisher I (1930) The Theory of Interest, MacMillan, New York.

Froberg DG and Kane RL (1989) Methodology for measuring health state preferences II: scaling methods, Journal of Clinical Epidemiology 42(5).

Froberg DG and Kane RL (1989) Methodology for measuring health state preferences III: population and context effects, Journal of Clinical Epidemiology 42, 585-592.

Froberg DG and Kane RL (1989) Methodology for measuring health state preferences IV: progress and a research agenda, Journal of Clinical Epidemiology 42, 675-685.

Fuchs VR (1982) Time preference and health: an exploratory study, in Economic Aspects of Health, Fuchs VR, University of Chicago Press: Chicago, 93-120.

Furlong W, Feeny D, Torrance GW, Barr R and Horsman J (1990) Guide to design and development of health-state utility instrumentation, CHEPA Working Paper, McMaster University, Hamilton, Ontario, Canada.

Gafni A (1994) The standard gamble method: what is being measured and how is it interpreted?, Health Services Research, 29 (2), 207-224.

Gafni A and Torrance GW (1984) Risk attitude and time preference in health, Management Science 30.

Gafni A, Birch S and Mehrez A (1993) Economics, health and health economics: HYEs versus QALYs, Journal of Health Economics 11, 325-339.

Gravelle H. and Rees R (1981) Microeconomics, Longman, London.

Greene WH (1992) LIMDEP version 6.0: User's manual and reference guide, Econometric Software Inc., New York.

Gudex C (1994) Standard gamble user manual: props and self-completion method, Centre for Health Economics, University of York, Occasional Paper Series.

Gudex C (1994) Time trade-off user manual: props and self-completion method, Centre for Health Economics, University of York, Occasional Paper Series.

Gudex C, Dolan P, Kind P and Williams A (1996) Health state valuations from the general public: using the visual analogue scale, Quality of Life Research.

Hausman JA (1993) Contingent valuation: a critical assessment, North Holland, New York.

Hornberger JC, Redelmeier DA and Petersen J (1992) Variability among methods to assess patients well-being and consequent effect on a cost-effectiveness analysis, Journal of Clinical Epidemiology 45(5), 505-512.

Hunt SM, Alonso J, Bucquet D, Niero M, Wiklund I and Mckenna S (1991) Cross-cultural adaptations of health measures, Health Policy 19, 33-44.

Johannesson M and Gerdtham UG (1996) A note on the estimation of the equity-efficiency trade-off for QALYs, Journal of Health Economics 15, 359-368.

Jones-Lee MW (1976) The value of life: an economic analysis, University of Chicago Press.

Jones-Lee MW (1989) The economics of safety and physical risk, Oxford, Basil Blackwell.

Jones-Lee MW (1989) The value of avoidance of non-fatal road injuries, Report to the Department of Transport.

Jones-Lee MW, Loomes G and Philips PR (1993) Valuing the prevention of non-fatal road injuries: contingent valuation vs standard gambles, Mimeo.

Jones-Lee MW and Loomes GC (1995) Discounting and safety, Oxford Economic Papers 47, 501-512.

Kahneman D and Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2), 263-291.

Kahneman D and Tversky A (1981) The framing of decisions and the psychology of choice, Science 211: 453-458.

Kahneman D, Knetsch JL, Thaler R (1990) Experimental tests of the endowment effect and the Coase theorem, Journal of Political Economy 98, 1325-48.

Kaplan RM, Bush JW, Berry CC (1978) The reliability, stability and generalisability of a health status index, American Statistical Association 704-709.

Kennedy P (1992), A guide to econometrics, Blackwell, p222.

Kind P (1990) Issues in the design and construction of a quality of life measure, in Quality of Life: Perspectives and Policies, Baldwin S, Godfrey C and Propper C.

Knapp RK, Kause RH and Perkins CL (1959) Immediate versus delayed shock in T-maze performance, Journal of Experimental Psychology 58, 357-62.

Lange M, Gerard K, Turnball D and Mooney G (1991) Economic evaluation of mammography screening: information reassurance and anxiety, CHERE Monograph, Department of Community Medicine, University of Sydney.

Laupacis A, Muirhead N, Keown P and Wong C (1992) A Disease-specific questionnaire for assessing quality of life in patients on hemodialysis, Nephron 60, 302-306.

Lipscomb J (1982) Value preferences for health: meaning measurement and use in program evaluation, in Values and Long Term Care, Kane RL and Kane RA, Lexington Books.

Lipscomb J (1989) Time preference for health in cost-effectiveness analysis, Medical Care, 27(3), S233-S253.

Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE and Boyd NF (1982) The measurement of patients' values in medicine, Medical Decision-Making 2, 449-462.

Llewellyn-Thomas H, Sutherland HJ, Thiel EC (1993) Do patients' evaluations of a future health state change when they actually enter that state?, Medical Care 31, 1002-1012.

Lockwood (1988) Quality of life and resource allocation, in Philosophy and Medical Welfare, Bell M and Mendus S, Cambridge Univ Press, Cambridge.

Loewenstein G (1987) Anticipation and the valuation of delayed consumption, The Economics Journal 97, 666-684.

Loewenstein G and Prelec D (1991) Negative time preference, Amercian Economic Review 81, 347-352.

Loomes G (1993) Disparities between health state measures: is there a rational explanation? in, The Economics of Rationality, Gerrard W, Routledge, London 1993.

Loomes G (1994) Valuing health and safety: some economic and psychological issues, mimeo.

Loomes GL and Sugden R (1982) Regret Theory: An Alternative Theory of Rational Choice under Uncertainty, Economic Journal 92.

Loomes G and McKenzie L (1989) The scope and limitations of QALY measures, Social Science and Medicine 28, 299-308.

Martin J and Elliot D (1992) Creating an overall measure of severity of disability for the office of population censuses and disability survey, Journal of the Royal Statistical Society A 155(1), 121-140.

McGuire A, Henderson J and Mooney G (1989) The economics of health care, Routledge.

Mehrez A and Gafni A (1991) Quality-Adjusted Life Years, Utility Theory, and Health-Years Equivalents, Medical Decision Making 11(2).

Meyerowitz BE (1983) Postmastectomy coping strategies and quality of life, Health Psychology 2, 117-132.

Mooney G (1994) Key issues in Health Economics, Harvester Wheatsheaf.

Mooney G and Lange M (1993) Ante-natal screening: what constitutes benefit?, Social Science and Medicine 37, 7, 873-878.

Mooney G and Olsen JA (1991) QALYs: where next?, in Providing health care: the economics of alternative systems of finance and delivery, McGuire A, Fenn P and Mayhew K, Oxford University Press.

Morris J and Durand A (1989) Category rating methods: numerical and verbal scales, Mimeo, Centre for Health Economics, University of York.

Munro S, Ferguson B, Sutcliffe E and Cooper A (1992) St James's University Hospital NHS Trust: health outcomes project, York Health Economics Consortium, University of York, England.

von Neumann J and Morgenstern O (1953) Theory of games and economic behaviour, Wiley, New York.

Ng Y (1983) Welfare economics: introduction and development of basic concepts, MacMillan.

Nord E (1992) Methods for quality adjustment of life years, Social Science and Medicine 34, 559-69.

Nord E (1993) The trade-off between severity of illness and treatment effect in cost-value analysis of health care, Health Policy 24, 227-238.

Nord E (1994) The QALY: A measure of social value rather than individual utility, Health Economics 3 (2), 89-93.

Nord E (1995) Health status index models for use in resource allocation decisions: a crtical review in the light of observed preferences for social policy, National Institute for Public Health Working Paper 6, Oslo,

Nord E (1995) The person trade off approach to valuing health care programmes, Medical Decision Making, 15, 201-208.

Nord E, Richardson J and Macarounos-Kirchmann K (1993) Social evaluation of health care versus personal evaluation of health states: evidence on the validity of four health state scaling instruments using Norwegian and Australian surveys, International Journal of Technology Assessment 9 (4), 463-78.

O'Brien BJ and Drummond MF (1994) Statistical versus quantitative significance in the socioeconomic evaluation of medicines, PharmacoEconomics 5(5), 389-398.

O'Connor AM, Boyd NF and Till JE (1985) Influence of elicitation technique, position order and test-retest error on preferences for alternative cancer drug therapy, Paper to 10th National Nursing Research Conference, University of Toronto.

Ohinmaa A and Sintonen H (1994) The effect of duration on the value given to the EuroQol states, in EuroQol Conference Proceedings, Busschbach J, Erasmus University, Rotterdam, 51-60.

Parsonage M and Neuberger H (1992) Discounting and Health Benefits, Health Economics 1, 71-79.

Patrick DL, Bush JW, Chen MM (1973) Methods for measuring levels of well-being for a health status index, Health Services Research 8, 228-245.

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA (1994) Measuring preferences for health states worse than death, Medical Decision Making 14, 9-18.

Pliskin JS, Shepard DS and Weinstein MC (1979) Utility functions for life years and health status, Operations Research.

Ramsey JB (1969) Tests for specification errors in classical linear least squares regression models, Journal of the Royal Statistical Society B 31:350.

Rawls J (1971) A theory of justice, Harvard University Press, Cambridge.

Read JL, Quinn RJ, Berrick DM, Fineberg HV and Weinstein ML (1984) Preferences for Health Outcomes: Comparison of Assessment Methods, Medical Decision Making 4(3), 315-329.

Redelmeier DA and Heller DN (1993) Time preference in medical decision-making and cost-effectiveness analysis, Medical Decision Making 13, 212-217.

Richardson J (1994) Cost-utility analysis: what should be measured? Social Science and Medicine 39(1), 7-21.

Robinson A, Dolan P, Loomes G and Williams A (1996) Valuing health states using the VAS and the TTO: What lies behind the numbers?, under review by Social Science and Medicine.

Rosser R and Kind P (1978) A scale of valuations of states of illness: is there a social consensus?, International Journal of Epidemiology 7, 347-358.

Sackett DL and Torrance GW (1978) The utility of different health states as perceived by the general public, Journal of Chronic Diseases 31, 697-704.

Schoemaker PJH (1982) The expected utility model: its variants, purposes, evidence and limitations, Journal of Economic Literature 20, 529-563.

Sculpher M, Bryan S, Dwyer N, Hutton J and Stirrat G (1993) An economic evaluation of transcervical endometrial resection versus abdominal hysterectomy for the treatment of menorrhagia, British Journal of Obstetrics and Gynaecology 100, 244-252.

Sen AK (1977) On weights and measures, Econometrica 44, 1539-72.

Sen AK (1982) Choice, welfare and measurement, Basil Blackwell, Oxford.

Shiell A, King M and Briggs A (1993) The consistency of rating scale and time trade-off techniques for eliciting preference weights for health states, HESG Conference, Strathclyde.

Singer PA, Tasch ES, Stocking C, Rubin S, Sieglerer M and Weichselbaum R (1991) Sex or survival: trade-offs between quality and quantity of life, Journal of Clinical Oncology 9(2), 328-334.

Slovic P (1995) The construction of preferences, American Psychologist 50(5), 364-371.

Stevens SS (1971) Issues in psychological measurement, Psychological Review 78, 426-450.

Sugden R (1981) The political economy of public choice, Martin Robertson, Oxford.

Sugden R and Williams A (1978) The principles of cost-benefit analysis, Oxford University Press.

Sutherland HJ, Llewellyn-Thomas H, Boyd NF, Till JE (1982) Attitude toward quality of survival: the concept of maximal endurable time, Medical Decision Making 2, 299-309.

Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Sheperd R, Battista RN and Catchlove BR (1981) Measuring the quality of life in cancer patients: a concise QL-index for use by physicians, Journal of Chronic Diseases 34, 585-597.

Thomas R and Thomson K (1992) Health related quality of life: technical report, Joint Centre for Survey Methods, London.

Torrance GW (1976) Social preferences for health states: an empirical evaluation of three measurement techniques, Socio-economic Planning Sciences 10, 129-136.

Torrance GW (1982) Preferences for health states: a review of measurement methods, Mead Johnson Symposium 20, 37-45.

Torrance GW (1984) Health states worse than death, Paper to Third International Conference on System Science in Health Care, Berlin.

Torrance GW (1986) Measurement of health state utilities for economic appraisal, Journal of Health Economics 5, 1-30.

Torrance GW (1987) Utility approach to measuring health-related quality of life, Journal of Chronic Diseases 40, 593-600.

Torrance GW, Boyle HH and Horwood SP (1982) Application of multi-attribute utility theory to measure social preferences for health states, Operational Research 30, 1043-1069.

Torrance GW, Zhang Y, Feeny D, Furlong W and Barr R (1992) Multi-attribute preference functions for a comprehensive health status classification system, CHEPA Working Paper 92-18, McMaster University, Hamilton, Ontario.

Wagstaff A (1991) QALYs and the equity-efficiency trade-off, Journal of Health Economics 10, 21-41.

Ware JE and Sherbourne CD (1992) The MOS 36-item short form health survey (SF36): concepual framework and item selection, Medical Care 30, 6, 473-483.

Warner KE and Luce BR (1982) Cost-benefit and cost-effectiveness analysis in health care: Principles, Practice and Potential, Health Administration Press, Ann Arber, Michigan.

Weinstein MC (1993) Time preference studies in the health care context, Medical Decision Making 13, 218-9.

Weinstein MC and Stason WB (1977) Foundations of cost-effectiveness analysis for health and medical practices, New England Journal of Medicine 296, 716-721.

Williams (1988) Ethics and efficiency in the provision of health care, in Philosophy and Medical Welfare, Bell M and Mendus S, Cambridge Univ Press, Cambridge.

Williams A (1985) The economics of coronary artery bypass grafting, British Medical Journal 291, 326-329.

Wolfson AD, Sinclair AJ, Bombardier C, McGeer A (1982) Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons, in Values and long term care, Kane RL and Kane RA (eds), Lexicon Books, Mass.