The Valuation of Health Outcomes Data from Clinical Trials for Use in Economic Evaluation

Volume 1

A thesis presented for the degree of PhD at the School of Health and Related Research, the University of Sheffield
by

Isabel Margaret Falcon Towers, BSc (York), MMedSc (Birmingham)

October 2005

Funded by the Medical Research Council

Contributions to the work in this thesis

John Brazier, as first supervisor, provided a great deal of support throughout the design phases of the studies, the analysis, and the subsequent writing of the thesis.

Paul Dolan, as second supervisor, provided support in the design phases of the studies, and made many helpful suggestions for analysis and discussion.

The interviews described in Chapters 5 to 8 were all carried out by the author.

The following individuals were involved with the studies described in Chapters 5, 6, 7 and 8:

Korina Karampela (Glaxo Wellcome) assisted in the design stages of Chapter 5.

Nigel Mathers and Maria Platts (Institute of General Practice and Primary Care, University of Sheffield) helped with the study described in Chapter 6. Nigel Mathers provided suggestions for the design of the project. Maria Platts was very helpful in putting the author in touch with GP practices and helping with recruitment of patients.

Jonathan Michaels and Kathryn Rigby (Vascular Institute, Northern General Hospital, Sheffield) helped with the recruitment of varicose veins patients for the study in Chapter 7. Both were involved in discussion at the design stages of the study. Kathryn Rigby also acted as moderator in the patient focus groups.

Steve Thomas (Vascular Institute, Northern General Hospital, Sheffield) provided support in the design of the aneurysm scenarios described in Chapter 8.

Contents

Summary

This thesis explored the extent of violations of the axioms underlying the Quality-Adjusted Life-Year (QALY), and constructed methods of eliciting health profile valuations holistically. In four studies, health states and profiles were constructed for irritable bowel syndrome (IBS), varicose veins, and abdominal aortic aneurysms (AAAs). Profiles were valued using QALY and holistic methods. Results were analysed to determine the extent of violations of the QALY, to compare QALY and holistic valuations, and to explore the potential of holistic valuation.

This thesis was methodological, and looked at three different approaches to holistic valuation, including a single-stage standard gamble and different approaches to using TTO in valuing profiles.

The results indicated that differences between QALY and holistic valuations depended on the types of profile. Holistic values for IBS profiles tended to be higher than QALY values, but QALY values were found to be more logically consistent and to have higher convergent validity with the original ranking of the profiles. No significant differences existed between QALY and holistic valuations for varicose veins profiles, although the QALY showed greater logical consistency. For AAA profiles, holistic values were consistently lower than QALY values, and the holistic method showed greater convergent validity.

This work highlights practical difficulties in the construction of holistic health profiles. There were indications that respondents used heuristics to aid valuation of profiles. The possibility that responses may not accurately reflect preferences because of cognitive overload is a concern with the holistic approach. However, there is strong evidence that respondents violated the QALY axioms, and therefore it is also of concern that QALY valuations may not accurately reflect preferences. More research is needed into how people value health profiles in order to understand the source of differences between valuation methods, and to examine variations between methods across a wider variety of health conditions.

Chapter 1

Introduction

A commonly used measure of benefit in economic evaluations is the number of quality-adjusted life-years (QALYs) that an intervention generates. The QALY approach combines the value of changes in health-related quality of life (HRQoL) with information on length of life into a single index number. The demonstration of cost-effectiveness in terms of an incremental cost per QALY ratio is becoming increasingly important for the public funding of new health care interventions in some countries including the UK (Commonwealth Department of Health, Housing and Community Service, 1994; National Institute for Clinical Excellence, 1999).

The use of QALYs to value the benefits of treatment requires a number of assumptions about the way people value a profile of health over time (Bleichrodt *et al*, 1997). For example, the number of QALYs associated with a particular health profile, consisting of a given health state, Q, and a given number of years, T, is typically calculated by multiplying the value of Q by T. Thus, the value of a health state is assumed to be a linear function of the time spent in that state. In addition, for health profiles characterized by changes in HRQoL, it is usually assumed that the value attached to a particular health state is independent of the state(s) that precede or follow it. This is equivalent to the estimation of the area under the curve for repeated assessments in a clinical trial (Matthews *et al*, 1990). If these assumptions are violated, then the QALY algorithm may give a false impression of the value associated with different health profiles, and hence the relative effectiveness of different interventions might then be misrepresented.

At this point it is worth defining the terminology that will be used throughout this thesis. The term "profile" will be used to describe health states occurring over time. The term "scenario" will be used to incorporate risk in addition to time.

Despite recent interest in the issue of whether summing QALY scores derived from the constituent health states of a profile gives an accurate value for that profile, there has been relatively little empirical research in this area and the existing evidence is limited in terms of the economic evaluation of health interventions. Studies in this area have focused largely on the sequence effects of the constituent health states. For example, Richardson *et al* (1996) found that profiles that deteriorate over time are valued lower than those that vary little over time. Lipscomb (1989) found that profiles which improve with time and finish on a relatively highly valued state, are valued higher than those which deteriorate or show little variation. This suggests that the sequences of states has a systematic effect on the value of a profile. However, a study by Mackiegan *et al* (1999) showed that, where there is very gradual deterioration, sequence has no significant effect on profile valuations.

There has been some relevant research in the psychological literature that has looked at how people's preferences over different profiles are affected by the sequencing of the events within those profiles (Varey and Kahneman, 1992; Redelmeier *et al*, 1993; Redelmeier and Kahneman, 1996; Ariely and Carmon, 2000). These studies were not examining specific health interventions, but consisted of profiles containing different forms of discomfort or pain. The findings of these studies indicated that people prefer profiles that improve over time and end on a good note to profiles with the reverse pattern. However, preferences for improving sequences are moderated by prior expectations of what is realistic (Chapman, 2000). A more general explanation is that certain aspects of the profiles may be given particular weight by the valuer. In particular, the final point, the peak point, and the rate of change in measurements over a profile of pain have been found to be significant predictors of the overall rating of the profile (Ariely and Zauberman, 2000).

The additive assumption of the QALY model leads to the assumption that temporary states within a profile will be given little weight relative to the rest of the profile. Utility attached to process of treatment may therefore be assumed to have little effect upon the valuation of a health profile. There has been little research to determine whether this assumption holds. This is an important area for research, because a high level of disutility attached to a particular process of treatment may alter the results of a cost utility analysis of different interventions.

The QALY model is based on Expected Utility Theory (EUT), and a general assumption of EUT is that people will show linearity over probability. Thus utility over probabilities is assumed to lie on an interval scale from 0 to 1, and the difference in utility between probabilities of 0 and 0.1 should be considered equal to the difference in utility between probabilities of 0.5 and 0.6. This has been refuted by the work of Allais (1953, 1979), who conducted a series of experiments with lotteries concerning money, and showed that people are more loathe to move from no probability of loss to even a small probability of loss than an equal interval move in the middle of the probability scale. The assumption of linearity over probability has been shown to be violated in a health economics context (Cook *et al*, 1994).

In general, the psychological literature has shown that each part of the profile is not of equal value. Despite previous research into the effects of duration and sequences, there has been very little research in the context of the QALY assumptions of linearity over time and probability.

1.1 Aims of this thesis

The aim of this thesis is to examine the practical implications for health services research of applying holistic profile methodologies for valuing the outcomes of health care to data generated from clinical trials. These have been recognised as important developments in health economics but there has been remarkably little empirical work on deriving holistic profiles, or on the extent to which they yield different results to the more convenient QALY method. The methodological and practical issues surrounding holistic profile valuation provide the topic of this thesis.

The aims of this thesis are to:

- Review the theory underlying the QALY and holistic valuation methods.
- Review the empirical literature on violations of the assumptions underlying the QALY method.
- Review the empirical literature on holistic valuations.
- Explore the practical issues around the construction of health profiles for each of the disease conditions of choice, and to construct these profiles.
- Explore the methodology of valuing these profiles with holistic methods.
- Determine whether valuations obtained by holistic methods differ from those obtained from the traditional QALY method.
- Explore the reasons for any differences in the results obtained by the different valuation methods.
- Examine the extent to which people violate the axioms of the QALY algorithm.

- Consider which valuation method might provide a more valid reflection of individual patient preferences.

1.2 Structure of the thesis

Chapter 2 outlines the theory underlying QALYs and healthy years equivalents (HYEs). The origin in welfare economics is discussed. Compensation analysis techniques are described, and the growth of cost utility analysis using the QALY tool are described. The bases of QALYs and holistic valuation methods in welfarism and non-welfarism are discussed. Criticisms of the assumptions of the QALY model and the HYE alternative are reviewed.

Chapter 3 reviews the empirical literature to determine the extent to which previous research has found the assumptions underlying the QALY algorithm to be upheld. The review covers mainly the health economics literature, but also considers psychological research. Following this review is a review of the empirical research into holistic valuations of health profiles.

Chapter 4 sets out the overall questions addressed by the four empirical studies and the research methods used in the studies.

Chapters 5 and 6 describe two studies of valuation methods for health states and profiles relating to irritable bowel syndrome (IBS). As explained in Chapter 5, IBS is an interesting area for study of economic evaluation, because treatments tend to be aimed at reducing the proportion of time spent suffering from symptoms rather than to eliminate the symptoms altogether. The construction of the profiles required detailed consideration, because the IBS health states often do not follow clear sequences, but vary in frequency. This condition provides an exciting forum in which to study the assumptions of additivity over time of the QALY algorithm, *i.e.* to determine whether the utilities for different health states are indeed separable and independent when it comes to IBS.

In Chapter 7 the effect of treatment process on valuations is explored. Treatments for varicose veins are of relatively short duration, ranging from a couple of days to several weeks for the affects of treatment to wear off. This raises the issue of how a process lasting such short durations should be valued. This issue is discussed at more length in Chapter 7. The issue of *ex ante* risk is also looked at, with respect both to treatment processes and risks of recurrence.

Chapter 8 seeks to develop profiles relating to large abdominal aortic aneurysms in unfit patients. Respondents are asked to consider the balance between their wish to maximise life expectancy, and risks of severe morbidity or mortality in the short-term. This study explores attitudes to risk, and tests the QALY axiom of a constant attitude to risk over survival duration. An attempt is made to measure time preferences, and to test the axiom of zero time preference. Health profile valuations obtained by the QALY method are adjusted for the risk attitudes and time preferences of the sample.

The results of the studies in Chapters 5 to 8 are discussed in Chapter 9. The implications of the findings are discussed in terms of construction of holistic health profiles and comparisons of the values obtained by QALY and holistic methods. The implications for health economics and economics more generally are also discussed. A programme of further research in this area is proposed.

The overall conclusions to the thesis are set out in Chapter 10.

Chapter 2

A Review of the Theory Underlying QALYs and HYEs

2.1 Overview of this chapter

This chapter provides a brief overview of the development of welfare economics and methods of economic evaluation in health care. Section 2.2 outlines the history of welfare economics, and discusses the two schools of thought of welfarism and extra-welfarism (or non-welfarism). Section 2.3 discusses the reasons health care is different to other market goods. The different methods of conducting economic evaluations are described in Section 2.4. The willingness to pay technique is described in Section 2.5. The development of QALYs is discussed in Section 2.6, along with valuation techniques. The development of the healthy years equivalent (HYE) is discussed in Section 7. The welfare economics foundations of QALYs and HYEs are discussed in Section 2.7, and Section 2.8 concludes.

2.2 Welfare economics

Welfare economics is the branch of economics that is concerned not only with the efficiency issues of resource allocation, but also with the normative concerns of equity or fairness. Thus it is concerned with how things *should be* rather than merely how they are.

A key concept in welfare economics is the notion of utility. The concept of utility was developed in the 18th and 19th century, but has evolved over the years, as reported in Richardson (1994). There are at least four definitions of utility. Firstly, utility was seen as a psychological concept of well-being or welfare, which was originally based on the ideas of Bentham (1789), who believed that utility could be derived from either pleasure or pain, and that a level of utility was absolute regardless of the source. This view was adopted by many economists in the 19th century (Viner, 1925). Secondly, utility was seen as an ordinal ranking system (Hicks and Allen, 1934). Thus the term "utility" was used to define the order of preferences between different options. As Graaff (1967) points out, this concept of utility is limited, and merely states that an individual would choose option A over option B if able. However, it is measurable in terms of revealed preferences (Richardson, 1994). Thirdly, utility was seen as having measurable cardinal properties (Allais, 1984), with a focus on the intensity of preferences. While being similar to the first definition in that it is basically a psychological definition, it is more of a measurement of preferences than of well-being. Thus individuals might have preferences for goods or activities that reduce their overall well-being or welfare (Richardson, 1994). Again, this concept of utility is measurable in terms of revealed preferences. Fourthly, "von Neumann-Morgenstern" utility is defined in connection with the standard gamble, in which an individual states the level of probability at which they are indifferent between the certainty of receiving a good and a probability of a better good and a worse good (these latter are often set at full health and death in health economics) (Richardson, 1994).

Economists generally agree that it is impossible to measure utility directly, and there are inherent problems with attempting to make interpersonal comparisons of utility (Friedman, 1984; Bleichrodt, 1997). However, efforts have been devoted to developing techniques of measuring reflectors of utility. One of the oldest of such methods is the willingness to pay concept, which was first suggested in the 19th century by Dupuit (1844), who suggested that a measure of the utility of an object should be taken as the maximum amount which each consumer would be willing to pay in order to acquire the object. The concept of willingness to pay was also famously taken up by Marshall (1890). The willingness to pay method will be discussed further below, and again in Section 2.5.

2.2.1 Welfarism

This sub-section describes the welfarist school of thought within welfare economics.

According to Sen (1979a), the welfarist view is that "the judgement of the relative goodness of alternative states of affairs must be based exclusively on, and taken as an increasing function of, the respective collections of individual utilities in these states". Sen (1979b) defines three basic attributes of welfarism. These are:

> Consequentialism – The outcome is what matters rather than the act that causes the outcome. Thus, utility should be maximised regardless of its source.
>
> Ordinalism – Social welfare judgements are to be based on ordinal ranking of individual utilities.
>
> Non-comparable utilities – Social ranking is independent of intra-personal comparisons of utility.

Welfare economics incorporates the concept of Pareto-efficiency. This is based on the ideas of Pareto (1909), who suggested that the market was inefficient if there was any point that could be reached, at which one or more individuals would be better off and nobody would be worse off. However, within a society there are an infinite number of Pareto-efficient distributions. The choice of which of these distributions is the "best" is a value judgment. Welfarists depart from Pareto in that they believe that social welfare is simply the sum of all individual utilities within a society (Dolan, 1998a), and a higher level of social welfare might be achieved by increasing the utility of some members of society while decreasing the utility of other members, if the overall utility increased.

Compensation tests were developed by Kaldor (1939) and Hicks (1939, 1941, 1943) as an extension to the Paretian criterion for the purpose of making judgments about the redistributions of goods or services. The principle behind this type of test is that, for any redistribution that involves a gain for some and a loss for others, the redistribution should go ahead if the maximum that the gainers are willing to pay is greater than the minimum that the losers are willing to accept.

Compensation variation (CV) begins from the standpoint of the original utility level. A redistribution may be proposed which would lead to a change in the level of utility. People for whom this would result in a gain would be asked for their maximum willingness to pay (WTP) for such a gain, and this would result in a restoration to their original level of utility. People for whom this redistribution would result in a loss would be asked for the minimum willingess to accept (WTA) value for this loss, and this would again restore the original utility level. In equivalence variation (EV), the perspective is from the new level of utility. Thus, a redistribution may be proposed which would again result in gains for some and losses for others. Individuals for whom it would be a gain would be asked for their minimum WTA value to compensate for not receiving this gain. Thus, their utility level would be increased to a level equivalent to that which would have been gained by the proposed redistribution. Individuals who would suffer a loss from the change would be asked for their maximum WTP for the change not to go ahead. These tests provide indications of the increases or decreases in utility attached to potential gains or losses. Thus, if the WTP for gains are greater than the WTA for losses, this is said to be an indication that the redistribution of goods or services provides a net increase in utility. The payments need not actually occur.

The welfarist will first seek the most efficient allocation of resources, based on welfarist premises. In welfarist terms, utility is reflected by factors that increase the individual's capacity to benefit from a unit of a good. These factors may be market productivity, or the ability to enjoy leisure activities. Utility may therefore be reflected by WTP. A rich person may obtain greater benefit from a unit of health than an impoverished person, because they are in a better position to enjoy good health (Sen, 1979a,b). According to welfarism, it is therefore more efficient to allocate health to a rich person than an impoverished person. However, having obtained this efficiency information, the welfarist may decide to balance efficiency with equity considerations by using some form of equity weight on the aggregated data.

Equity deals with how fairly resources are distributed. Horizontal equity deals with equal treatment of equals in terms of economics, and is therefore usually thought to rule out discrimination on grounds of racism, sexism or other "isms". Vertical equity deals with the unequal treatment of unequals, thus allowing the concept of redistributing resources according to different economic characteristics and need (Begg *et al*, 1994). Taxing the high-income groups at a higher rate to provide a benefit or welfare service for the less well off is an example of vertical equity. Pareto-efficiency provides no guidance about choices over redistributions of goods based on equity, such that some members of society would gain while others would lose out. Such a choice is purely a value judgment. However there is evidence that people are not only concerned with the amount of utility, but also the distribution of that utility (Nord *et al*, 1995). It is possible that people may prefer less utility to be obtained if it is distributed "fairly" than for utility to be maximised for a minority.

In summary, the broad welfarist belief is that utility is ordinal and cannot be compared between individuals (Gowdy, 2004). Welfarists take the stance that the measure of social welfare is equal to the sum of the utilities of all the individuals in society (Bleichrodt, 1997). According to the Kaldor-Hicks welfarist viewpoint, upon which CV and EV are based, any increase in total social welfare is desirable, regardless of its source. The welfarist decision-maker may decide to use equity weights or value judgements. For example, the welfarist may not wish to base resource allocation strictly along the lines of income-based measures, and may therefore give different weights to each income group and base resource allocation on efficiency weighted by equity (Sen, 1979a,b).

2.2.2 The non-welfarist viewpoint

As already stated, welfarism does not take into consideration the methods used in the increase in social welfare, or the source of utility. Sen (1979b) makes the point that an increase in total social welfare could be attained by one person torturing another if the increase in utility for the culprit was greater than the decrease in utility for the victim. Sen makes the point that certain non-welfare aspects may matter, such as justice. In other words, justice in itself may be so important that it cannot just be assumed to fit implicity into the utility functions of individuals, but has to stand separately.

Following on from these ideas is the implication that there may be more to social welfare than simply the sum of the utilities of all the individuals in society. Such things as justice and equity are seen by many to be important, and attempts have been made to model social welfare functions based on such precepts as equity (Bleichrodt, 1997). For more on this see Dolan (1998a, 1999) and Johannesson (1999).

Sen (1979a) explored the effects of weakening the three welfarist rules listed in the previous sub-section. He argued that, even if utilities could be non-ordinal (*e.g.* cardinal, or ratio), and interpersonal comparisons of utility could be made, there is often still a concern for the source of utility. For example, Sen argues that A's utility obtained from A's enjoyment of B suffering is a different kind of utility to that which A obtains from reading a book he enjoys. In the first instance, A's utility status is affected by the utility status of B. In the second instance, A's utility status is totally connected to A's own activites. Sen argues that these are two different kinds of utility, and that many people feel that the outcome is not the only thing that matters. Thus, people may weight utility according to its source.

The above concept is not based in welfarism, and in health economics is commonly referred to as "extra-welfarism". However, non-welfarist views do not necessarily add anything "extra" to welfarism, and the more correct term is "non-welfarism". In this chapter it will therefore be referred to as non-welfarism.

It would be desirable to begin this sub-section by defining non-welfarism. However, the concept of non-welfarism is difficult to define, because there are many views of what is meant by non-welfarism. Some work in the field of non-welfarism has involved merely rejecting one or more welfarist concept (Sen, 1979a,b). However, in health economics a feature of non-welfarism is often the introduction of varying amounts of decision-maker freedom, because the welfarist restrictions are not necessarily applied (Culyer and Wagstaff, 1993a).

Sen (1980, 1993) further developed the theory of non-welfarism by exploring the issues of basic capabilities and well-being (see also Sugden, 1993). According to Sen, people have a range of functionings. These functionings describe well-being, and include being adequately nourished, avoiding premature mortality, and being happy. Thus, Sen saw individual utility as part of well-being, rather than well-being in its entirety. The basic capabilities of an individual describe the capability of that individual to achieve well-being as defined by his set of functionings. Sen suggests that it is the fact that a paraplegic is unable to perform certain activities that gives her the classification of having "special needs". Thus it would seem that needs, according to Sen's notion of basic capabilities, might be defined differently in different cultures. If everyone was and always had been a paraplegic, no one would feel deprived of their basic capabilities.

Culyer (1991) expands on Sen's non-welfarism to suggest that a paraplegic individual, who has adapted to his condition and feels that he has a high level of utility, may none the less have *needs* as a result of being paraplegic over and above what is in his utility function. The implication is that society owes this individual a duty of service to needs over and above that which the individual may feel he needs or desires, and therefore the individual's utility function is not a sufficient deciding factor. According to Culyer (1989b), "the idea of utility focuses too much on mental and emotional responses to commodities and characteristics of commodities and not enough on what they enable you to do".

Following from the ideas of Culyer is the non-welfarist notion that the individual is not responsible for the decision-making process, but the decision of what that individual needs should be in the hands of an "expert" decision-maker. This notion is supported by Brouwer and Koopmanschap (2000), who suggest that "it seems important to assess assumptions used in welfare economic models in terms of their ability to reflect the values and judgements endorsed by society or policy makers". This goes back to the non-welfarist notion that there may be a social welfare function which is different from simply the sum of individual welfare functions (see above).

A different non-welfarist view is the communitarian view proposed by Mooney and Jan (1997) and Mooney (1998a, 1998b), according to which the community should decide what is important for the allocation of health care resources. This might, for example, be based on the degree of individual responsibility for ill health. For example. the community may be of the overall opinion that it is the fault of the smoker that he developed lung cancer, and therefore this smoker is less deserving of resources than a non-smoker. However, it would sometimes appear that, as more medical research is conducted, the more it seems that people are responsible for their own ill health. For example, using the same argument it could be argued that people who do not eat the recommended daily amount of fruit and vegetables do not deserve health care resources to be allocated to their treatment after development of colorectal cancer. This view would also support a lesser allocation of resources to the health care of car accident victims if these victims drive over the legal speed limit and have therefore put themselves at risk. There are many other such examples, and it may end up being the case that there are relatively few groups who are not responsible for their ill health, and therefore there may be a relatively small proportion of people who "deserve" to have resources allocated to their health care. However, the non-welfarist decision-maker could weight different groups according to the degree to which they are "at fault".

The non-welfarist communitarian notions of Mooney (1998a, 1998b) could also have an impact on the distribution of health care resources among different community groups. Thus majority groups within the community may have a greater power in the decision-making process than minority groups within the community. This could lead to serious reductions in well-being within some groups. Mooney (1998a) himself suggests that this could, in theory, have been the case in Nazi Germany leading up to World War II.

In health economics, broadly speaking, the non-welfarist is concerned with the source of utility, and therefore does not agree with the welfarist view that it may be more efficient to allocate health to the rich than the poor (Brouwer and Koopmanschap, 2000). The non-welfarist therefore may choose to ignore most of the utility function, just choosing to focus on that part of the utility function relating to health (although, due to the decision freedom within non-welfarism, the non-welfarist could take into account whatever aspects of the utility function she chooses). The non-welfarist prefers to see health as equal for all, rather than acknowledging that any given health state could have a different value depending on the income group of the recipient (Dolan, 2000). This decision is based on concepts of justice and equity. Having obtained data on values of different health states, the non-welfarist may then weight the results according to non-health concerns such as equity.

One non-welfarist view, in a health economics context, is that every individual's health is seen to matter because good health above all other things is seen as "necessary for an individual to 'flourish' as a human being", and "in so far as health care is necessary to 'good health', this provides a strong ethical justification for being concerned with the distribution of health care and not with the distribution of, say automobile spares, and for using the word 'need' in the context of health care and not in the context of, say, skiing holidays" (Culyer and Wagstaff, 1993a). Thus, according to this concept of non-welfarism, health economists should focus on valuing health rather than measuring utility, which incorporates things other than health. The non-welfarist in health economics typically only considers the output of health from the use of health care. Rather than arguing for equal access to health care, the non-welfarist may argue for equal health between individuals (Dolan, 2000), or equal changes in health between individuals.

Health care stands out as different from other market goods, because of the externalities relating to health care or health, which cause individuals to benefit from the provision of health care to others in need of it. Culyer and Simpson (1980) put forward the argument that it is the health status of other individuals, rather than their access or not to health care, which impacts on each individual's utility function. Evans and Wolfson (1980) described individuals as "wishing [each other] well, but not necessarily happy". In a publicly-funded healthcare system such as the UK NHS, the tax-payers may prefer to take the non-welfarist approach, and their caring externalities may only go so far as to maximise the health status of others rather than to maximise the utility function of others. Whereas an individual may wish to maximise her own utility function, taking into account all aspects of it, she may wish only to maximise the health status of others across society.

Welfarists and non-welfarists may also differ in their opinions about valuing health states. Because of the welfarist belief that the whole utility function should be taken into consideration rather than just the health argument, welfarists suggest that the same health state might be given different values between different people (Dolan, 2000). Indeed, it has been shown that people will often adapt to their ill health state (Meyerowitz, 1983, Cassileth *et al*, 1984). As Dolan (2000) points out, this could lead to the decision to allocate resources to a less seriously ill person who has not adapted so well, rather than the apparently more seriously ill person who has adapted and gives a higher value to their health state (see also Sen, 1987). One non-welfarist view is that health is the only relevant outcome, and therefore suggests that each health state should be given the same value whoever is in it.

The welfarist approach allows the utility derived from treatment process to be taken into consideration in the valuations of the intervention, whereas the non-welfare approach traditionally merely focuses on health status outcomes (although it would be possible for the non-welfarist to give weight to other parts of the utility function, such as those aspects relating to dignity, treatment process, or whatever else the policy maker felt to be of importance). Preferences over process may be directly related to the treatment itself, such as preferences between two different procedures. Depending on the remit of the non-welfarist decision-maker, she may expect valuations of the procedures to be based only on health status outcomes, such as the level and duration of pain involved. The welfarist would allow values to incorporate any aspects of the utility function affected by the treatment, not restricting valuations to health outcomes. This welfarist definition means that utility derived from a whole range of attributes relating to treatment process may be taken into account, including the "hotel" care received during in-patient hospital stays. Even propounders of the non-welfarist view admit that "it seems unlikely that any extra-welfarist would assign zero weights to such factors as consumer choice, privacy, speed of service, hospital hotel service, and other factors that may be only remotely causally linked to health" (Culyer, 1991). In fact, the breadth of the ranges of sources of utility to be taken into account by the non-welfarist might be as broad as the welfarist, strictly depending on the view of the non-welfarist decision-maker as to what is important. Birch and Donaldson (2003) express the concern that non-welfarism allows decisions to be based not on the preferences of individuals, but upon the preferences of one person or a relatively small group of people involved in making policy decisions.

In summary, there are several non-welfarist views, and there is not one overall consensus as to the meaning of non-welfarism either within the broader economics literature, or within the health economics literature. However, broadly speaking, the non-welfarist approach appears to be more paternalistic and less trusting of the individual's ability to look after their own best interests (Birch and Donaldson, 2003). Dolan (2000) is correct when he says that choosing between these views is a normative decision.

2.3 Health care is different

Most studies of markets are based on the neo-classical economical theories of supply and demand. Smith (1776) likened the market mechanism to an "invisible hand", which dictates prices and thus causes an equilibrium price to be reached for each good in each market. Neo-classical market theory requires the following assumptions:

- Utility maximisation - Consumers act to maximise their utility, and suppliers act to maximise their profits (Donaldson and Gerard, 1993)
- Certainty – Consumers know what they want, where they can obtain it, and when they will want it (Donaldson and Gerard, 1993), and they should be able to determine the effect a good will have on their utility
- Perfect knowledge of the market (Begg *et al*, 1994)
- No supplier-induced demand – Consumers should be able to act free of self-interested influences of suppliers (Donaldson and Gerard, 1993)
- No externalities – Consumer utility should not incorporate the effects of transactions in which he did not voluntarily participate (Donaldson and Gerard, 1993)
- Completeness – Consumers should be able to state ordinal preferences over each good (Gravelle and Rees, 1983)
- Transitivity – If X is preferred to Y, and Y is preferred to Z, then X is preferred to Z (Gravelle and Rees, 1983)

Before entering into discussion about the application of welfare economics methods to health care, it is necessary to explain that the market for health care is different from markets for other goods. Unlike most markets, the full price of health care is rarely paid for at the point of consumption by consumers in the developed world. In the USA, for example, members of the population are expected to take out health insurance, and pay premiums. In the UK, health care is largely funded by the National Health Service (NHS). The arguments for government intervention are set out in the following paragraphs.

In conventional markets, consumers are expected to have a high level of certainty over their demand for market goods, in that they should know what they want and when they will want it. For example, a consumer who shops at the supermarket on a weekly basis will learn as the weeks go by what items he will need to purchase. The shopping list will probably be quite similar on a week-to-week basis. Based on this regularity, or certainty, the consumer will be able to plan a household budget. However, the nature of demand for health care is uncertain, reflecting the uncertain nature of health due to illness.

Consumers in a perfect market should have perfect knowledge of the market. In most markets they are able to gain this knowledge from frequent use, learning from experience. However, in the health care market use is often infrequent. The exception to this is in cases of chronic conditions, in which patients use the health care market more frequently and therefore have the opportunity to become very knowledgeable about their condition (although some sufferers of chronic conditions may not wish to become knowledgeable about their condition). However, even for chronic conditions, medical professionals often learn about new technologies before patients do. There is an association between health and health care, which is impossible for the majority of consumers to fully understand. The consumer often cannot tell what effect a unit of health care will have on his health status. This is one of the reasons the consumer will visit the doctor.

There must be independence between supply of, and demand for, a good. However, there is an asymmetry of information between doctors and patients, which means that the patient has to rely upon the expertise of the doctor during the decision-making process. Thus the doctor acts as the patient's "agent". An agency relationship is one in which the consumer (principal) does not have access to sufficient information in order to make an efficient decision regarding demand, and therefore approaches the supplier (agent) for advice (Donaldson and Gerard, 1993). If the doctor is a perfect agent he will provide the patient with all the information the patient needs in order for the patient to make a decision. However, the reality is often that the patient provides the doctor with all the information that the doctor needs in order for the doctor to make a decision upon the patient's course of action and demand for health care (Williams, 1988). Agency relationships assume independent utility functions, such that the agent and consumer each try to maximize their own utilities. Because of their greater knowledge, doctors are in a position by which they could potentially act in their own interests rather than those of the patient, should there be a conflict of interests. The patient may well be unaware of it if this occurs. In non-publicly financed systems where the doctor is paid for services, this could lead to a situation of supplier-induced demand, in which the supplier of health care would also be the demander of health care (Evans, 1974; Rice, 1983; Cromwell and Mitchell, 1986). In the UK, doctors may not generally accrue direct in-pocket benefits from inducing demand, but there may be advantages to them in inducing the demand of health care beyond that which is necessary if, for example, it helps meet government targets.

The operation of a perfect free market assumes that there are no externalities. Externalities occur when consumption of goods by one individual affect the utility of another, and this is beyond the control of the latter. For example, if a vaccination is taken up by the majority of the population this can provide a positive externality on an individual's health because the disease is less common and therefore less likely to be caught by any individual. This would be coined a "selfish externality" (Donaldson and Gerard, 1993). A caring externality is demonstrated by someone who gains benefit from the knowledge that another individual is being looked after by the health care system, even though this does not affect anyone else's health (Donaldson and Gerard, 1993). Donaldson and Gerard (1993) point to ozone depletion as a negative externality. A negative externality occurs when the consumption of goods by one individual has a negative effect on the utility of another. However, if externalities are not taken into account by market forces, the good that results in a negative externality will continue being produced, and it will be over-produced. There are indications that externalities occur over health care (Culyer, 1971).

In order for a market to operate there should be many small producers, such that no individual producer can exert an undue amount of power on the market. However, doctors must be registered with the General Medical Council to practice in the UK (GMC, 2003). Thus present doctors have some control over who can practice. This licensure exists to maintain standards of medical practice, but results in fewer producers and increased prices.

Uncertainty regarding their demand for health care leads individuals to seek health insurance (this may be replaced by taxation in publicly-financed systems such as in the UK). This means that they pay regular premiums rather than paying for health care at the point of consumption. With a price of zero at consumption, the consumer would have no motivation to stop consuming health care. This overuse is known as consumer moral hazard (Pauly, 1968, 1983). The marginal cost of producing the health care would then become greater than the marginal benefit, meaning that resources spent on health care would be of greater benefit spent elsewhere. There is also provider moral hazard, which occurs if health care providers induce as much demand from an episode as possible in order to increase their income (see the above discussion of supplier-induced demand). There are no incentives for the consumer to prevent this from happening as there would be if the consumer was paying at the point of consumption (Donaldson and Gerard, 1993; Mooney, 1994).

Adverse selection results from the asymmetry of information between insurers and the insured. The insured party has the informational advantage, because he has a better knowledge of his own risk status. One way to get around the difficulty of ascertaining the true risk status for each individual in the population is for insurance companies to set a community-rated premium. The community-rated premium is an estimation of the average risk for the insured population. The problem with this is that the community-rated premium is higher than the fair premium for low-risk individuals, and lower than the fair premium for high-risk people. This will lead to low-risk individuals dropping out of the insurance scheme. The average risk level of the insured population will then increase, resulting in a higher community-rated premium. More low-risk people will drop out, and this will go on. This will ultimately lead to two groups of uninsured:

1) Low-risk people who would pay a reasonable premium for their level of risk, but feel that the premiums asked are too high.

2) High-risk people who cannot afford to pay the required premiums even if the premiums were to fit the magnitude of their risk status.

This leads to market failure for (1), but not for (2) which is usually considered inequitable. In the latter instance they simply cannot afford the prices, like they cannot afford to buy a Mercedes (Evans, 1984). However, for (1) this low-risk group cannot obtain insurance at reasonable prices.

Another problem with private insurance schemes is that, if there are many small companies operating, there will be diseconomies of small scale. Each company will have its own administrative costs for handling insurance schemes for each consumer. This could lead to market failure because consumers may be unwilling to pay premiums that incorporate administrative costs, and are therefore greater in price than the true value of the premium. Indeed, there is a great deal of evidence to suggest that administrative costs are proportionately higher when health care is financed by numerous competitive small-scale companies (Evans. 1984: Himmelstein and Woolhandler, 1986; Evans, 1987; Richardson, 1987; Quam, 1989: Evans. 1990; Woolhandler and Himmelstein, 1991). However, if there was a monopoly insurance company, this would solve the problem of diseconomies of small scale at the expense of allowing the opportunity for exploitative pricing.

Thus there are several reasons for government intervention in the funding of health care services. One reason is the market failure of private funding in terms of adverse selection. Another reason is market failure due to moral hazard. A third reason is due to diseconomies of small scale, as a taxed system provides economies of scale while avoiding the problem of exploitation. A fourth reason is caring externalities. Government intervention by taxation would cover the costs of transfer of benefits to the disadvantaged. The necessity for licensure of doctors and asymmetry of information would also lead to market failure.

In fact, as Donaldson and Gerard (1993) point out, in many cases the public health care funding systems were set in place before these arguments were clearly defined (Culyer, 1971; Evans, 1984). However, they may have underlain the foundations of these systems in less well-defined forms. The formation of systems such as the NHS in the UK were correlated with strong labour movements and socialist governments (Navarro, 1989). Equity and social justice were issues at the heart of the formation of the NHS. Initially it was said by the UK government that "... everybody in the country ... should have an equal opportunity to benefit from ... medical and allied services" (HMSO, 1944). It was later said that "... equal opportunity of access to health care for people at equal risk" (HMSO, 1976). The characteristics of health care have led to large scale public funding, with the result that the government has to ensure efficiency in resource allocation. This has led to an interest in economic evaluation in health care.

2.4 Methods of economic evaluation

Economic evaluation in health care exists on the basis that governments, hospitals, and other bodies wish to know whether different programmes are efficient, cost effective, or how they compare with other programmes. For example, the National Health Service in the UK may wish to consider the question of whether the Breast Screening Programme should be extended to women between the ages of 40 and 50 years. In order for policy makers to make this decision, they need information about how much it would cost to routinely invite women in this age group and the benefits of doing so compared to the next best alternative programme.

When considering the cost of inviting each woman for screening, relevant costs include screening staff time and use of equipment. But there are also costs to the women. These may include travel expenses, childcare facilities, or time away from paid work. There may also be emotional costs accruing for example from false positive results. The benefits accrued from inviting these women would hopefully include a reduction in mortality from breast cancer over time and the reassurance to women screened. The costs of not inviting these women would be the benefits forgone. The policy maker needs to decide from whose perspective costs and benefits will be considered.

In order to determine the opportunity costs of this programme, information would be required on other alternative uses of the resources. An economic evaluation must compare two or more alternatives and consider both costs and benefits in a systematic manner (Drummond *et al*, 1997). Some studies only consider either costs or benefits between two programmes. These are cost or outcome descriptions. Other studies consider costs and benefits, but only within one programme. Again, these studies are not an economic evaluation, but a cost-outcome description. Some studies compare costs between different programmes. Since benefits or outcomes are not compared, these are cost analyses. When the benefits of two programmes are compared, but the costs are not, this is called an effectiveness or efficacy evaluation (Drummond *et al*, 1997).

There are four basic types of economic evaluation. These are known as cost minimisation analysis (CMA), cost effectiveness analysis (CEA), cost benefit analysis (CBA), and cost utility analysis (CUA).

A CMA compares the costs between two alternative programmes when the differences between the outcomes are identical, or at least do not differ significantly (Drummond *et al*, 1997). A CEA compares alternatives in terms of both costs and outcomes. This involves determining either the cost per unit of the relevant outcome measured in natural units, or the amount of the relevant outcome per unit cost. The outcome used varies with the type of study. Thus outcomes may be medical, such as measurements of blood sugar levels for drugs to treat diabetes. Non-medical outcomes include measurement of life years gained. However, a problem with CEAs is that they focus on only one measure of outcome. A drug to treat diabetes may have effects other than lowering blood sugar levels, such as significant side effects. A further limitation of CEAs is that two alternative programmes can only be compared in terms of a common outcome. For example, two programmes that result in number of lives saved could be compared. However, if the outcomes from two programmes were quite different (*e.g.* prevention of premature deaths and prevention of disabled days), a comparison in terms of CEA would be difficult (Drummond *et al*, 1997).

In CBAs the outcomes of programmes are valued using a common denominator, *i.e.* units of money (Drummond *et al*, 1997). Thus monetary values are placed on, for example, the prevention of premature deaths or disabled days, and then the costs and outcomes of alternative programmes are comparable. WTP is a frequently-used valuation method in CBAs, since it allows researchers to place a monetary value of the potential benefits of a programme, which can then be compared to the costs. This technique has the closest link to welfare economics.

Rather than putting a monetary value of the outcomes of health care, which may incorporate things other than effects on health, CUA measures outcomes in terms of values based on length of life and health-related quality of life (HRQoL). A common unit used to describe outcomes of health care interventions is the quality-adjusted life-year (QALY). Alternative programmes are compared in terms of cost per QALY. CUAs have the advantage over CEAs in that all the health affects of a programme can be incorporated into this single index measurement. The focus of this thesis will be on methods used to measure value in CUAs.

2.5 Willingness to pay and accept

The usual welfare economics approach to valuing goods and services would be to use market data on WTP or WTA. However, health care is rarely paid for at the point of consumption. It is usually covered either by private insurance or by government funded health care programmes. This means that it is practically impossible to gauge revealed preferences using actual WTP or WTA for health care. Studies have attempted to gauge revealed preferences over life years, assessing the relationship between levels of risk individuals are willing to take in their employment and their wages (*e.g.* Marin and Psacharopoulos, 1982; Viscusi, 1992). The problem with this approach is that, for the evaluation of health care interventions, it would be necessary to study revealed preferences for the particular health outcome of interest. Particular modes of employment tend to be fairly risk-specific. and the revealed preferences may be employment-specific.

Another method of use for compensation tests is the stated preference method. also known as contingent valuation. In this method individuals are presented with a hypothetical scenario, and WTP or WTA methods are used to measure the value placed on the scenario. Mishan (1971) advocated the use of the Kaldor-Hicks compensation test in connection with health services. These methods were used initially in transport studies (Jones-Lee, 1976). Jones-Lee suggested the necessity for testing under uncertainty rather than certainty in the context of valuing life, on the basis that under certainty the values given, for example, to loss of life might be infinite. Jones-Lee also argued for valuations being made in terms of monetary value for life to risk trade-offs.

Most CBAs in health economics rely on WTP and do not include WTA. One of the reasons for this is that the two types of measure have been found to give different results for the same question, namely that WTA values are always greater than WTP values. This has been shown in studies of hypothetical lottery choices (Knetsch and Sinden, 1984 and 1987; Coursey *et al*, 1987; Singh, 1991), environmental economics (Gordon and Knetsch, 1979; Rowe *et al*, 1980; Brookshire *et al*, 1980; Knesch, 1990), and the field of health care (Garbacz and Thayer, 1983). While there are sound theoretical reasons why this might be so in principle for the studies mentioned above. Donaldson (1995) found that WTA had not been used in publicly funded health care systems. Respondents to WTP questions involving publicly-financed health care systems such as in the UK are told that the questions are hypothetical. and they will not be required to pay the named sum. Donaldson (1995) conducted a small study. in which 82 parents of schoolchildren covered by the Grampian Health Board were randomly allocated to WTP or WTA questions on publicly-funded child health services. Donaldson found that WTA questions received a higher number of invalid responses (*i.e.* no value, but a comment indicating that the respondent did not understand the purpose of the questions, such as protest comments). In the whole sample. 43 WTA questions were asked (involving WTA to lose the existing type of care) and 51% of responses were invalid. Two types of WTP questions were asked: WTP for a new type of care (133 questions), and WTP to keep an existing type of care (55 questions). The WTP for existing care performed best with an invalid response rate of just 2% compared to 18% for the WTP for new care. The WTA questions elicited comments such as "The money would come too late to be of any use" and "I would feel guilty about accepting the money – I would rather it were spent on services". Such responses were seen as invalid, because they reflected a lack of understanding of the WTA questions (*i.e.* the hypothetical nature of the question, which was asked in order to assess values rather than to actually make payments). However, they suggest that such problems could be more common for WTA than WTP in publicly-financed health care systems.

Mitchell and Carson (1989) argued for a property rights approach to the difficulty of inequalities between WTP and WTA. Since people are unable to sell their rights to health care, the concept of willingness to accept for losses or decreases in health care would not normally enter into consideration. However, WTP values may be asked for current access. Mitchell and Carson suggested use of WTP for gains and losses on these grounds. Thus people would be asked for WTP values for gains and also asked for WTP values for access to health care to remain the same.

An advantage of the WTP method is that it can place values on an entire package of goods and their attributes, for example the processes of treatment in addition to the treatment outcome. However, there are several problems with this methodology. One of these is that WTP may reflect ability to pay. Thus the rich may give higher WTP scores than the poor simply because they are able to pay more. In addition to the fact of their having more money and thus being able to pay more, it is probable that the rich are subject to diminishing marginal utility of monetary units, so they may state a higher level of WTP partly because each monetary unit is worth less to them that it would be to a poor person. It may be possible to use weights to adjust for different income groups (Friedman, 1984; Donaldson, 1995). However, in order to use adjustment weights based on income group, sufficient numbers of people from each income group are required to respond to the WTP survey. It is a well-known fact that lower response rates may be expected from people of lower income groups.

WTP is also open to strategic responses. People may give higher values than they actually would feel to be merited in order to try to ensure that a favoured policy is implemented. (This is also a potential problem with other non-monetary measures of HRQoL.)

It has been observed that WTP scores for the sum of parts do not always equal the WTP score given for the whole programme (this is known variously as part-whole bias, the embedding effect, and insensitivity to scope). For example, Kahneman and Knetch (1992) found that Toronto residents gave a WTP value for preserving the fish stock in all the lakes in the province which was only slightly higher than that given for a small proportion of the lakes. The authors suggested that this was due to the "warm glow effect" which results from expressing WTP for public goods in itself, no matter how great or small the benefit from those goods.

In summary, the WTP technique is prone to income bias, such that the rich may give higher WTP values than the poor because they have a higher ability to pay. Although this problem could, in theory, be alleviated by weighting responses according to income group, there are problems of recruitment in lower income groups. Since economic evaluation of health care usually follows the non-welfarist school, other measures of HRQoL have been developed based on valuing health, as described below.

2.6 Introducing the QALY

The QALY is a measure of life years adjusted for health-related quality of life. It combines quality of life and quantity of life into a single index measure. For example, consider a case where it is said that 10 years spent in a state of chronic renal failure are equivalent to 8 years in full health. It can then be said that the QALY value for 10 years with chronic renal failure is 8 QALYs. In order to calculate the number of QALYs associated with a health state of a given duration, a quality weight must first be obtained for that health state. Methods used to obtain quality weights are described in Section 2.6.1.

The QALY was developed by Fanshel and Bush (1970) under the name "function years". Researchers used the measure increasingly throughout the 1970s under the terms "index day", "health day", "health status unit day", "health status unit year" (Torrance, 1971; Torrance *et al*, 1972; Torrance, 1976b). It was Weinstein and Stason (1977) who first coined the QALY acronym (Drummond *et al*, 1997). The QALY concentrates on HRQoL rather than quality of life in a broader sense, which might incorporate the more material aspects such as number of cars per household, or types of holiday that may be taken by the individual. As such, the QALY fits into the non-welfarist view of externalities relating to health status (Brazier, 1998).

In the simplest scenario the individual remains in the same condition or health state throughout the rest of his life expectancy. Data on life expectancy and HRQoL are combined into a single index figure by multiplying the two as described in Equation 2.1 (Pliskin *et al*, 1980; Miyamoto and Eraker, 1985), where U(Q) is the value of health state Q. T is the duration of the state, and U(Q.T) describes the overall value of the two in combination.

$$U(Q,T) = U(Q) * T \qquad (2.1)$$

A series of health states (or one health state) occurring over a specified time period (*e.g.* life expectancy) is sometimes referred to as a health profile. For the more complex profiles that involve several periods in different health states occurring over the life expectancy of the individual, the HRQoL weight for each health state is multiplied by the duration of each health state, and the QALY score is the sum of the products as shown in Equation 2.2, where the QALY score is the sum of the health state values $U(q_i)$ where i = 1 to T, and t is the period of time in each health state q_i.

$$U(Q_T) = t \sum_{i=1}^{T} U_i\,(q_i) \qquad (2.2)$$

The result of the QALY calculation may be presented as a comparison between the life expectancy of the individual in the expected health profiles for the individual versus a shorter period in full health. Thus if the life expectancy of an individual with chronic renal failure is 10 years, the QALYs attached to the profile may be 8 years (figures purely for example purposes, not derived from literature). Thus 8 years in full health would be equivalent to 10 years with chronic renal failure. QALY values may also be calculated by taking serial measurements of health state values over time, plotting them, and estimating the area under the curve (Matthews *et al*, 1990; Diehr *et al*, 1997).

Information about condition-specific life expectancies can usually be obtained from consulting the epidemiological literature. HRQoL data for the health states may be obtained either from direct preference elicitation or indirectly via the application of weights obtained by preference-elicitation techniques, such as EQ-5D (Brooks. 1996) and HUI-III (Feeny *et al*, 1995).

The methods used to elicit preferences for health states directly are described in detail in Section 2.6.1. However, in brief the main methods are standard gamble (SG), time trade-off (TTO), and rating scales or visual analogue scales (VAS). Health states are described to respondents, who provide a quality weight between 0 and 1 using one of the elicitation methods just mentioned, where 0 is usually death and 1 is usually full health. Another option of direct elicitation is to obtain valuations of current health from patients who are in the health states. Preference-based weights are often obtained via the application of a generic questionnaire (*e.g.* EQ-5D), and in this method responses to a health status questionnaire are weighted using reference weights, which have been obtained from a general population using one of the preference elicitation techniques mentioned above.

The great advantage with the QALY methodology is that it has an enormous level of flexibility. If the preference weights for most of the health states associated with a given condition are known, they can be applied to any health profile containing these states. Thus it is not necessary to individually assess each possible health profile.

The QALY has been considered the "gold standard" for economic evaluations by the Washington Panel (Gold *et al*, 1996), and has been used by agencies such as the National Institute for Clinical Excellence (2002) in the UK, and the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia (Commonwealth Department of Health and Ageing, 2002). QALYs are widely used in the economic evaluation of health care resources.

2.6.1 Valuation techniques

As mentioned in the above section, a quality weight is needed for each health state that is entered into the QALY algorithm. The following sub-sections describe commonly used methods for obtaining quality weights for health states.

Rating scales

There are several different variants of rating scales. Some consist of categories such as "not at all", "moderately", "extremely" (Ware and Sherbourne, 1992). Other variants of rating scales frequently consist of a scale with values ranging from 0 to 1. When being used to value health states, the value of 1 is usually the best imaginable state of health, and 0 usually corresponds to death. The scale is assumed to have interval properties. This means that the intervals are of equal width everywhere on the scale, so that the difference between 0.6 and 0.7 is the same as the difference between 0.3 and 0.4 (Drummond *et al*, 1997). A commonly used variant of the rating scale is the visual analogue scale (VAS), which consists of a line with end points (*e.g.* full health and death) and with or without other defined points on the line (Drummond *et al*, 1997). Rating scales are a very simple form of measurement involving the valuation of states under conditions of certainty. Figure 2.1 shows the version of the rating scale technique used in Chapter 7 to obtain values for varicose veins treatments.

Standard gamble

The SG technique is commonly considered as the gold standard for health state valuation (Gold *et al*, 1996). It is based on the axioms of the expected utility theory (von Neumann and Morgenstern, 1944; Bell and Farquhar, 1986; Drummond *et al*, 1997), which are described in Section 2.6.2. The reason for the popularity of the SG technique is that, unlike VAS, it attempts to take the riskiness of decision-making into account. In the SG the individual is presented with the following type of choice:

> You must decide whether to accept a given treatment. If you do not accept the treatment you will be in a chronic health state *X* with certainty. If you accept the treatment, there are two possible outcomes: there is a probability (*p*) that you will end up in full health; there is a probability (1-*p*) that you will end up dead.

The individual must choose a value of *p* such that she is indifferent between the certainty of state *X* and the probability *p* of being in full health (Figure 2.2). The choice may be adapted so that best reference state may be a state less than full health, or the worse reference state may be a state better than death. There are also adaptations for valuing states worse than death (Drummond *et al*, 1997).

Time trade-off

The TTO methodology was first developed for use in evaluation of health states by Torrance *et al* (1972). It does not take account of risk in the way that SG is alleged to, but is concerned with preferences for trade-offs at the end of life. Individuals are presented with the following type of choice:

> You must decide whether to accept a given treatment. If you do not accept the treatment you will be in health state *N* for time *t*. If you accept the treatment, you will be in health state *H*, which is better than *N*, for time *x*. Of the two time periods, $t > x$.

The TTO method uses the formula $U(H) = x/t$.

Although TTO does not have a fundamental basis in the axioms of expected utility theory as the standard gamble does (Feeny and Torrance, 1989; Gafni *et al*, 1993), it is commonly used to derive QALYs (Figure 2.3). SG and TTO are compared in terms of reliability, feasibility and validity in Chapter 4.

Magnitude estimation

In magnitude estimation respondents are asked to compare pairs of health states and indicate how much better or worse one state is compared to another state. Thus it is a form of ratio scaling (Brazier *et al,* 1999). Magnitude estimation was developed as an alternative to VAS (Stevens, 1966), and was the basis of scoring for the Rosser Index (Rosser and Kind, 1978; Rosser and Watts, 1978). However, magnitude estimation has not been widely used in health economics.

Person trade-off

Whereas the techniques described in the previous sub-sections ask respondents to consider choices at the level of the individual, the person-trade-off (PTO) technique employs a social perspective. The PTO asks respondents to consider two groups of people. Group A contains x people in a certain health state, and Group B contains y people in another health state. The task of the respondent is to choose to which group health care resources should be allocated. The values of x and y may then be varied until the respondent is no longer able to choose between the two groups. If the condition described in Group B is considered worse than that in Group A, the undesirability of condition B is said to be x/y times that of condition A (Brazier *et al,* 1999).

Discrete choice experiments

In discrete choice experiments scenarios are described in terms of their characteristics or attributes (Ryan and Farrar, 2000). The levels of each attribute vary between scenarios, and respondents choose between the scenarios. This enables researchers to elicit the relative importance of the different attributes. Since there are often hundreds or even thousands of possible combinations of attributes, one way of incorporating as many choices as possible is to divide the number of choices between the sample of respondents and conduct the analysis by choice sets rather than at the level of the individual respondent (Viney and Savage, 2003). However, this method for valuing health states is still under development (Viney *et al,* 2005).

2.6.2 Theory underlying the QALY model of preferences

The QALY model is based upon a theory of rational decision making under uncertainty called expected utility theory (EUT), which was developed by von Neumann and

Morgenstern (1944). The EUT is seen by some as a normative model, which describes how people *should* behave under uncertainty. not necessarily how they *do* behave (Drummond *et* al, 1997). The theory assumes that people will maximise expected utility, and has the following four axioms (Bell and Farquhar. 1986: Drummond *et al*. 1997):

- Completeness of preferences – For any two risky prospects. one is preferred to the other or they are liked equally (indifference).

- Transitivity of preferences – For any three risky prospects A, B and C, if A $\succ$ B and B $\succ$ C, then A $\succ$ C. (The "$\succ$" sign depicts "is preferred to".) Similarly. if there is indifference between A and B, and indifference between B and C. then there is also indifference between A and C.

- Independence – In the words of Basili (2000), this axiom "states that given two alternatives (lotteries in technical language), each composed of an action and a common act, preferences between them should be independent of any common consequence with identical probability".

- Continuity of preferences – If A $\succ$ B, and B $\succ$ C, there is a probability *p* at which an individual is indifferent between B under certainty and the gamble where there is probability *p* of outcome A and probability 1-*p* of outcome C (Drummond *et al*, 1997).

These axioms allow utility to lie on a linear scale from 0 to 1. where 0 and 1 may be arbitrarily defined (Schoemaker, 1982; Friedman and Savage, 1948). It follows from these axioms that a greater level of utility is synonymous with a greater level of preference. In other words, the expected utility of lotteries preserves the rank order of the lotteries (Schoemaker, 1982). These axioms allow any point on the utility function to be determined (Friedman and Savage. 1948). For example. if a person is faced with a choice between a gamble involving a probability p of receiving £500 or 1-p of receiving £1000. £500 can be assigned a utility value of 0. and £1000 a utility value of 1. The probability p at which the person is indifferent between this gamble and the certainty of receiving £600 indicates that the utility of receiving £600 is p. For example. if this indifference point is p=0.6. the utility of receiving £600 is 0.6. The above axioms allow any point on the utility curve to be determined in this way. incorporating the individual's attitude to risk.

There is plenty of evidence that preferences are not pre-formed and complete. In health economics, Shiell *et al* (1997) have argued against the view that preferences are pre-formed data waiting to be collected by researchers. They suggest that for most goods preferences are formed by experience of those goods. Obviously, the more experience a person has of a set of goods, the better position he is in to formulate a set of preferences. For mild health states such as headaches, Shiell *et al* propose that most people have wide and relatively frequent experience of these, and may therefore have a formed set of preferences. However, most people do not have wide experience of a variety of the more serious health states. For such health states, therefore, preferences are unlikely to be pre-formed. Rather, they are likely to be incomplete and unstable. This hypothesis is supported by the work of Shiell *et al* (2000), Bernstein *et al* (1997, 1999), Chapman *et al* (1998), Bleichrodt and Johannesson (1996), and Bazerman *et al* (1992).

Life is full of uncertainty, and people often do not know the outcome of a choice until the choice has been made. The expected utility of a choice can be calculated (provided the risks are known) by multiplying the utility of each possible outcome by the probability of the outcome occurring. If all risks for each outcome are known, the sum of the probabilities should be equal to one. EUT assumes that individuals will maximize expected utility by choosing the option with the highest expected utility (Friedman, 1984).

For the QALY equation described in (2.2) to hold, the additional assumptions set out at the end of this chapter in Table 2.1 must be upheld (Pliskin *et al*, 1980; Bleichrodt and Johannesson, 1996). These assumptions are over and above the four EUT axioms just described. Each of the assumptions is described in detail in the following subsections of this chapter.

2.6.2.1 Constant proportional trade-off and its implications

The axiom of constant proportional trade-off implies that the proportion of remaining life that a person is willing to forgo is independent of the amount of life remaining. Thus if a person is indifferent between four years in full health and five years in a state less than full health, he should also be indifferent between 16 years in full health and 20 years in the lesser state. This assumption is specifically relevant to the application of the TTO method of valuing health and may be relaxed in the more general form of the QALY (Dolan, 2000).

Zero effect of duration

A direct implication of the constant proportional trade-off assumption is that the value of a health state is independent of its duration. Thus, according to this assumption, the weight given to a health state is the same regardless of whether it lasts one year or 15 years.

Zero time preference

Another implication of the constant proportional trade-off assumption (relating specifically to the TTO-QALY) is that individuals should have a time preference of zero. Time preference may be defined as a preference for an event to occur with particular timing, for example sooner in the future rather than later or *vice versa*. A person may have a choice between receiving a present in one year or in five years. If she chooses to receive it in one year, she prefers to receive goods sooner, and is said to have a positive time preference. If, however, she chooses to receive it in five years, she prefers to receive goods later, and is said to have a negative time preference. Indifference between receiving goods sooner or later implies a zero time preference.

For the application of the TTO method of valuing health, the QALY algorithm assumes a zero time preference. This ensures that valuations are stable with regard to the point in time at which the health occurs. Thus an individual should place the same value on any given health state regardless of the point in their life at which it would occur. This is a necessary addendum of the constant proportional trade-off axiom. There is evidence that a significant proportion of people do not demonstrate a zero time preference (Cairns and van der Pol, 1999). This has led to attempts to adjust QALYs by discounting for time preferences (Cairns and van der Pol, 2000).

Zero quantity effect

The quantity effect, or strength of preference (Dyer and Sarin, 1982), is the effect of diminishing marginal returns upon utility. This might have the effect of causing people to value future years of life lower than earlier years because they more urgently want life if they have less. However, it could work the opposite way around. People may place greater values on later years of life, because they then have fewer years remaining. The constant proportional trade-off axiom leads to an assumption of zero quantity effect.

2.6.2.2 Mutual independence of life quality and quantity

This means that if a person is indifferent between nine years in full health and the gamble where there is 50% probability of 15 years in full health and a corresponding 50% probability of five years in full health, this person should also have this indifference point if the state of full health is replaced by another health state. As Dolan (2000) explains, "… U(Q) depends only on the health state irrespective of the number of life years and U(T) depends only on the number of life years irrespective of the health state". This ensures that the weighting given to a health state during an SG valuation procedure will remain constant regardless of the time horizon under consideration, and can be applied to the state no matter what the duration of that state just by multiplying the weight by the duration. This axiom is necessary to the SG procedure.

2.6.2.3 Constancy of risk attitude

It was once assumed that risk neutrality was necessary in order for the QALY algorithm to reflect preferences, thus survival with respect to life years would be linear. This would mean that values obtained by TTO should be equivalent to those obtained by SG. However, there is evidence that many people are not risk neutral (see Chapter 3), and the two methods often give rise to different results (Dolan, 1998b; Torrance, 1976a; Reed *et al*, 1993; Bleichrodt and Johannesson, 1997; Wolfson *et al*, 1982; Read *et al*, 1984; Stiggelbout *et al*, 1994; Lenert *et al*, 1998).

Risk neutrality is not now considered to be essential, and methods have been developed to adjust QALY valuations for risk attitude (Miyamoto and Eraker, 1985). However, individuals should demonstrate a constant attitude to risk through time, and risk attitude should be independent of the health state (Dolan, 2000).

2.6.2.4 Additive separability of utility over time

The additive utility axiom follows from the axiom of independence, and leads to the conclusion that one health state is totally independent and separable from another, and that utility weights are directly proportional to the duration spent in a given state of health. This means that it should be valid to value profiles indirectly using the values given to discrete states, as is done in the QALY method. This makes the job of valuing health profiles easier for health economists and their subjects, because the number of valuation exercises is reduced since the valued states can be combined to produce many different profiles.

It follows that preferences should be context independent, which means that the values an individual places on a particular health state should not depend upon preceding or succeeding health states. In other words, the sequence of a series of events should not affect the utility attached to the series. Nor should the value given to a state depend upon the individual's current health.

However, prospect theory suggests the opposite, that the value given to a health state very much depends upon the current status of the valuer, with the utility function being S-shaped (Treadwell and Lenert, 1999). One consequence of the S-shaped function is that losses and gains are valued differentially. Losses lead to a steeper loss in utility than the increase in utility produced by a gain of equal magnitude. Secondly, the value given to a particular health state will depend upon whether that state is viewed as a gain or loss in health to the valuer, and how much of a gain or loss the state represents.

The additive utility axiom implies that small duration effects should be negligible. The implications of this last statement are that even if a health state of very short duration is extremely unpleasant, it will have negligible effects on the QALY score of the overall sequence occurring over a long duration, because the overall effect on utility is a function of the duration of the state. This is of relevance to temporary health states associated with processes of treatment. According to the QALY algorithm, the overall value of a profile of health should be only negligibly affected by alternative treatment processes and the short-duration, temporary states associated with them.

2.6.2.5 Conclusions

As Table 2.1 and the preceding sub-sections make abundantly clear, the application of the QALY algorithm assumes that people have pre-formed preferences following clear-cut and restrictive guidelines. There is increasing evidence that the underlying assumptions of the QALY algorithm are regularly violated at an individual level (see Chapter 3). This has led to concerns that true preferences may not be represented by the QALY model. If this is the case, there could be serious implications for economic evaluation and the distribution of health care resources.

2.7 The healthy years equivalent

2.7.1 Background

The healthy years equivalent (HYE) method was first suggested by Mehrez and Gafni (1989) as an alternative to QALYs. Also based on EUT, it was intended to provide a holistic method for the direct valuation of lifetime health profiles, avoiding some of the restrictive assumptions involved in the QALY procedure. A lifetime health profile would typically consist of a sequence of different health states, each of which would last a period of time before giving way to the next state. The HYE method may also be applied to the valuation of a chronic health state.

Mehrez and Gafni proposed that HYEs should be obtained by a two-stage SG procedure. The first stage would consist of a typical SG in which the utility of the health profile would be ascertained. Thus for a profile consisting of a chronic state (Q,T), where Q is the health state and T is the duration, the individual would be presented with a choice between the certainty of (Q,T) versus probability *p* of full health (Q*) or probability 1-*p* of death (Q^0). The probability *p** at which the individual is indifferent between the gamble and the certainty is taken as the utility of the health profile (Q,T). The second stage of the HYE elicitation process is a certainty equivalent exercise in which the individual is asked to state the number of years in Q* which are equivalent to the gamble *p** Q* : 1-*p** death (Figure 2.4). The number of healthy years elicited in this stage is the HYE value (H*). The two stages of the HYE elicitation procedure are outlined in equations (2.3) and (2.4).

$$(Q,T) \sim p^* (Q^*,T) + (1-p^*)(Q^0) \qquad (2.3)$$

$$(Q^*, H^*) \sim p^* (Q^*,T) + (1-p^*)(Q^0) \qquad (2.4)$$

As indicated in (2.3) and (2.4), (Q,T) and (Q*,H*) are both equivalent to the same gamble. It therefore follows that, according to the law of transitivity, there must be equivalence between (Q,T) and (Q*,H*) (see (2.5)). According to Mehrez and Gafni, there is equivalence between the utility of the profile and the number of healthy years that are equivalent to the profile. This is shown in (2.6). Mehrez and Gafni (1989) argue that the equivalence between these two profiles means that they are on the same isoutility curve.

$$(Q^*,H^*) \sim (Q,T) \qquad (2.5)$$

$$U(Q^*,H^*) \sim U(Q,T) \qquad (2.6)$$

The HYE method would more commonly be applied to health profiles consisting of a sequence of health states of differing durations. If the health states at each period 0 to T are represented by Q_0 to Q_T, the formula for the HYEs of any profile is shown in (2.7).

$$U(Q^*,H^*) \sim U(Q_{0,\ldots,T}) \qquad (2.7)$$

Mehrez and Gafni favour HYEs over QALYs because of their belief that their two-stage process allows valuations to be made which truly reflect the preferences of individuals by using information obtained directly from their utility functions. Because the profile is valued as a whole, the evaluation process is not restricted by all the assumptions underlying the QALY algorithm. For example, the additive utility function is no longer necessary because the constituent health states are no longer valued as discrete states and multiplied by their duration. Mehrez and Gafni (1989) and Gafni *et al* (1993) also argue that the assumption of risk neutrality is not required, because they claim that the first stage of the elicitation procedure accounts for the individual's risk attitude. It is also argued that no assumptions are necessary about time preferences, or constant proportional trade-off, because the individuals time preferences will be taken into account in their utility function.

2.7.2 Criticisms of the HYE

Many have argued that the two-stage HYE algorithm is equivalent to the TTO-QALY method for valuing health profiles (Buckingham, 1993; Culyer and Wagstaff, 1993b; Johannesson *et al*, 1993; Loomes, 1995). The TTO-QALY method of valuing a chronic health state (Q,T) consists of describing state Q and explaining that it will last for duration T. The individual is asked to state how many years (X*) in good health (Q*) would be equivalent to (Q,T), where Q < Q* and T > X*. The profile (Q,T) is said to be equivalent to (Q*,X*), as set out in (2.8). This is a similar outcome to the HYE procedure, in which the individual is also asked via the two-stage process for the number of years in good health (Q*,H*) equivalent to a health profile (Q,T), where Q < Q* and T > H* (see (2.5)). The EUT axiom of transitivity states that if (Q,T) is equivalent to (Q*,X*) and (Q,T) is equivalent to (Q*,H*), then (Q*,X*) must be equivalent to (Q*,H*) (see (2.9)). Thus X* must be equivalent to H* (2.10). It has therefore been argued that the two-stage HYE algorithm and the TTO-QALY as applied to profiles are theoretically equivalent.

$$(Q,T) \sim (Q^*,X^*) \qquad (2.8)$$

$$(Q^*,H^*) \sim (Q^*,X^*) \qquad (2.9)$$

$$H^* \sim X^* \qquad (2.10)$$

Gafni *et al* (1993) responded to the transitivity argument by suggesting that the TTO method evaluates health states under certainty, whereas their two-stage algorithm for HYEs evaluated health states and profiles under uncertainty. They argued that the TTO method obtains values, but the two-stage HYE method obtains utilities. This argument is based on the fact that the two-stage HYE uses SG in the first stage, which is based on the continuity of preferences axiom of EUT. Evaluations using SG take into account uncertainty, and there is seldom certainty in medical decision-making. The TTO, however, obtains valuations under certainty, and does not account for attitudes to risk. Thus, according to the view of Gafni *et al*, the axiom of transitivity does not apply to this debate because U(Q,T) and V(Q,T) need not be equal.

However, Dyer and Sarin (1982) argued effectively that two goods which are equal in value will also be equal in utility (see (2.11)). Equations (2.6) and (2.11) give rise to (2.9) and (2.10).

$$V(Q,T) \sim V(Q^*,X^*) \therefore U(Q,T) \sim U(Q^*X^*) \qquad (2.11)$$

Johannesson *et al* (1993) also argued that HYEs did not take risk attitudes into account, even though they allowed that HYEs were free from the assumption of constant proportional trade-off and time preferences. They argued that in the first stage of the process to elicit HYEs, a risk averse individual would give a higher value for *p* than a risk neutral individual, because she would tolerate a lower risk of death. A risk averse individual would, however, give a lower value for H* than a risk neutral individual. Johannesson *et al* (1993) argued that the higher value of *p* in the first stage, and the lower value of H* in the second stage would produce an equal and opposite effect on the overall HYE value of a risk averse individual. They argued that this effect would cancel out any sensitivity to risk attitude. From this argument they concluded that the two-stage algorithm for HYEs assumed risk neutrality, and that health states/profiles were in fact being measured under certainty, as is done in the TTO.

2.7.3 Alternative HYEs

Even though the general consensus among health economists is that the two-stage approach to measuring HYEs is equivalent to the TTO-QALY, the concept of valuing whole health profiles as an alternative to the QALY model with all its restrictive assumptions has attracted interest in many quarters (see for example Buckingham, 1993; Mackeigan *et al*, 1999).

Johannesson *et al* (1993) suggest an alternative method for measuring HYEs to the two-stage procedure. This would use a version of the TTO in which the profile contained uncertainty. The respondent would then decide the number of certain years that would be equivalent to the risky profile (as used in Chapter 8). Johannesson *et al* referred to this as the *ex ante* HYE, and made the distinction between this and the *expected* HYE. The expected HYE takes the *ex post* perspective, in which the respondent provides values for each health profile, which are then entered into a decision tree in which they are multiplied by the probability of that profile occurring. The original two-stage method put forward by Mehrez and Gafni (1989) is an example of the expected HYE. Drummond *et al* (1997) suggest valuing health profiles holistically by using the SG. This would comprise simply using the first stage of the original two-stage process.

However, as Johannesson *et al* (1993) rightly pointed out, replacing QALYs with HYEs would vastly increase the measurement burden. The obvious advantage of the QALY algorithm is that it allows researchers to measure HRQoL for a limited number of separate health states, and then use these weights to calculate QALYs for any health profile that could be obtained from mixing these states in any combination of sequence and duration. The weights obtained in this way can be used in complex decision models involving probability distributions of health profiles. However, in order to enter HYEs into such models it would be necessary to obtain valuations for every possible profile. Even providing an individual with descriptions of all the health states within one lifetime profile may leave him with a very demanding task. Each health state might take as much as half a page of text (Drummond *et al*, 1997).

Despite the disadvantages of the holistic approach to valuation of health profiles, this approach has theoretical superiority to QALYs, because the restrictions required for the QALY are not required for the holistic approach. As will be demonstrated in Chapter 3, empirical studies have shown there to have been widespread violations of the axioms underlying the QALY algorithm. If the holistic approach is found to provide a more accurate reflection of preferences than the QALY approach, it is worth putting more research into it to determine whether it can become a viable alternative to QALYs for the estimations of preferences over health profiles.

Because the term "HYE" was used originally to refer specifically to the two-stage standard gamble approach originally suggested by Mehrez and Gafni, this thesis will use the term "holistic" to refer to the general approach of valuing health profiles directly rather than using their composite health states. This thesis will be exploring other approaches to valuing health profiles than the two-stage gamble.

2.8 QALYs and HYEs: Where do they fit into the welarism versus non-welfarism debate?

As already explained, QALYs are not generally considered to be utilities. QALYs traditionally relate just to the part of the utility function relating to health (Garber, 1999). They are therefore used as a measure of HRQoL rather than QoL more generally. They are often seen as non-welfarist in nature (Culyer, 1991; Brouwer and Koopmanschap, 2000), and follow from the view that the public wish each other well, but are not necessarily prepared to pay for each other to be happy (Evans and Wolfson, 1980). In not allowing the entire utility function to be considered, they ignore utility derived from other things not directly related to health (*e.g.* utility from continuing to smoke instead of giving it up for health reasons). Externalities such as wishing others healthy are particularly relevant to publicly financed healthcare systems such as the NHS in the UK.

Rather than including the whole of the individual's utility function, just the health-related argument is allowed in QALYs and a decision maker may be left to make decisions regarding important non-health items such as justice. The non-health items that the non-welfarist may include depend on the political and societal views of the day (Garber, 1999).

Holistic valuations are also based in non-welfarism. Like QALYs, holistic valuations are concerned with the source of utility, and are also concerned with health outcomes rather than the complete utility function. They may include a wider range of factors than QALYs. For example, holistic valuations are more easily able to incorporate individual risk attitudes and time preferences than QALYs, because respondents can see the whole profile and the way in which it extends over time when they value it. When valuing the composite health states using the QALY method, the respondents are not shown the way in which the health profile will unfold over time (*e.g.* ordering within the sequence), and the *ex ante* risks involved. Holistic valuation methods are theoretically better at reflecting preferences than QALYs, because they rely on fewer restrictive assumptions.

Both QALYs and the holistic valuation method have the underlying non-welfarist aim to avoid the effects of income on valuations of health. In the case of the holistic method, respondents are asked to consider profiles of health. In some cases these may last years, or even for the rest of the respondent's lifetime. The use of the holistic method to elicit values, and the freedom from the restrictive assumptions of the QALY, may allow slightly more of the utility function to be taken into consideration than the traditional use of QALYs. This is because, although the focus of the holistic approach is still on health, there may be a greater opportunity for respondents to consider certain non-health aspects in the specific context of the profile (*e.g.* preferences over treatment procedures *per se*, rather than merely their effect on health). The extent to which non-health outcomes should be taken into account is to some extent a matter for policy-makers choice (hence non-welfarist).

2.8.1 Attempts to link cost utility analysis with welfarism

Despite the traditional basis of QALYs in non-welfarism, there have been attempts to create a link between QALYs and welfarism (Dolan and Edlin, 2002; Hansen *et al*, 2004; Edlin, 2004). Bleichrodt *et al* (2002) discuss the possibility of devising "personalised QALYs". These would take into account the fact that the same health state might be given a different value by different people. Some non-welfarists believe that health state x should have a particular value regardless of who values it (Dolan, 2000). However, Bleichrodt *et al* attempted to define equivalent states for different groups. For example, this could be applied between sub-populations with different capacities for health attainment, and between each sub-population and the general population. However, there has not been much empirical progress in linking QALYs with welfarism. Most studies involving QALYs are based in non-welfarism, concentrating more on health than the broader utility function (Drummond *et al*, 1997).

This theoretical work is equally relevant to the holistic valuation approach, which also has a non-welfarist foundation.

2.9 Conclusions and the aim of this thesis

Welfarism is concerned with efficiency issues in relation to resource allocation, and also with normative issues relating to fairness and equity. According to welfarists, social welfare is the sum of individual utilities within society. Any improvement in social welfare is therefore desirable in the eyes of a welfarist, regardless of the source of the increase in utility. Welfarists believe that health is just one argument in the utility function and all the different arguments of the individual's utility function should be taken into account when judging the value of health care (Dolan, 2000; Birch and Donaldson, 2003).

There are several non-welfarist views within welfare economics. Many non-welfarists hold the view that the source of utility should be taken into account. The non-welfarist view in a health economics context tends to be that health is the output of importance, because when attempting to allocate a healthcare budget health is the output that matters. According to the non-welfarist school, therefore, devices are required that measure HRQoL rather than QoL more broadly. However, some in the non-welfarist school believe that the decision-maker should have the freedom to incorporate factors outside individual utility into the resultant value.

Cost benefit analysis is one method of obtaining values for health. However, as discussed in Section 2.5, the CBA approach has certain problems. One problem is that WTP values reflect ability to pay, and difficulties have been encountered in applying equity weights to responses based on income group.

Due to market failures associated with health and health care (see Section 2.3), there is widespread government funding of health care around the world (Dolan, 2000). This large scale public funding of health care, alongside the issue of caring externalities, leads to the argument that the tax-paying public may care enough to pay for increases in the health argument of the utility function of others, but not to pay for increases in other parts of the utility function of others (see Section 2.2). Thus a valuation system is required that values health, rather than other sources of utility (*e.g.* leisure activies).

The QALY was developed as a measurement of HRQoL. It is traditionally based in non-welfarism, and pays attention to the source of utility (health). Its non-welfarist framework also allows decision makers to weight QALYs according to other factors they may think important, such as sources of health improvements (*e.g.* whether a health improvement in one person is directly based on a decline in the health of another).

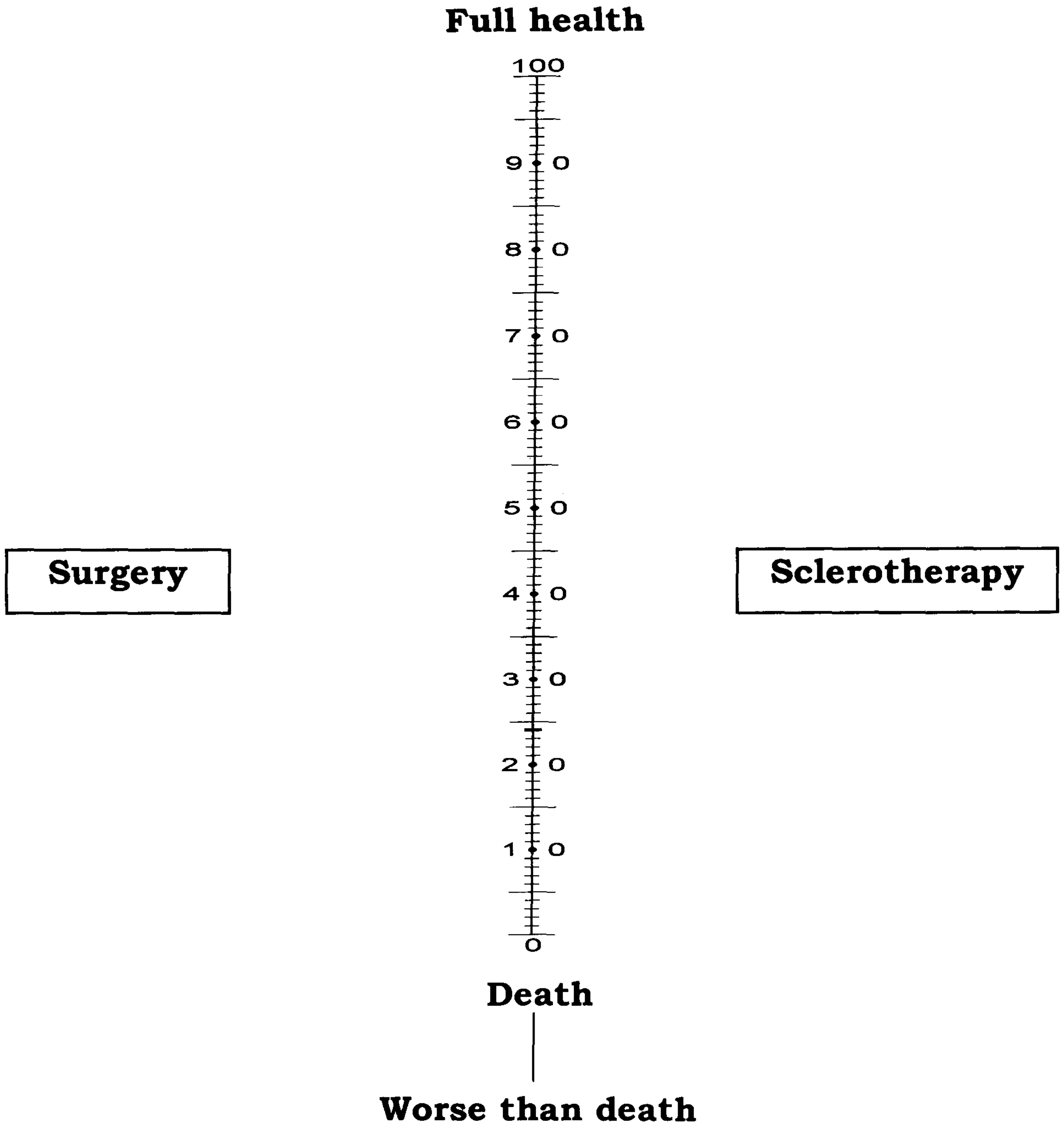
The QALY is based on a number of restrictive assumptions (see Section 2.6). There is evidence that there are violations of the assumptions underlying the QALY (see Chapter 3 for more detail of this evidence). Concerns over the validity of the underlying assumptions of the QALY have led to the development of the holistic approach as an alternative means to valuing health. The holistic approach still follows the non-welfarist school of thought, in that it is also concerned with valuing HRQoL rather than being all-inclusive of the entire utility function.

It is true that the holistic approach may in some cases reflect non-health preferences to a greater extent than the QALY. For example, in a profile describing a sequence of health states including the process of treatment, the respondent to the holistic approach might more easily incorporate non-health preferences over treatment process in the context of that profile of health. These are not usually taken into account with the QALY. However, both methods are too greatly removed from the welfarist school for this to be called a move towards welfarism. If the entire utility function should be taken into account, it may be better to divert resources to improve CBA methods than to develop the holistic health profile valuation approach.

This thesis contains four empirical studies that test the axioms underlying the QALY, and compares QALYs and holistic valuations within a non-welfarist framework. The specific QALY axioms being tested are described in detail in Chapter 4. In these studies, values are collected for health states and profiles relating to irritable bowel syndrome, varicose veins, and abdominal aortic aneurysms. The QALY and holistic valuation methods are both used to value health relating to specific health profiles. The difference between them is in the underlying methods. Thus, rather than comparing welfarist and non-welfarist valuation methods, this thesis compares two valuation methods that are both based in non-welfarism. For reasons described above, the welfarist framework is not considered acceptable by many in health economics.

Chapter 3 reviews the existing literature on empirical research into the extent to which people violate the QALY axioms, and empirical studies using holistic profile valuation methods. Chapter 4 describes the overall research methods used within the thesis. Different methods will be used to measure holistic profiles in Chapters 5 to 8. The advantages and disadvantages of each method will be discussed in the context of the disease conditions under study. This thesis will describe comparisons between holistic and QALY valuations of the same health profiles. The extent to which people violate the axioms of the QALY algorithm, and the extent to which QALYs and holistic valuations deviate empirically will be examined.

Figure 2.1 The rating scale and instructions used in Chapter 7 for eliciting values for treatments associated with varicose veins.



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

We would like you to indicate on this scale how good or bad the above processes are in your opinion. Please do this by drawing a line from the boxes beside the scale to whichever point on the scale indicates how good or bad the process is.

If you consider one or both of the above processes to be worse than death, please draw the line to the "worse than death" point at the bottom of the scale.

Figure 2.2 The standard gamble.

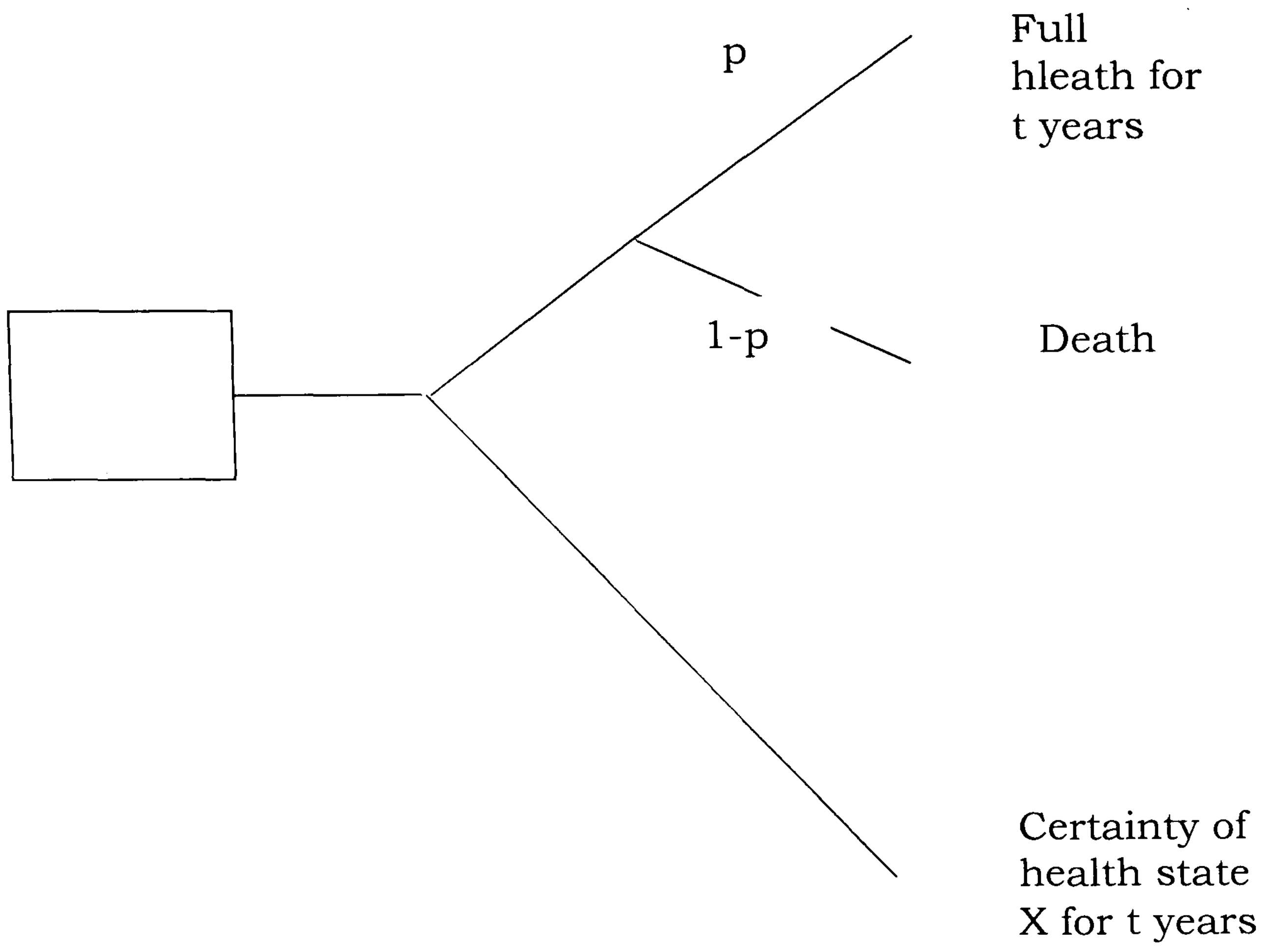

p
Full hleath for t years
1-p
Death
Certainty of health state X for t years

Figure 2.3 A diagram of TTO, with value of health on the vertical axis and time on the horizontal axis. The individual may choose between health state N for time t and helath state H for time x.

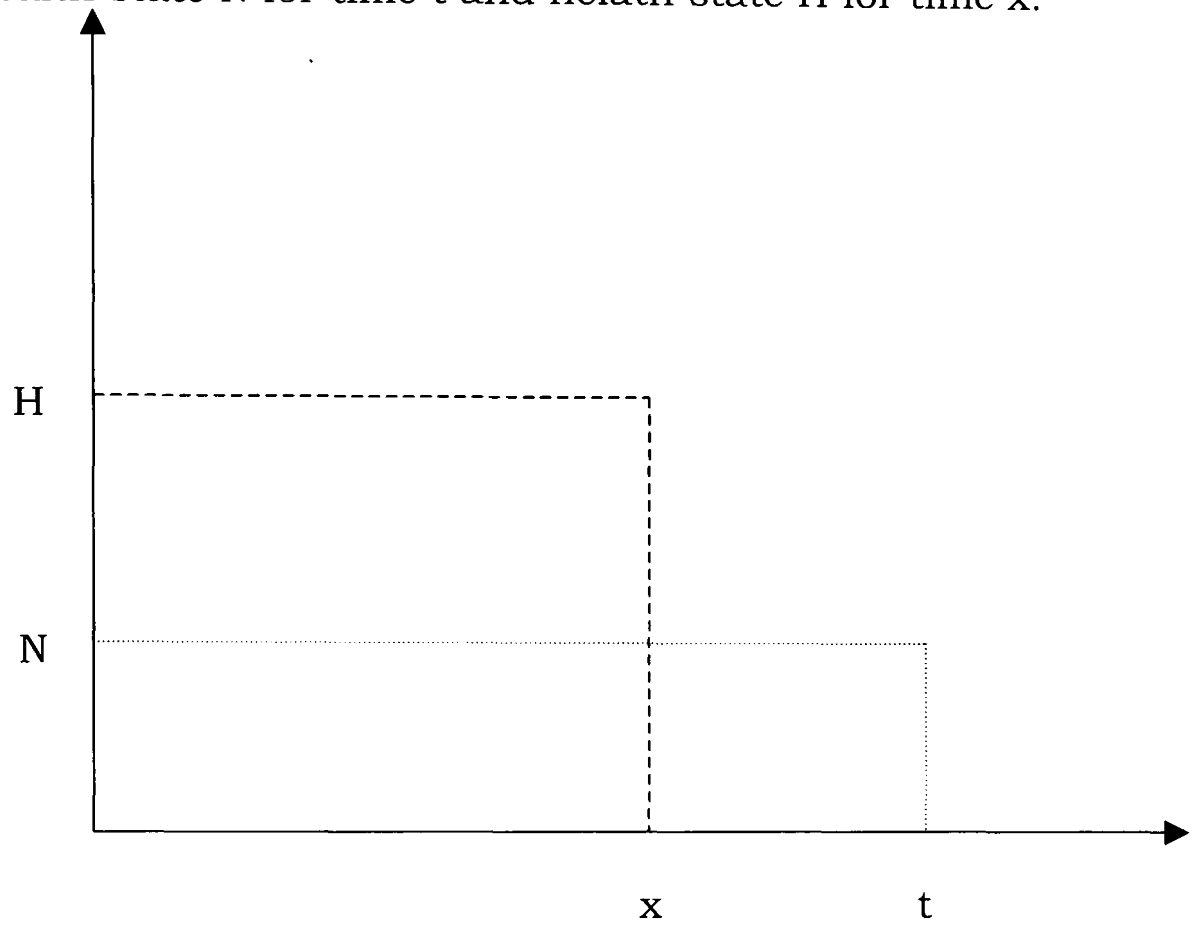


Figure 2.4 Diagrammatic representation of the two-stage HYE elicitation process.

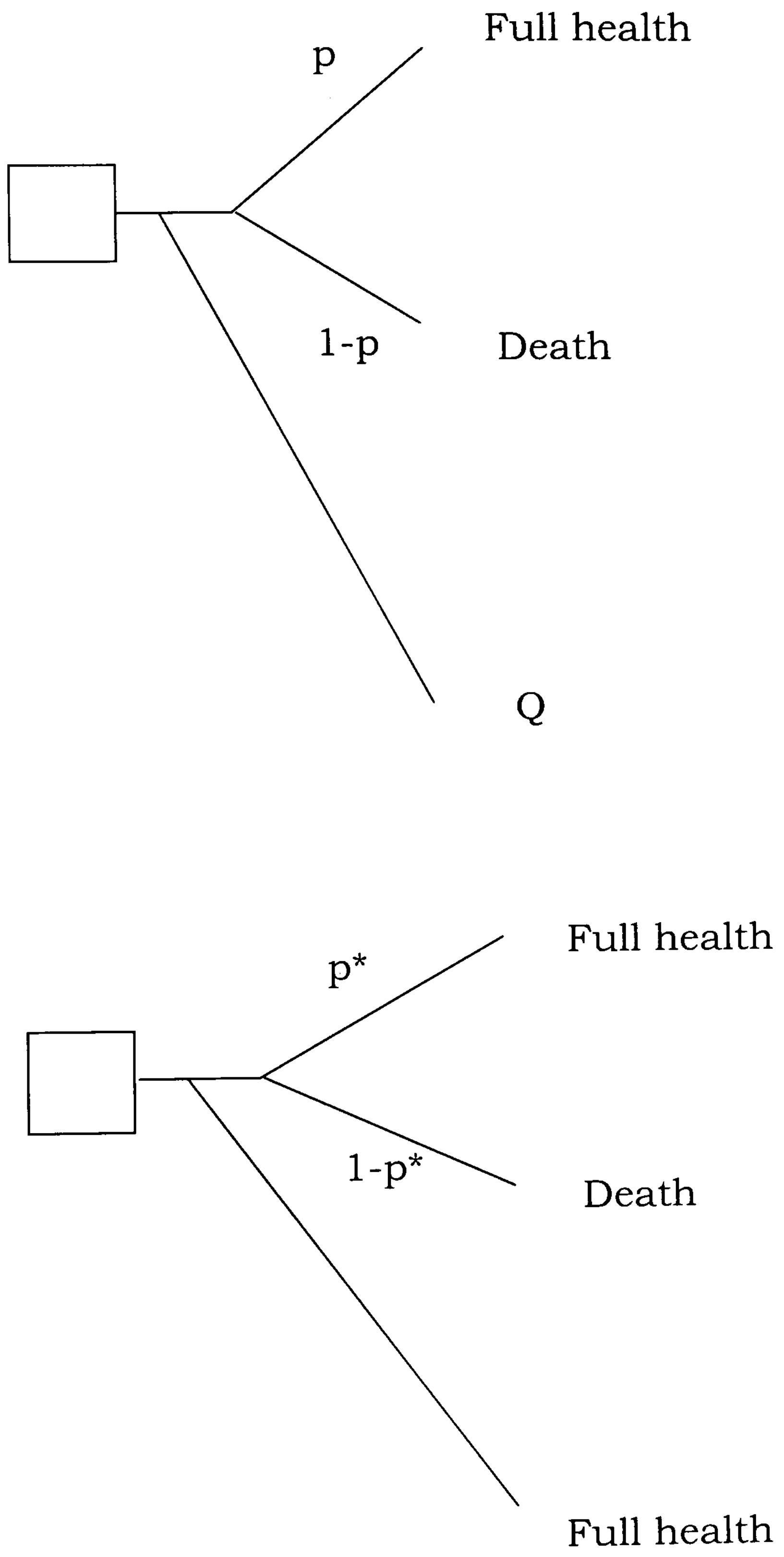


Table 2.1 Additional assumptions to EUT underlying the QALY algorithm. The main assumptions are presented in bold font, and the implications of these assumptions are presented below in regular font.

1. Constant proportional trade-off Duration Zero time preference Quantity effect
2. Mutual independence of life quality and quantity
3. Constancy of risk attitude to survival duration Risk neutrality under all health states
4. Additive separability of utility over time Valuations of health state independent of succeeding/preceding health states Small duration effects (*e.g.* temporary health states and short-term states associated with treatment processes) are negligible to the patient

Chapter 3

Reviews of the Empirical Literature on QALYs and HYEs

This chapter reports two literature reviews. The first was of the empirical literature into violations of the axioms underlying the QALY algorithm. The second review was of empirical research into HYEs or holistic valuations of health profiles more broadly.

3.1 A review of the empirical literature into violations of the axioms underlying the QALY

3.1.1 The literature search

Methods of the search

Before embarking on a study to examine the extent to which the QALY assumptions are violated, a review of the literature was conducted to determine what previous research had been conducted and the findings. This literature review consisted of three parts:

1) A review of empirical literature in ATHENS databases.

2) A search of the ATHENS databases and the bibliographies of articles obtained from (1) for the names of researchers in the field of QALY/holistic profiles valuation research.

3) A bibliographical search for empirical literature in non-health economics sources from the bibliographies of (1) and (2).

There were many words that could have been chosen to be used in the ATHENS search strategy, but words relating to QALY were used. This was because any article containing information on empirical research into the assumptions of the QALY model would almost certainly refer to the QALY. The following search terms were used:

- qaly$

- quality adjusted life year$

Results of the search

Databases were searched from 1966 to the end of 2002. A total of 2376 papers were found using the above search terms. However, the majority of these were not empirical studies of the extent to which the QALY axioms hold. A total of 18 papers were selected from the ATHENS search for inclusion in this review.

A great deal of empirical work has been conducted in the fields of psychology, for example, and this has not been efficiently married to research on underlying preferences in health economics. It was felt that a review of empirical studies into the assumptions of the QALY model would be incomplete without including the relevant non-health economics literature. The bibliographies from the ATHENS searches were searched for relevant empirical research using names of authors suggested by colleagues and in the bibliographies of papers.

Papers were selected for inclusion in this review if they described empirical findings relating to behaviour of relevance to the axioms of the QALY algorithm. This was a review of general findings rather than a critical review of the studies involved. The review is divided into sections that discuss research in the areas of each QALY axiom. The results are also shown in tabular form in Table 3.1 at the end of the chapter.

Sample sizes

Also of concern within this thesis are the issues of sample sizes and difficulties in recruitment of the desired sample populations for valuation studies. These issues will be discussed subsequently in the thesis, and it is worth briefly noting the range of sample sizes used in these studies.

For studies looking at the effects of duration on valuations of health state or scenarios there was quite a wide range of sample sizes, ranging from 20 (Sutherland *et al*, 1982) to 246 (Sackett and Torrance, 1978). The sample populations were also varied, including convenience samples such as undergraduate students (Bleichrodt and Johannesson, 1996; Ohinmaa and Sintonen, 1994) or colleagues (Sutherland *et al*, 1982), members of the general public (Dolan, 1996; Sackett and Torrance, 1978), and a mixture of members of the general public and patients (Hall *et al*, 1992).

There was an even wider variation in study sample sizes for studies looking at time preference rates, ranging from 10 (Pliskin *et al*, 1980) to 5120 (Cairns and van der Pol, 2000). These studies sampled members of the adult general public (Cairns and van der Pol, 1997a, 1997b, 2000; Johannesson and Johansson, 1996; Cropper *et al*, 1991; Dolan and Gudex, 1995; Cairns, 1994), mixtures of students and health professionals (Olsen, 1994), mixtures of general public and health professionals (Olsen, 1993), undergraduate students (Cairns, 1992; Chapman, 1996; Chapman and Elstein, 1995), and professional colleagues (Pliskin *et al*. 1980), workforces (Chapman and Coups, 1999). and mixtures of patient groups and students (Chapman *et al*. 1999).

Tests of mutual independence used sample sizes ranging from 64 (Miyamoto and Eraker, 1988) to 189 (Duru *et al*, 2002). These samples consisted variably of people with serious illnesses (Miyamoto and Eraker, 1988), students (Bleichrodt and Johannesson, 1996), and healthy individuals (Duru *et al*. 2002).

Studies into risk attitude were carried out in relation to health in samples of 14 (McNeil *et al*, 1978) to 40 (Mehrez and Gafni, 1987). However, Laughhunn *et al* (1980) examined the risk-seeking behaviour of 224 business executives in relation to investment risks.

Finally, the axiom of additive utility has been tested on sample sizes of 29 (Spencer, 2000) to 121 (Kupperman *et al*, 1997).

3.1.2 The results of the review

The empirical literature into the QALY axioms is reviewed in the following sub-sections. Each axiom is examined in its own sub-section.

3.1.2.1 Constant proportional trade-off of time

Constant proportional trade-off of time means that the proportion of time a person is willing to trade over any given health state is independent of the time horizon under consideration. Thus a person who is indifferent between four years in full health and five years in state x should also be indifferent between 16 years in full health and 20 years in state x. This section looks at studies examining the effects of duration and time preferences, and their implications for the constant proportional trade-off axiom.

Effect of duration

If the axiom of constant proportional trade-off holds, the value of a health state should be independent of its duration. There have been several studies examining this axiom of the QALY algorithm, and this literature review found varied results from these different studies.

The majority of studies (Kirsch and McGuire, 2000; Dolan, 1996; Ohinmaa and Sintonen, 1994; Sutherland *et al*, 1982; Sackett and Torrance, 1978) suggest that valuations of health states do depend on the duration of that health state. These studies asked participants to value health states relating to a wide variety of adverse health conditions. Each health state was valued over shorter and longer durations. The results were that participants tended to give lower values to the longer durations than the shorter durations. Thus the longer the duration of the health state, the lower its value. Sutherland *et al* (1982) referred to this effect as maximal endurable time. Most of these studies examined this issue in relation to small numbers of health states, but the study by Dolan (1996) included a total of 45 health states based on the EQ-5D health state classification system. Dolan (1996) found that duration had differing levels of effect on health state valuation, depending on the severity of the health state being valued. The more severe the health state, the more prominent was the effect of duration. Differences in ratings of health states over longer and shorter durations were found to range from 0.03 for milder health states to 0.07 for the more severe health states. As Dolan points out, these differences could be considered economically important in some contexts.

The studies by Bleichrodt and Johannesson (1996) and Hall *et al* (1992) both supported the axiom of constant proportional trade-off. However, the two studies were very different in terms of methodology. Bleichrodt and Johannesson (1996) explored the issue with respect to the valuation of health states relating to back pain and rheumatism. Hall *et al* (1992) investigated the issue with respect to the valuation of holistic health scenarios relating to breast cancer. Both studies found that, overall, valuations of health states or scenarios did not depend upon the durations or time horizons of those states or scenarios.

Bleichrodt and Johannesson (1996) introduced their concept of "personal confidence intervals", based on the premise that preferences are not preformed and complete, but are developed during the valuation process. Their study sample had no prior experience of the health states under valuation, which they were asked to value using the TTO method. Members of the sample were given the opportunity to provide personal confidence intervals, which corresponded to those values over which they could not decide whether to trade, because their preferences had not been preformed. The indifference value would be within this range of uncertainty, but it would be impossible to determine exactly where the indifference value fell. An initial analysis of individual responses, which disregarded personal confidence intervals, showed that 23% of respondents exactly satisfied constant proportional trade-off when values for the health states over different durations were compared. However, when the authors adjusted values for personal confidence intervals, the proportion of respondents who satisfied constant proportional trade-off rose to 37%. The authors further attempted to adjust responses by allocating those respondents who had not given a personal confidence interval an arbitrary range of +/- 0.075 to their health state values. This caused the proportion of respondents satisfying the axiom of constant proportional trade-off to rise to 90%. The authors' argued that the lack of complete and preformed preferences, which led to some respondents giving personal confidence intervals, was also applicable to those respondents who did provide personal confidence intervals. There were no significant differences between average values for health states over different durations. This suggests that, on average, the participants upheld constant proportional trade-off. However, on an individual basis only a little over a third of this sample did so. The assumption that respondents who had not provided personal confidence intervals could be allotted a personal confidence interval of +/- 0.075 seems questionable. The "personal confidence interval" approach has not been used in other studies to the knowledge of this author.

The TTO valuation technique has been the main method of choice for examining the effect of duration on valuation of health states and scenarios (Kirsch and McGuire, 2000; Bleichrodt and Johannesson, 1996; Hall *et al*, 1992; Sackett and Torrance, 1978), although the VAS has also been used (Dolan, 1996; Sutherland *et al*, 1982).

The different studies have varied in their use of time horizons. Valuations of health states over long-term time horizons of 10 or 30 years have been explored (Bleichrodt and Johannesson, 1996). Other studies have confined their research to time horizons of under 10 years (Kirsch and McGuire, 2000), with some studies looking at comparisons involving the very short-term, *e.g.* one month and one year (Dolan, 1996; Ohinmaa and Sintonen, 1994). Sackett and Torrance (1978) and Sutherland *et al* (1982) explored valuations over three months, eight years and a lifetime.

The weight of evidence suggests that duration does have an effect upon valuation of health states, thereby implying a violation of the axiom of constant proportional trade-off (Kirsch and McGuire, 2000; Dolan, 1996; Ohinmaa and Sintonen, 1994; Sackett and Torrance, 1978). However, there has also been some research upholding this axiom (Bleichrodt and Johannesson, 1996; Hall *et al*, 1992). Bleichrodt and Johannesson

(1996) took an interesting approach by suggesting that people may have a range of uncertainty surrounding their values of health states. They showed that there were differing levels of support to the axiom of constant proportional trade-off, depending on the assumed range of uncertainty surrounding individual values. In most valuation studies, an exact indifference value is provided by respondents, and it is assumed that their preferences are preformed and complete.

Time preferences

Another implication of the constant proportional trade-off axiom is that people should exhibit zero time preference rates. Thus individuals should be indifferent between receiving goods sooner or later. However, there are theoretical reasons for assuming a positive time preference. It is common to discount valuations for time preference, and the UK Government presently recommends a universal discount rate for costs and benefits of 3.5% (HM Treasury, 2003).

There has been a great deal of very varied empirical research into time preferences. It has not only been of interest to members of the health economics community, but there has also been extensive research reported in the psychological literature. This section will summarise the main points.

There is a large mass of empirical research showing that time preferences do not have a discount rate of zero (Cairns and van der Pol, 1997a, 1997b, 2000; Krabbe and Bonsel, 1998; Johannesson and Johansson, 1996; Dolan and Gudex, 1995; Cairns, 1992, 1994; Olsen, 1993, 1994; Cropper *et al*, 1991; Pliskin *et al*, 1980; Kirby and Hernstein, 1995; Loewenstein and Prelec, 1993). Not only have people been found to have non-zero discount rates for time, but average discount rates have been found to vary considerably between studies, ranging from –0.029 (Dolan and Gudex, 1995) to 1.24 (Chapman and Elstein, 1995).

The types of experiment have been very varied in terms of data collection techniques and the periods over which time preferences have been researched. Time preferences have been studied using TTO (Dolan and Gudex, 1995), Person Trade-Off (PTO) (Cropper *et al*, 1991; Olsen, 1994), and various forms of telephone (Johannesson and Johansson, 1996) and postal (Cairns and van der Pol, 2000, 1997a, 1997b; Cairns, 1994) surveys. The types and sizes of samples have also been very varied. For example, Pliskin *et al* (1980) reported results from a mere 10 colleagues who were asked to provide TTO values relating to anginal pain. This contrasts with a much larger study reporting extensive data on a sample of 5120 members of the UK general public in terms of time preferences relating to health for self and others (Cairns and van der Pol. 2000). The time horizons over which studies have examined time preferences have also varied, ranging from as short as three months (Chapman and Coups, 1999) to 100 years from the time of interview (Johannesson and Johannson, 1996). Time preference rates have been found to vary considerably over different time horizons (Cropper *et al*, 1991; Olsen, 1993; Johannesson and Johannson, 1996).

One possibility is that time preference may depend upon context. Chapman and Coups (1999) suggest that time preferences may even be inconsistent within individuals, depending on the length of a delay, the magnitude of outcomes, and whether they are described as gains or losses, single outcomes or sequences (Loewenstein, 1988; Benzion *et al*, 1989; Loewenstein and Thaler, 1989; Thaler, 1991; Loewenstein and Prelec, 1993; Redelmeier and Shafir, 1995; Chapman, 2000). Life stage may be another explanatory factor to differences in time preferences even within individuals (Pliskin *et al*, 1980). Chapman and Coups (1999) also suggest that individuals may be time neutral over outcomes about which they are not very bothered, but display time preferences when the outcome is important to them. For example, it has been shown that people may express a zero time preference over the choice of flu in the present or in three months time, or paying a parking fine in the present or at some future date. Yet this same sample demonstrated a negative time preference over cough symptoms set in a sequence, preferring that sequence to improve over time (Chapman and Coups, 1999). This suggests, not surprisingly, that people prefer to have an illness get better than get worse even if the two sequences would contain the same quantity of illness.

Another context in which time preferences have been found to vary is between different domains. For example, Chapman (1996) showed that her three student samples had different time preferences over health and wealth. Chapman and Elstein (1995) showed that discount rates were higher for health than for wealth. Chapman *et al* (1999) explored the possibility that these differences in time preference rates between the domains of health and wealth may reflect differences in individual familiarity with decision-making between different domains. However, upon testing this hypothesis, these authors found that time preferences for the financial domain did not correlate any closer to time preferences over familiar health states than unfamiliar health states.

Adding to the complexity is the fact that there is some evidence that time preferences for both health and wealth depend on whether the states in question are to be postponed or expedited (Cairns, 1992; Chapman and Elstein, 1995). It seems that discount rates are a function of magnitude of outcome and length of delay (Chapman and Elstein, 1995).

In the psychological literature, Loewenstein and Prelec (1993) demonstrated the complexity of time preferences. They found that individual time preferences for meals at different restaurants varied according to the context. For a single outcome, people preferred to receive the goods sooner rather than later, thus demonstrating a positive time preference. However, when a sequence of outcomes was considered, people tended to prefer a sequence that improved over time to one that deteriorated over time. This finding is in agreement with the finding reported above, that people preferred improving sequences in the health domain (Chapman and Coups, 1999). This suggests a negative time preference, preferring to receive the good later. Loewenstein and Prelec (1993) suggest that, when considering sequences of outcomes, people prefer improving sequences that have the improvements spread evenly over time.

The complexity of the time preference data so far collected has led to efforts to determine a reliable functional form for application in time preference studies. The results of a large UK study suggest that individual time preferences are best represented by a hyperbolic discounting model (Cairns and van der Pol, 1999, 2000). This study also found evidence for decreasing time aversion, implying that the discount rate is a function of the period of delay, lending support to the findings of Chapman and Elstein (1995). However, Cairns and van der Pol (2000) found no significant differences between implied discount rates when individuals were considering their own health or the health of others.

In summary, there is a mass of literature describing empirical studies indicating that people do not have a time preference of zero. Time preference rates elicited from different studies vary a great deal. The rate of time preference appears to be a function of delay and magnitude of outcome.

3.1.2.2 Mutual independence of life quality and quantity

There should be a mutual independence of life quality and quantity. This means that utility of life years and utility of quality of life should be separable and independent of each other. This ensures that the weight given to a health state during an SG exercise is independent of the duration of the state. Thus the weight can be applied to the state over any duration.

This is one of the lesser-tested axioms of the QALY model, and the empirical evidence for or against this axiom is varied. For example, Miyamoto and Eraker (1988) found that most of their sample of seriously ill patients satisfied the axiom of independence even over very short durations. Treadwell (1998) examined the axiom of mutual independence of life years and health status, which holds if preferences between profiles that contain the same health state in period i do not depend upon the severity of that health state. He concluded that independence held in 36 out of 42 tests.

However, there is also evidence refuting the mutual independence. For example, Duru *et al* (2002) tested the hypothesis that if t years in health state i $\succ$ x years in health state i then t years in health state j $\succ$ x years in health state j, where $x < t$. For mutual independence to hold, these conditions should apply whatever the health states. Duru *et al* found that only 30% of their sample upheld mutual independence.

In addition to their work on constant proportional trade-off, Bleichrodt and Johannesson (1996) explored the issue of how the SG is used to value health states over 10 and 30 years. Only 13% of their sample exactly satisfied mutual independence of quality and quantity of life. This rose to 18% when responses were adjusted for those who had given personal confidence intervals (see Section 3.1.2.1). However, a significant proportion of respondents gave higher values to health states when valued over a 10-year period than when valued over a 30-year period. It is of interest to note that only 6% of the sample satisfied both constant proportional trade-off and mutual independence simultaneously. Adjustment for personal confidence intervals for those that provided them resulted in a higher figure of 11%.

In summary, the axiom of mutual independence of utility is one of the lesser-tested QALY axioms. Whereas there is some evidence in favour of this axiom (Miyamoto and Eraker, 1988; Treadwell, 1998), the findings of other studies refute it (Duru *et al*, 2002; Bleichrodt and Johannesson, 1996).

3.1.2.3 Risk attitude

Many economic evaluations make the assumption that people are risk neutral, and therefore that SG and TTO values for the same state should be equal. Risk attitude with respect to expected survival would be linear. However, it is now accepted that risk attitude is frequently not neutral, but that people may be either risk seeking or risk averse. None the less, the QALY algorithm operates under the assumption that people should demonstrate a constant attitude to risk through time, and risk attitude should remain constant over different health states and scenarios. This section first explores the evidence relating to constancy of attitude to risk, and then goes on to explore the evidence for and against the assumption that risk attitudes are linear in form.

Constancy of attitude to risk

Even though Miyamoto and Eraker (1985) have suggested a method for adjusting valuations of health states to incorporate risk attitude, most studies do not do so, thus taking on the assumption of risk neutrality. However, there is abundant evidence that people are not necessarily risk neutral (Gaskin *et al*, 1998; Brealey and Myers, 1988; Boyd *et al*, 1982; O'Connor, 1989; Sackett and Torrance, 1978). This sub-section aims to determine if there is evidence on whether people do or do not show a constant attitude to risk.

The SG has two forms: the probability equivalence and the certainty equivalence formats. The probability equivalence format is typically used in economic evaluations employing the SG technique. However, the certainty equivalence format has been a useful tool for studying risk attitude. The certainty equivalence question involves asking for the number of years for certain that would be equivalent to a gamble involving probability p of x years and 1-p of death. McNeil *et al* (1978) and Stiggelbout *et al* (1994) both used the certainty equivalence format to examine risk attitudes.

The study by McNeil *et al* (1978) used a small sample of 14 individuals, and the majority showed risk aversity, lending overall support to the assumption that risk attitude is constant within individuals. However, a minority in this sample (two respondents) showed risk seeking behaviour over the short-term in contrast to risk aversity over the longer term. Stiggelbout *et al* (1994) explored risk attitudes among 30 disease-free testicular cancer patients, and found that risk attitude depended upon health state. Those patients who had received chemotherapy showed greater risk aversity than those who had merely been under surveillance.

The results of most studies in this area suggest that individuals are likely to exhibit non-constant attitudes to risk, and that their risk attitudes may depend upon various attributes of the situation, such as time horizon and severity of health states. Even so, there is variation in the empirical findings, with some studies demonstrating opposite results. For example, Verhoef *et al* (1994) explored risk attitudes in a sample of healthy women and found evidence that, although these women were risk averse overall, individuals showed risk seeking behaviour over gambles for short durations. However, Mehrez and Gafni (1987) found the opposite. Using a sample of undergraduate decision science students, they showed that risk-seeking behaviour was more likely to occur over increased durations, whereas over short durations risk averse behaviour was more likely to occur. Thus an individual could demonstrate both risk-seeking and risk averse behaviour depending on the duration of the state. Mehrez and Gafni (1987) suggested that the reason for the finding that risk-seeking behaviour was more likely over longer durations was that the students were more willing to take risks when the consequences would occur too far into the future to be considered at the time of making the hypothetical choice. If this is the case, this could indicate that older people may be more likely to be risk averse over longer durations, perhaps explaining the opposite findings of the studies by Verhoef *et al* (1994) and Mehrez and Gafni (1987).

In the broader field of economics, Friedman and Savage (1948) suggest that any individual may show both risk seeking and risk averse behaviour over domains for both losses and gains (see Loomes and McKenzie, 1989). In a non-health domain, Laughhunn *et al* (1980) demonstrated that business executives showed risk-seeking behaviour over gambles involving outcomes below the break-even level of investments (*i.e.* losses), but they were risk averse for gambles that concerned ruinous losses. Returning to the domain of health, Gaskin *et al* (1998) introduce the health stock risk adjustment model as an explanation of why people opt for very risky treatments when seriously ill. According to this model, an individual's risk attitude depends on their relative health stock. The lower it becomes, the more likely a normally risk averse person is to become risk seeking. Gaskin *et al* (1998) suggest that there is a critical health stock point for each individual at which their risk attitude changes.

The findings of Stiggelbout *et al* (1994) reported above led these authors to hypothesise about the possible factors underlying risk attitudes. These authors suggest that there are two factors underlying risk aversion. Firstly, Stiggelbout *et al* suggest that people have a decreasing marginal utility of time, such that some people place a higher utility on the immediate future than on the distant future (positive time preference). This is related to the concept of quantity effect or strength of preference described by Dyer and Sarin (1982). Dyer and Sarin suggest that when risk attitude is measured, *e.g.* by use of the certainty equivalent methods described in the Stiggelbout *et al* study, the result is risk attitude relative to the effect of diminishing marginal utility of time. They suggest that, when an S-shaped utility function is observed over risky choices, it is the measurable value function that changes while the relative risk attitude remains the same. Dyer and Sarin define a measurable value function as a "preference function that may be used to order the differences in the strength of preference between pairs of alternatives". In practice, it would be difficult to separate an individual's given value into their measurable value function and their relative risk attitude. This is discussed further in Chapter 8.

The second factor suggested by Stiggelbout *et al* (1994) to underlie risk aversion is an aversion to gambling with life. This is one explanation for SG values being higher than TTO values (Stiggelbout *et al*, 1994). The health state being valued is given a higher value in order to reduce the risk of death. The ideas of Stiggelbout *et al* (1994) are supported by Loomes (1988) and Nord (1992), who suggest that the gaps between SG and TTO valuations are larger for longer time horizons. See also Bleichrodt (2002).

Another complexity of the issue of individual attitudes to risk is the way in which risk is incorporated into valuations. The application of the standard gamble method in the QALY model takes the *ex post* perspective. In other words, risks presented are based upon the percentage of success. It does not take into account the *ex ante* risks attached to entering or leaving a health state, nor all the risks intrinsic to medical decision-making. Yet there is evidence that such risks affect the preferences of individuals. For example, Cook *et al* (1994) found that, when a risk of 1 in 1000 of death during an operation to remove gallstones was taken into account in an *ex ante* QALY valuation, the utility of respondents tended to differ significantly from the *ex post* valuation. Since most medical decision-making takes place from the *ex ante* perspective, these findings are relevant to the use of QALYs (which mostly adopt the *ex post* perspective) and HYEs (which may adopt an *ex ante* or an *ex post* perspective).

In summary, the majority of empirical research has found evidence that individuals' risk attitudes are non-constant. The evidence is mixed, with some studies suggesting that risk aversity occurs over shorter durations while people are more likely to be risk seeking over longer durations, and other studies suggesting the opposite. It is possible that background characteristics such as age may affect risk attitude.

Linearity of attitude to risk

More generally, it is assumed that people should demonstrate linearity with regard to probabilities. Thus the difference in utility between probabilities of 0.4 and 0.5 should be the same as the difference in utility between probabilities of 0.8 and 0.9.

Allais (1953, 1979) conducted a series of experiments with lotteries concerning money, which appear to refute the axiom of constancy of risk attitude (Machina, 1987). Allais used the following four lotteries, *a1* to *a4*, to demonstrate non-linearity of attitude to risk (from Rabin, 1997):

	a1	*a2*	*a3*	*a4*
Prob $0	0.00	0.01	0.90	0.89
Prob $1 million	1.00	0.89	0.00	0.11
Prob $5 million	0.00	0.10	0.10	0.00

If attitude to risk is linear, people who choose *a1* over *a2* should also choose *a4* over *a3*. By studying the lotteries, it can be seen that the trade-off between *a1* and *a2* represents exactly the same trade-off as that between *a4* and *a3*. The probability of gaining $5 million is the same, so ought not to affect the gambling choice if the axiom of independence holds (see Section 3.1.2.2). Thus the subject should only consider the probabilities of gaining $0 or $1 million. Lottery *a2* has a 0.01 increase from *a1* in the probability of gaining $0, and a 0.11 decrease in the probability of gaining $1 million. Likewise, lottery *a3* has a 0.01 increase from *a4* in the probability of gaining $0, and a 0.11 decrease in the probability of gaining $1 million. However, Allais' experiments showed that people who chose *a1* over *a2* were more likely to choose *a3* over *a4*. This phenomenon is known as Allais' paradox.

Allais' paradox indicates that risk attitude does not lie on a linear scale as formulated by the QALY model. It appears that people are very reluctant to move from *no probability of loss* to even a small probability of loss. This is known as the "certainty effect"

(Schoemaker, 1982). It means that for a small increase in a low probability of loss, people demand a large increase in the probability of gain (Machina, 1982; Dekel, 1986; Stalmeier and Bezembinder, 1999). Thus the changes at the extreme ends of the probability scale will have a greater influence on utility than changes of equal magnitude in the central regions of the probability scale.

This was borne out by the results of a survey by Viscusi *et al* (1987), which indicated that people were willing to pay $1.04 extra for a reduction from 15/10,000 to 10/10,000 harmful toxic reactions in a pesticide, whereas they were willing to pay $2.41 extra for a reduction from 5/10,000 to 0/10,000. The implication of these results is that a perceived total elimination of risk is worth more than a same magnitude risk reduction, thus supporting the work of Allais and refuting the assumption that risk attitude is linear.

In summary, there is evidence that attitude to probability is non-linear. This has been shown in the field of finance by Allais (1953, 1979), and also in the health domain by Viscusi *et al* (1987). The empirical evidence indicates that changes at the extremes of the scale have a greater impact on utility than changes elsewhere on the scale.

3.1.2.4 Additive separaibility of utility over time

The additive separability of utility assumption means that a health state may be seen as totally independent from any succeeding or preceding health states. Thus a health profile may be valued by multiplying the quality weights of each composite health state by the duration of each state, and summing these products over the health profile. One implication of this assumption is that two profiles, each consisting of the same set of states over the same durations, should be valued equally regardless of whether the profile contains an improving sequence or a declining sequence. Another implication of the additive assumption is that states lasting only a short duration should only have an impact proportionate to their duration upon the value of the profile.

The results from the health-related research literature are varied, with some studies supporting the axiom of additive utility (MacKeigan *et al*, 1999; Cook *et al*, 1994), and other studies offering evidence to refute this axiom (Hall *et al*, 1992; Kupperman *et al*, 1997). However, to add to the complexity, some studies have provided mixed results, with some support for this axiom and some evidence refuting it (Krabbe and Bonsel, 1998; Spencer 2000, 2003).

The studies have been varied in nature. Some researchers explored the issue in the context of holistic valuations of health profiles involving sequences. For example, MacKeigan *et al* (1999) performed a comparison of QALY values for health profiles that gradually deteriorated over 30 years with holistic valuations of those profiles. These authors found no significant differences between the two methods of valuing the profiles. On the other hand, Hall *et al* (1992) found that prognosis affected valuations. Specifically, health profiles ending in death from cancer were given lower values than those ending in non-cancer-related deaths. The findings of Hall *et al* (1992) were supported by a later study by Richardson *et al* (1996), who constructed a 16-year health profile consisting of a moderate health state for five years followed by a good health state for 10 years followed by a bad state for one year followed by death. These authors compared the holistic valuations of this profile with QALY valuations constructed from the composite health states, and found that QALY values were significantly higher than holistic values for the majority of the sample. These results do not explicitly refute additive utility, because the authors were unable to prove that the holistic valuations were a truer reflection of preferences than the QALY values. However, the fact that the two methods gave different results, and that the QALY values were higher, suggests that the QALY underestimates the effect of sequence on preferences. Richardson *et al* (1996) suggested that dread may have played a part in the lower values given to holistic valuations.

Kupperman *et al* (1997) followed the route of comparing holistic valuations of health profiles with QALY valuations using composite health states, but applied an econometric analysis to the valuations obtained. The authors concluded that individuals' preferences over health profiles were not accurately reflected by the traditional fixed coefficient model usually applied by the QALY, in which each discrete health state is weighted according to its duration. A multiple regression model was applied, which maintained the linear additive attribute but did not weight each health state by its duration. This model allowed examination of the relative contribution of each health state to the absolute score of the profile to be examined. This model was found to fit the data better than the traditional QALY model. Thus the evidence from this econometric analysis is that preferences do not follow a simple additive model.

Prior to the research using holistic health profiles, Cook *et al* (1994) tested the axiom of additive utility by asking respondents to value short-term treatment health profiles over 12 months and 12 years in order to investigate whether the different time horizons would affect valuations. No significant differences in values were found across the different time horizons.

Another method of examining the axiom of additive utility was employed by Spencer (2000, 2003), who constructed health profiles consisting of three distinct time periods. The health states within the profiles may be abbreviated to "good", "medium", and "bad", and each time period contained just one health state. Profiles good-good-good and good-medium-bad were compared, and profiles bad-good-good and bad-medium-bad were compared. According to additive utility, the differences between the profiles in the two comparisons should be equal. However, there was a greater difference between the profiles in the latter comparison than in the former comparison for 21/29 respondents. This finding disputes the additive utility axiom, and provides some evidence for sequence effects. The results from similar tests in which other parts of the profile were varied were inconclusive, and Spencer was unable to conclude that her study categorically refuted additive utility. However, these results do cast doubt upon the axiom of additive utility.

There is evidence that the additive utility axiom is upheld by some individuals, whereas other individuals are affected by sequences. For example, Krabbe and Bonsel (1998) showed that, although the majority of their sample of 104 were unaffected by sequence, a minority of sequence-affected individuals fell into two categories: those who prefer better health earlier, and the "happy ending" group who prefer the better state to occur at the end.

In addition to the research that has been carried out in health economics, there has been a great deal of research in the psychological literature. Much of the psychological research has little direct connection with health, and hence has been under-used by health economists. However, the psychological literature provides a great wealth of empirical evidence that can be related to the extent to which the QALY axioms are violated.

The work of Lowenstein and Prelec (1993) on preferences over sequences has already been mentioned in Section 3.1.2.1 in relation to time preferences. Research in the financial domain has supported the findings that many people prefer sequences that improve over time. For example, Loewenstein & Sicherman (1991) found a preference for wage profiles that increased over time to other possible profiles where it would decrease or remain level, even though the total amount earned would be the same. This finding was confirmed by Ross and Simonson (1991) in an investigation of attitudes to financial gains and losses, who found a preference for a sequence involving a loss of $15 followed by a win of $85 over the reverse sequence.

There is also evidence that individual expectations may affect preferences over sequences. For example, although improving sequences may be preferred for sequences involving changes in athletic ability or facial acne, the reverse may be true for sequences relating to facial wrinkles (Chapman, 2000). Chapman suggested that preferences for improving sequences may be tempered by beliefs about what is realistically likely.

There have been several experiments exploring preferences over sequences of adverse events such as pain or unpleasant noise. The results from these experiments confirm a preference for improving sequences (Ariely and Zauberman, 2000; Ariely, 1998, Hsee and Abelson, 1991; Fredrickson and Kahneman, 1993; Kahneman *et al*, 1993; Redelmeier and Kahneman, 1993; Ariely and Carmon, 2000; Hsee and Abelson, 1991; Hsee *et al*, 1991). However, these experiments provided additional insights into the way people value sequences. There is evidence of a phenomenon known as duration neglect when valuing sequences. For example, Ariely (1998) showed that for sequences involving pain of constant intensity, valuations did not depend on the duration of the sequence. However, when there were changes in pain intensity over the sequence, the duration of pain at each level of intensity had a greater effect on overall valuations of the sequence. Ariely suggested that this may be due to adaptation, such that people adapt to a single intensity of pain, but when changes in intensity occur the lack of adaptation causes them to be more aware of duration.

Experiments such as these in the psychological literature have shown that certain characteristics of sequences have a particularly important effect on the overall valuations of those sequences. These characteristics were the rate at which the intensity of the stimulus changed, the directional nature of the sequence (*i.e.* whether it improved or deteriorated), and the level of stimulus towards the end of the sequences. For example, Varey and Kahneman (1992) found that trend of sequence was an important factor in overall valuations, with sequences showing more rapid increases in pain being given lower values to sequences with a more gradual increase in pain. This was to such a degree that a sequence could be significantly greater in duration, but if it was a great deal more gradual in its increases in pain it would not be valued much lower than a shorter duration steeper pain increase sequence. Hsee and Abelson (1991) also found that people preferred higher velocity improving sequences (see also Hsee *et al*, 1991).

This research reported in the psychological literature does not relate specifically to the domain of health, but is nonetheless relevant to the QALY model and in particular the axiom of additive utility. It goes to show that, when it comes to adverse stimuli such as pain or unpleasant noises, there is a tendency for people to value sequences in a manner which is not linearly additive as is supposed by the axiom of additive utility. As it has been shown that overall valuation of a sequence is affected by trends and patterns within the sequence when those sequences relate to non-health domains, it is reasonable to suppose that those same factors may influence judgements over sequences of health states.

In summary, the findings of some health economics studies have supported additive utility, whereas other studies in health economics have provided evidence against the axiom of additive utility. However, the psychological literature provides very little support for this axiom. According to a large body of psychological literature, patterns and trends within a sequence affect overall valuations of those sequences, with improving sequences being preferred to declining sequences. There is also evidence of duration neglect. Although this psychological research does not relate specifically to health states, it is reasonable to suppose that the same decision-making and judgement processes that are used by people in these situations might be used in a health setting also.

3.1.2.5 Conclusions

A total of 69 studies have been listed in this review. Out of this number, 7 (10.3%) upheld or seemed to support one or more of the axioms of EUT and the QALY algorithm. However, 66 (95.6%) provide evidence of violations of one or more of these axioms.

There has been relatively little empirical research in the health economics literature demonstrating the extent to which people violate the QALY axioms. However, there has been an abundance of research in the psychological literature that is of relevance, but has largely been ignored by health economists. Much of this psychological literature does not relate directly to health states and profiles. Yet this work provides evidence of how people deal with sequences, and suggests that violations underlying the QALY axioms may be widespread in preferences over non-health outcomes.

Previous research has demonstrated that there is a significant degree of publication bias in health care. A study by Easterbrook *et al* (1991) indicated that there is a greater likelihood of publication if a study reports statistically significant results. Callaham *et al* (1998) demonstrated that studies which had been published in journals of highest prestige were more likely to be cited than those published in less prestigious journals, which would lead to their greater dissemination within the research community.

It is impossible to state the degree to which this review has been subject to publication bias. However, it is worth taking note of the issue, and it cannot be assumed that all the relevant research relating to the axioms underlying the QALY axiom has been captured by this review. For example, some of the literature was obtained by bibliographic searches, which would be prone to publication bias.

Research is required to bridge the gap between the psychological literature and the field of health economics. It is important to thoroughly research the extent to which the QALY axioms are violated in the context of judgements over sequences of stimuli or states as this is of relevance to valuing health profiles.

3.2 A review of the empirical research into holistic valuations of health profiles

This section reviews the empirical research into what has already been discovered in the area of holistic valuations of health profiles. Of particular interest are methods of construction of health profiles, and methods used to obtain holistic valuations. Although the original method suggested for valuing HYEs was the two-stage procedure described by Mehrez and Gafni (1989), this review aims to examine empirical research into the holistic valuations of health profiles more generally. Studies reporting two-stage gamble methods will be included, as also will studies reporting other valuation techniques such as TTO or single-stage standard gambles.

This literature review consisted of three parts:

1) A review of empirical literature in ATHENS databases.

2) A search of the ATHENS databases and the bibliographies of articles obtained from (1) for the names of researchers in the field of holistic profiles valuation research.

3) A bibliographical search for empirical literature in non-health economics sources from the bibliographies of (1) and (2).

The following search terms were used (in brackets are the numbers of papers found):

- HYE (29)
- HYEs (23)
- Healthy year equivalent (2)
- Healthy years equivalent (14)
- Healthy year equivalents (6)
- Healthy years equivalents (15)
- holistic valuation (0)
- holistic valuations (0)
- holistic value (0)
- holistic values (3)

Results of the search

Databases were searched from 1966 to 4 June 2005. A total of 51 papers were found using the above search terms. A total of 34 of these papers related to HYEs, the majority of which were theoretical rather than empirical.

The literature review of empirical studies relating to QALYs (Section 3.1) included papers that were relevant to the HYE, and the bibliography search from Section 3.1 also produced relevant HYE papers. This present review includes these papers plus papers from other sources such as personal communications and journal searches.

A total of seven papers describing empirical research into HYEs were detected.

Sample sizes

Only seven studies involving the elicitation of holistic valuations of health profiles were detected, and the sample sizes in these studies ranged from 60 (Sculpher, 1996) to 194 (Llewellyn-Thomas, 2002). The sample sizes per study are listed below:

- 104 (Hall *et al*, 1992)
- 96 (Cook *et al*, 1994)
- 63 (Richardson *et al*, 1996)
- 60 (Sculpher, 1996)
- 121 (Kuppermann *et al*, 1997)
- 101 (MacKeigan *et al*, 1999)
- 194 (Llewellyn-Thomas *et al*, 2002)

3.2.1 Techniques used in the valuation of holistic health profiles

Only one study has been conducted using the two-stage gamble elicitation procedure suggested by Mehrez and Gafni (1991). Llewellyn-Thomas *et al* (2002) set out to test the feasibility of this approach in the context of eliciting HYEs from patients who had undergone major joint replacements. They obtained HYEs for personalised health profiles consisting of a pre-surgery state, surgery, and post-surgery remainder of life state. These authors found that this procedure was feasible and comprehensible. However, they reported a sample split into two types: HYE-invariant and HYE-variant. The HYE-invariant group refused to gamble, and comprised 43% of the sample.

Five studies used some kind of TTO formulation to elicit holistic valuations (Hall *et al*, 1992; MacKeigan *et al*, 1999; Cook *et al*, 1994; Richardson *et al*, 1996; Sculpher, 1996). Reid (1998) refers to the elicitation of health profile values by TTO as the "generalised" TTO to differentiate it from the TTO-QALY, which is used to elicit values for discrete health states. There are two possible postures of the generalised TTO (Johannesson, 1995b): the 'expected HYEs' and '*ex ante* HYEs'. According to Johannesson (1995b), the expected HYE approach uses a single-stage SG to value each health profile. The expected HYE of a treatment is then calculated by combining these values with the risk of occurrence (i.e. 'expected utility' of the risky treatment is the utility of the health profiles multiplied by its probability of them occurring). The expected HYE takes the *ex post* perspective. The two-stage gamble method as formulated by Mehrez and Gafni (1991) produces expected HYEs, because *ex ante* risks are not incorporated into the profiles. Likewise, the single-stage TTO can also be used in this way. The *ex ante* HYE approach asks respondents to value health profiles which include a description of the risks associated with each stage of the health profile. Johannesson suggests that a single-stage generalised TTO could be used to elicit *ex ante* HYEs.

Previous research has demonstrated the use of the single-stage generalised TTO in eliciting "expected" or *ex post* holistic values of profiles. The study by Hall *et al* (1992) was among the first to value health profiles holistically, and used the generalised TTO approach to value hypothetical health profiles relating to breast cancer. This was followed by further research into breast cancer health profiles using this approach to valuing health profiles holistically (Richardson *et al*, 1996). MacKeigan *et al* (1999) used this approach to elicit holistic valuations of health profiles relating to type II diabetes.

There has only been one study utilising an *ex ante* TTO-based approach to valuing holistic profiles. This was a study by Sculpher (1996) to elicit holistic values for health profiles relating to treatment pathways for menorrhagia. Each profile contained differing levels of risk. Prior to the study by Sculpher, Cook *et al* (1994) used a partial *ex ante* TTO approach to elicit values for a risky treatment for gallstone disease. One of the potential problems with the QALY approach is the way with which risk is dealt. The QALY traditionally takes an *ex post* perspective, in which the risks attached to an outcome are multiplied by the value of the outcome in a decision tree. However, there are concerns that people may have subjective values attached to the actual risks of entering or leaving a health state. Cook *et al* (1994) did not value holistic health profiles *per se*, but incorporated a small risk of mortality into their treatment health state. This was a risk of operative death of 1 in 1000. This tiny risk carried very little weight in the *ex post* QALY calculations, but was given greater weight by respondents in the *ex ante* approach, and was in fact given a weight similar to the adverse health states themselves.

In addition to using the generalised TTO approach, Richardson *et al* (1996) also obtained holistic values by a single-stage SG and a rating scale approach. Kuppermann *et al* (1997) also used SG and VAS to obtain holistic valuations for health profiles. These authors were interested in choices between screening for Down's syndrome in unborn babies by either amniocentesis or corionic villus sampling.

3.2.2 Construction of scenarios

Another issue of importance to valuation studies is the way in which health state and health profile descriptions are constructed. One of the aims of this thesis is to examine the methodology for constructing condition-specific health profiles. A combination of sources for descriptive health profiles has been used, *e.g.* literature reviews (Hall *et al*, 1992; Sculpher, 1996), information from health professionals (Hall *et al*, 1992; Richardson *et al*, 1996), and patients themselves (Hall *et al*, 1992; MacKeigan *et al*, 1999; Llewellyn-Thomas *et al*, 2002; Richardson *et al*, 1996; Sculpher, 1996).

3.2.3 Comparisons of QALYs and holistic valuations

Sculpher (1996) discusses a trade-off between the accessibility and flexibility of "off the shelf" QALYs with their restrictive assumptions, and the less flexible HYEs that may be stronger theoretically. At present, generic HRQoL data (*e.g.* EQ-5D) may be collected during clinical trials alongside medical data. This can be entered into the QALY algorithm to produce QALY values associated with any health profile over the course of the clinical trial. However, if holistic valuations of experienced profiles take the place of QALYs in CUAs, the construction and valuation of health profiles will be able to take place only after completion of the clinical trial, because it will only be at the end of the trial that the complete health profiles associated with the courses of treatment are established (Sculpher and Barbieri, 2001). Whereas preferences for current health indicated by the EQ-5D may be collected while patients are actually in that health state, the construction of holistic health profiles subsequent to the clinical trial will result in asking respondents to value health profiles in retrospective. This allows values to be affected by cognitive limitations, such as being unable to imagine a hypothetical health profile, or being unable to remember what HRQoL was while undergoing a profile previously experienced.

As Sculpher (1996) points out, one important area for research is the issue of whether QALYs and holistic valuations give different results. If the results from these two methods do not differ, it would seem reasonable to continue using the QALY rather than opt for holistic valuations, which have their own limitations.

There has been a small amount of empirical research comparing the results of health profile valuations from QALY and holistic methods. The results of these studies have varied with the type of health profile. Richardson *et al* (1996) found that holistic values tended to be lower than QALY values for breast cancer health profiles in which HRQoL deteriorated over a 16-year time horizon. This fits in with the work of Loewenstein and Prelec (1993), who researched preferences over sequences of non-health goods (see Section 3.1.2.1) and found that people tended to prefer sequences that improved over time to sequences that got worse over time. However, MacKeigan *et al* (1999) compared QALY and holistic values for health profiles relating to type II diabetes. These profiles deteriorated very gradually over a 30-year time horizon. These authors found no significant differences between the two methods in terms of health profile values. This is in contrast to the findings of Richardson *et al* (1996), but the profiles in the study by MacKeigan *et al* deteriorated in a far more gradual fashion than those in the Richardson *et al* study, in which the deterioration was more marked.

Other research has focussed on differences between the two methods in relation to the way in which risk is incorporated into health profiles. As discussed above, risk may be incorporated either *ex post* or *ex ante*. Sculpher (1996) obtained values for health states using the *ex post* TTO-QALY technique, and the *ex ante* TTO-HYE to obtain holistic values for health profiles. Cost per benefit was found to be lower by holistic valuations than QALYs, but the direction of preferences between the two treatment pathways was the same between the two methods. However, holistic valuations were found to have a greater consistency when compared to a previous question on treatment choices elicited from his sample of respondents (see also Sculpher and Barbieri, 2001).

Finally, Kuppermann *et al* (1997) elicited QALY and holistic values for health profiles, and found the results to differ. These authors looked at the predictive properties of discrete health state values in terms of correlations with holistic profile values. Certain health states were more predictive of profile values than other states. A higher correlation between health state and holistic values was demonstrated by VAS scores than SG scores. Kuppermann *et al* concluded that, although there was a relationship between QALYs and holistic valuations, this relationship was not linearly related to duration in each health state as would be predicted by the QALY algorithm. They developed a regression model which applied weights to the means of the discrete health states, which would then be predictive of holistic profile mean values. This would be for use in population studies for which means were used. It would not be applicable to individual data.

3.2.4 Further research

In conclusion, there has been very little empirical research into holistic valuations of health profiles. Although methods of valuing profiles holistically have progressed little since the idea of the two-stage gamble procedure for eliciting HYEs was first introduced by Mehrez and Gafni (1991), the notion of holistic valuation has been an important contribution to the health economics literature.

This thesis will explore several aspects relating to methodology of holistic profile valuation. One aspect of research will be the use of different valuation techniques, such as single stage SG and TTO. Ways in which *ex ante* risk can be incorporated into health profiles will also be examined. One area which has received no research as yet is how short-term terminal health profiles should be valued holistically, and this will also be explored. Another area that will be researched in this thesis is the construction and valuation of health profiles that vary in an unpredictable way over time. This thesis will also investigate ways to construct and value health profiles relating to different types of health conditions. Holistic valuations of health profiles will also be compared to QALY values for the profiles, and any differences between the two methods in terms of results will be investigated.

Table 3.1 Results of the literature review.

Do people obey the QALY axioms?	**Yes**	**No**
Constant proportional trade-off		
Duration	Hall *et al* (1992)[1]	Sackett and Torrance (1978)[2] McNeil *et al* (1981)[2] Sutherland *et al* (1982)[2] Ohinmaa and Sintonen (1994)[2] Bleichrodt and Johannesson (1996)[1] Dolan (1996)[3] Kirsch and McGuire (2000)[3]
Zero time preference	Chapman and Coups (1999)[2]	Pliskin *et al* (1980)[2] Horowitz and Carson (1990)[2] Cropper *et al* (1991)[2] Cropper *et al* (1992)[2] Cairns (1992)[2] Loewenstein and Prelec (1993)[2] MacKeigan *et al* (1993)[2] Olsen (1993)[2] Redelmeier and Heller (1993)[2] Cairns (1994)[2] Olsen (1994)[2] Chapman and Elstein (1995)[2] Dolan and Gudex (1995)[3] Kirby and Herrnstein (1995)[2] Chapman (1996)[2] Johannesson and Johansson (1996)[2] Richardson *et al* (1996)[3] Cairns and van der Pol (1997a)[2]

[1] This paper was obtained via a personal communication.

		Cairns and van der Pol (1997b)[2] Johannesson and Johansson (1997a)[2] Johannesson and Johansson (1997b)[2] Krabbe and Bonsel (1998)[3] Cairns and van der Pol (1999)[4] Chapman *et al* (1999)[2] Cairns and van der Pol (2000)[3]
Mutual independence of life quality and quantity	Miyamoto and Eraker (1988)[2] Treadwell (1998) [1]	Bleichrodt and Johannesson (1996)[1] Duru *et al* (2002)[5]
Constancy of risk attitude to survival duration	Miyamoto and Eraker (1985)[2]	Friedman and Savage (1948)[2] Mehrez and Gafni (1987)[2] Cook *et al* (1994)[3] Stigglebout *et al* (1994)[2] Verhoef *et al* (1994)[2] Gaskin *et al* (1998)[2]
Risk neutrality under all health states		McNeil *et al* (1978)[2] Sackett and Torrance (1978)[2] Laughhun *et al* (1980)[2] Boyd *et al* (1982)[2] Miyamoto and Eraker (1985)[2] Brealey and Myers (1988)[2] O'Connor (1989)[2]
Linearity of risk attitude		Allais (1953, 1979)[2] Dekel (1986)[2] Viscusi *et al* (1987)[2]

[2] This paper was obtained from bibliographical sources.

[3] This paper was obtained from Medline.

[4] This paper was obtained via a journal search.

[5] This paper was obtained via a search using the internet search engine Google.

		Stalmeier and Bezembinder (1999)[2]
Additive utility function		Kupperman *et al* (1997)[2] Ariely (1998)[2]
Valuations of health state independent of succeeding/preceding health states	MacKeigan *et al* (1999)[3] Spencer (2000)[1]	Hsee and Abelson (1991)[2] Hsee *et al* (1991)[2] Loewenstein and Sicherman (1991)[2] Ross and Simonson (1991)[2] Hall *et al* (1992)[1] Varey and Kahneman (1992)[2] Fredrickson and Kahneman (1993)[2] Kahneman *et al* (1993)[2] Loewenstein and Prelec (1993)[2] Redelmeier and Kahneman (1993)[2] Richardson *et al* (1996)[3] Krabbe and Bonsel (1998)[3] Chapman and Coups (1999)[2] Ariely and Carmon (2000)[2] Ariely and Zauberman (2000)[2] Chapman (2000)[6]
Small duration effects are negligible to the patient	Cook *et al* (1994)[3]	

[6] This paper was obtained via a search for the author's name.

Chapter 4

An overview of the empirical questions addressed in the four studies and the research methods used

One of the aims of this thesis is to test the assumptions which underlie the QALY algorithm, and to determine the extent to which these assumptions are violated. This chapter describes the assumptions being addressed in this thesis, and outlines why these assumptions were selected for testing. The chapter then goes on to describe how these axioms are tested in each of the four studies in Chapters 5 to 8 of this thesis.

Issues relating to the construction of condition-specific health profiles and scenarios are also discussed.

4.1 Overview of studies

Three illness conditions were chosen for the bases of the studies described in this thesis. These were irritable bowel syndrome (IBS), varicose veins, and abdominal aortic aneurysms (AAAs). Each condition was chosen for its particular characteristics.

IBS is usually considered to be at the mild end of the utility scale. However, there is evidence that sufferers find that it has a significant reduction on their quality of life (see Chapters 5 and 6). It has the characteristic that its symptoms may vary considerably in any given time period, from perhaps being very mild or absent to being very severe. The frequency and severity of symptoms can be very unpredictable. This illness provides the opportunity to study the axiom of additive utility in terms of scenarios in which the defining characteristic is the proportion of time spent in each health state rather than the sequence of health states.

Like IBS, the condition of varicose veins is usually considered to be comparatively mild. Sufferers endure reductions in quality of life to varying degrees, but rarely does the condition of varicose veins lead to a shortening of life expectancy. Thus this condition facilitates the comparison of different methods of valuation of health profiles in which HRQoL varies but life expectancy does not. The main focus of this study is a comparison of the treatments of sclerotherapy and surgery for varicose veins, in particular with respect to process, HRQoL, and risk.

AAAs are of interest because sufferers may expect a reduction in life expectancy without a reduction in HRQoL, because the condition is often asymptomatic (see

Chapter 8 for a more complete discussion of the condition). Once their AAA is discovered, patients may face a choice. One option would be to undergo a process of treatment (endovascular repair) which may lengthen their life, but has a relatively high chance of serious immediate morbidity or mortality. The other option may be to forgo this treatment and opt for the next best watchful waiting alternative. in which case the life expectancy would not be lengthened but the immediate risks of surgery would be avoided. This condition facilitates a study of the way in which people balance their wish to maximise their life expectancy against serious *ex ante* risks in the short-term.

4.2 QALY axioms under test

The first column of Table 4.1 lists the assumptions of the QALY model, and the next four columns indicate which assumption was tested in each study. It would have been unfeasible to test all the axioms of the QALY model within the time and resources available. A decision was made to test the axioms relating to constant proportional trade-off, risk attitude, and additive separaibility. The reasons for these choices are discussed in the following sub-sections.

4.2.1 Zero time preference

As shown in Table 3.1, much of the previous research over the past two decades has refuted the axiom of zero time preferences (*e.g.* Pliskin *et al*, 1980; Cairns, 1994; Chapman, 1996). If a zero time preference cannot be assumed, a comprehensive exploration of how and whether QALY values should be adjusted for time preferences is required.

Cairns and van der Pol (2000) reported results of measuring time preferences over a life time. Despite the zero time preference axiom, there are theoretical arguments for a positive rate of time preference. The findings of Cairns and van der Pol (2000) supported the hypothesis of positive time preferences.

In Chapter 8 of this thesis the Cairns and van der Pol method of eliciting time preferences was adapted to explore time preferences in the context of very short life expectancies of three years. Previous studies have focussed on rather longer time horizons. However, short time horizons are relevant for many profiles relating to conditions for which life expectancy is short. The study described in Chapter 8 elicits valuations for scenarios relating to large AAAs in unfit patients. for which a three-year life expectancy is typical (see Chapter 8). Time preferences are elicited, and there is an attempt to adjust QALY valuations for individual time preference.

4.2.2 Risk attitude

Most of the previous research has also refuted the axiom of risk neutrality (*e.g.* McNeil *et al*, 1978; Stiggelbout *et al*, 1994), and independence and constancy of risk attitude to survival duration (*e.g.* Mehrez and Gafni, 1987; Verhoef *et al*, 1994). McNeil *et al* (1978) used a certainty equivalence method to elicit risk attitude from individuals, and Miyamoto and Eraker (1985) described a method of adjusting QALY values for individual risk attitude.

McNeil *et al* (1978) elicited risk attitudes over expected survivals of 10 or 25 years. They devised a method by which respondents always valued a 50:50 gamble, to make the procedure as easy as possible for respondents (Figure 4.1). Respondents were first asked to state the certain number of years that were equivalent to a 50:50 gamble of 25 years (or 10 years, depending on age) in full health and immediate death. This certainty equivalent value was called *a* years. Respondents then stated the certain number of years *b* in full health that were equivalent to a 50:50 gamble of *a* years in full health or immediate death. Finally, respondents stated the number of certain years *c* in full health that were equivalent to a 50:50 gamble of 25 years in full health or *a* years in good health. The value of 25 years in full health was assigned a utility of 1. The utility value of *a* years in full health was assumed to be 0.5, of *b* years the utility would be 0.25, and of *c* years the utility would be 0.75. If expected survival was 25 years, a risk neutral individual would give a value of 12.5 years for *a*, 6.25 years for *b*, and 18.75 for *c*.

In Chapter 8 risk attitude and the effects of *ex ante* risks were examined in the context of scenarios relating to AAAs. In the AAA scenarios, the expected life expectancy was either two or three years, and there were substantial risks of immediate mortality and serious morbidity that would last for the remainder of life in the three-year scenario. An attempt was made to measure risk attitude using a method similar to that used by McNeil *et al*. However, rather than adopting the 50:50 gamble for all the elicitations, the utility values were elicited directly. Thus respondents were asked directly for their certainty equivalents of a 25:75 gamble of current health and death, a 50:50 gamble, and a 75:25 gamble. The risk attitude values obtained from this elicitation process were used to adjust QALY valuations and determine the affect of risk attitude on preferences over the AAA scenarios.

The AAA scenarios of Chapter 8 contain descriptions of risk. The affect of including descriptions of risk in health profiles is also explored in Chapter 7, which describes a study in which patients with varicose veins value varicose veins health profiles. Each profile consists of a pre-treatment state (*e.g.* moderate varicose veins) lasting six months, followed by a treatment process, followed by a state that would last the rest of their life. The final state might be an improvement or remain the same as the pre-treatment state. Composite health states are used to construct QALY values for the profiles, and patients also value the profiles holistically. Two of the profiles contain descriptions of mortality and recurrence risks related to the treatment processes. Valuations of these risky profiles are compared to valuations of profiles that are equal except for the inclusion of risk. This study seeks to explore how perception of risks alters the valuations given to health profiles.

4.2.3 Small duration effects

One implication of the additive utility axiom is that experiences of short duration, however intensive, will be weighted according to the product of the utility of the experiences and its duration. This is relevant if preferences are being sought over different treatment processes, for which patients may have strong preferences regardless of the small weight that may be given to the treatment by the QALY algorithm because of short duration.

The issue of process utility was examined in Chapter 7. Process may be seen as a non-health outcome, because it deals with utility derived from the process of care rather than changes in health. As such, it would be given no weight by the traditional QALY model. However, it has also been argued that process affects health-related aspects such as pain and anxiety, and as such it could be considered to be a part of health utility (Donaldson and Shackley, 1997). The treatment processes of sclerotherapy and surgery for varicose veins were valued using a method for valuing short-term temporary health states. Process was therefore treated as a health outcome in that respondents were asked to state how they believed the hypothetical treatment process used would affect their HRQoL. Varicose veins patients may have strong preferences about what type of treatment they wish to receive prior to consultation. QALY valuations of the health profiles were compared to holistic valuations in order to determine how valuations were affected by treatment process. It may be that the treatment process is given greater weight by patients than is suggested by the QALY algorithm.

4.2.4 Valuations should be proportional to time in each health state

It is notable that, of the large number of studies into aspects of the additive utility function shown in Table 3.1, the majority are from the psychological literature. Indeed many were not even specifically related to health or health improvements. It is important to do comprehensive research into decision-making behaviour. and explore the extent to which the QALY axioms are violated in a health setting.

An implication of the additive utility axiom is that valuations of a profile should be proportionate to time in each state within the profile. This is put to the test in Chapters 5 and 6, in which IBS patients use SG to value health profiles relating to IBS. Each profile consists of IBS health states occurring at random over a 12-week time period. The number of weeks overall spent in each health state (*i.e.* the proportion of time) is described for each profile. However, it is made clear that the order in which the states occur would not be known in advance. The randomness of the occurrence of each health state was impressed upon respondents by stating that if *x* weeks in total was spent in one state, this total time would be randomly dispersed over the profile rather than automatically occurring in one block. Thus the assumption regarding succeeding and preceding states, which has already been widely explored elsewhere (Richardson *et al*, 1996; Chapman and Coups, 1999; Ariely and Zauberman, 2000), was not relevant in this study.

This type of profile has not been studied previously to the knowledge of this author. Yet IBS is just one of several conditions in which different health states may occur with unpredictable frequency for unpredictable and variable durations. The QALY model assumes that the value of each profile would be the sum of the products of the durations and values of each health state, thus quantifiable by the proportion of time spent in each state. However, it is possible that there could be disutility attached to the very unpredictable nature of the health states, which is not taken into account by the QALY algorithm.

4.3 Study design issues

This section outlines several issues relating to the designs of the studies described in this thesis, and the relevance of each of these issues to each of the QALY and holistic models of valuation.

4.3.1 ex ante *versus* ex post

It is possible to value profiles either *ex post* or *ex ante*. The QALY model usually adopts the *ex post* perspective (Cook *et al*, 1994). Respondents are asked to value individual health states. The values of these health states can then be incorporated into any hypothetical health profile. This is done simply by multiplying the weight for the health state by its duration. Any risks involved, for example risks of entering an adverse health state due to treatment, are simply incorporated by multiplying the value of the adverse state by its probability and duration (Cook *et al*, 1994). Respondents are not able to make a choice based on their perception of risks. Rather, risks are treated as objective probabilities.

The holistic approach to valuation can also take an *ex post* perspective. In this format, a respondent would be presented with a health profile describing a series of health states. For example, if the result of a treatment could result in four different profiles with different probabilities attached, the respondent could be asked to value each of these profiles. The resultant values could be used in a decision-tree with probabilities attached (Johannesson, 1995b). Alternatively, the holistic approach can also be used to value profiles retrospectively. For example, respondents can be presented with health profiles that have already occurred and asked to value these profiles. These profiles may either be hypothetical, have happened within a separate group of patients, or be the health profile that each of these respondents has experienced.

One of the potential advantages of the holistic approach is that a profile can be valued *ex ante*. It may be valued prospectively, in which case the individual is asked to value a profile that is meant to occur over a future time period (*e.g.* "for the rest of your life"). Risks are described and respondents are allowed to make their own judgement of perceived risks (Crocker *et al*, 1988). This is on the basis that there may be disutility associated with particular risks. The respondent may, for example, prefer a less risky scenario that has a lower life expectancy, even if a QALY valuation incorporating objective probabilities implies that the higher risk scenario is preferable. Cook *et al* (1994) found a significant disutility associated with a risk of operative mortality described as 1/1000. The varicose veins and AAA studies described in Chapters 7 and 8 respectively compare *ex ante* holistic valuations of profiles involving risks with *ex post* QALY valuations of those same profiles.

4.3.2 Whose values

There has been debate about whose values should be used in valuation studies (Daniels. 1991; Gold *et al*, 1996; Dolan, 2000). The debate centres on whether valuations should come from people who have experience of the conditions in question and hence may have a better understanding, or members of the tax-paying general public.

One argument is that it is the views and preferences of the patients that should be considered, because it is they who know what it is like to have the condition, and they who can say how it affects utility. However, there is the possibility that patient responses could be biased if the patients suspect that giving the "right" answer may increase their chances of receiving the desired treatment.

The results of cost utility analyses are used to inform decision makers of the costs and benefits of potential treatments, including quality of life. Decisions about the allocation of health care resources are based on this information. These decisions affect the whole of society, because for every intervention funded there is the opportunity cost of another intervention that could have had the funding. Thus it is not only the interests of a particular patient group at stake, but the interests of the population as a whole.

Rather than just relying on patient valuations, members of the general public could provide values for health states relating to particular conditions. However, the health state would need to be described adequately for the respondents to be able to imagine the health state for themselves and the impact it would have on their lives. If this was the case, they may be able to make a more informed decision that would be unlikely to be biased by strategic answers. However, it has been shown that people who experience ill health often give the health state higher values than the general public would. because they have adapted to the condition (Meyerowitz, 1983; Cassileth *et al*. 1984). There is a likelihood that the adaptation effect would not to be incorporated into valuations made by the general public. However, whether or not adaptation should be incorporated into valuations is itself a matter for discussion (Dolan, 1998b, 2000). As pointed out by Menzel *et al* (2002), it may not be desirable to take into account factors of adaptation that reflect a lower sense of expectation in the patient.

It is the present opinion of the author that patient perspectives should be taken. If it is worth obtaining valuations for illnesses, it is worth obtaining valuations that reflect the preferences of people in those conditions, because other people may not have so much understanding of the affects on quality of life. Also members of the society as a whole do not know to which illness they may fall prone during their lifetime. Thus the perspectives of patients could end up being anyone's perspectives. Hence it could be argued that members of the general population may be better off if patient values were used.

However, in some cases it is impossible to obtain valuations from the patients themselves. For example, the scenarios described in Chapter 8 related to risks of mortality and serious morbidity associated with treatments of large AAAs in unfit patients. It would have been considered unethical to ask people facing these risks to value them. It may have caused them distress. The next best alternative sample would have been a sample of AAA patients with small aneurysms, who would not yet have faced these risks but would have more knowledge of AAAs and their risks than the general population. As is described in more detail in Chapter 8, it turned out to be impossible to recruit AAA patients, and it was necessary to recruit a convenience sample.

As this last paragraph has touched upon, the timing of valuations can impact upon ethical issues. For example, it may have been inappropriate to ask patients with large AAAs, who were actually facing the risks involved and the potential treatments, to value the scenarios. An ethics committee faced with a proposal to subject patients to potentially distressing health state/health profile value elicitation exercises would no doubt weigh up the potential gains from conducting such a project against the distress that could be caused. If the project was part of a clinical trial to test a new drug that could have a significant impact on the subjects, such distress might be considered justifiable. However, the study in Chapter 8 was methodological in nature, and would not have a direct impact on the respondents or that patient group as a whole in terms of potential improvements in health. In this case it may not have been considered ethical to subject them to potentially distressing valuation exercises.

Thus there may be a problem with approaching seriously ill patients with *ex ante* valuation exercises. However, it might be possible to approach patients suffering from life threatening conditions and ask them to value profiles retrospectively *ex post* after they have come through their own treatment. However, there may be situations in which it would simply not be worth the potential distress or confusion caused to patients by offering them valuation exercises. In such situations it might be appropriate to obtain valuations from carers, or where this is also too distressing, the general public may have to be approached with descriptions of the illness and asked to value these health states or profiles.

The AAA study highlights the problems that may be encountered in attempting to obtain patient samples. These problems are likely to apply particularly to studies seeking valuations from populations where there is dependence, such as elderly populations, the mentally ill, and children. However, these problems may also relate to patients in general.

The fact that AAA patients were difficult to recruit was due not so much to patient characteristics as to logistical problems at the hospital. Thus this was a problem with access to patients rather than the patients choosing not or being unable to take part. The health professionals who looked at the first draft of the questionnaire were of the opinion that it could be generalized to anyone as it did not clearly seem to relate to AAA patients in particular. Whereas the other studies related to specific conditions, the scenarios described in the AAA study described levels of risk that might be faced by people other than AAA patients. Thus in this case it was not felt that the necessity for a convenience sample detracted from the results.

Problems with recruitment were also encountered in the varicose veins study reported in Chapter 7. Even when patients responded to the initial letter of invitation, it was often very difficult to make appointments that were convenient for them. Many of the people to whom invitations were sent worked full-time. Making appointments in the evening helped, but even this did not entirely get rid of the problem. It was not possible to recruit as many patients as would have been wished for.

General population surveys are in a sense easier, because there is such a large sample from which to recruit. However, given sufficient resources it should be possible to recruit patients in most cases. Where an illness is relatively rare, and there are relatively few patients, it might be necessary to include non-patient groups in order to obtain a sufficient sample size.

The issue of whether samples are drawn from the general population or patient groups is equally relevant to the QALY and holistic approaches.

4.3.3 Own state versus hypothetical state

Even if patients are used as sources for valuations of health states and profiles, there is still a decision to be made regarding whether they should value their own state or hypothetical states. There is evidence that people who are in a state less than full health tend to value that state higher than people who are asked to value that state hypothetically (Salomon and Murray, 2002).

Salomon and Murray suggest three possible factors which might explain this phenomenon. The first of these is "adaptation", which means that the health domain changes over time. For example, a right-handed person who loses the use of his right hand may initially report lower health in terms of dexterity. However, after a while he becomes more dexterous with his left hand, and therefore he reports his dexterity as being equal to that before his loss. The second factor is "coping". This involves a change in norms and expectations rather than an actual change in health. Thus the man did not learn to be so dexterous with his left hand, but his expectations changed so that now he reports dexterity as being as good as previously. The third factor is "adjustment". This involves neither a change in health nor a change in norms, but the individual has adjusted so that there has been a change in relative importance of things in his life. Thus dexterity may have been of prime importance to the man in our example, but as time passes the relative importance of things in his life changes so that dexterity is no longer so important.

All this supports the valuation of patients' own states. However, if patients are being asked to value *ex ante* profiles consisting perhaps of risks and changes in health states occurring over a future stream of time, these patients are being asked to value the hypothetical. In other words, they may have no experience of the future health states, and they may not be able to predict their feelings. The valuation of *ex ante* profiles therefore has an impact on the argument that patients should be used because they know what it is like to be in the health state being valued. However, it is possible that a patient who has already undergone changes in health in relation to their illness may have a more informed view of what the impact of the *ex ante* profile would be. For example, a patient undergoing kidney dyalisis in hospital knows that this state impacts on their life in terms of the amount of time taken up by this form of treatment during each weak. If asked to value an *ex ante* profile involving a kidney transplant, this patient would perhaps be better informed than a member of the general public in terms of what aspects of his life might be affected by the transplant. For example, if the transplant was successful, the inconvenience of undergoing dyalisis. and all the factors associated with this, would be removed.

Hypothetical states or profiles are used commonly in valuation studies, even when patients are sampled (Brazier *et* al, 2003; Laupacis *et al*, 1993). It can vastly simplify studies to be able to value a standardised set of health states and profiles, because there would be variations in health for each individual patient. It may also be the case in some studies that it is very specific health states/profiles in which the researchers are interested, but they may not always be able to find patients in these precise states. However, as suggested above, it is possible that patients may be able to relate better to hypothetical states for their particular condition than a member of the general public.

The issue of whether respondents value their own or a hypothetical state or profile of health certainly applies to the holistic approach, in which respondents will be expected to value a health profile perhaps in the *ex ante*. It also applies to the QALY model, in which respondents are often expected to value hypothetical health states. However, if patients are valuing their own health states, these issues do not apply. Likewise, if a patient is valuing their own health profile retrospectively they do not apply.

4.3.4 Generic versus condition-specific

There are, broadly speaking, two types of health status measure (Patrick, 1997). Generic measures are intended to cover a broad range of factors relating to general health. The EQ-5D and SF-36 are two different types of generic measure. They deal with "domains" of generic health. For example, the EQ-5D deals with the five broad dimensions of mobility, self-care, usual activities, pain/discomfort, anxiety/depression.

The second type of health status measure is the condition-specific measure. As their name implies, condition-specific measures relate to specific conditions or illnesses. A condition-specific measure explores HRQoL with relation to a specific condition.

Generic measures of HRQoL are often used in order to produce values that may be compared across different health care programmes. These are useful tools to aid decisions regarding distribution of health care resources across different health care programmes. However, generic measures may be insensitive to small changes in HRQoL and may miss certain specific dimensions, and as such may not pick up some of the disutility associated with the condition (Donaldson *et al*. 1988; Bowling, 1991: Brazier and Fitzpatrick, 2002). Condition-specific measures are useful for providing information of specific interest to the condition or its treatment rather than in making broader decisions across different health programmes. They have fewer domains than generic measures, and can lead to a greater degree of refinement of detail (Brazier and Fitzpatrick, 2002). Both generic and condition-specific measures are useful in different ways, and it is generally recommended that both a generic and a condition-specific measure should be used (Brazier and Dixon, 1995; Patrick, 1997). Indeed, condition-specific measures exist for many illnesses.

Both condition-specific and generic measures can be used to obtain QALYs and also for holistic valuations. Although previous studies using the holistic approach have favoured condition-specific measures, health profiles could be devised using generic classification systems. For example, the EQ-5D domains could be used to describe a series of health states occurring over a period of time. However, in order to fully incorporate disutility derived from a condition and its treatment into any assessment of HRQoL, a condition-specific perspective would seem appropriate in addition to a generic measure. It is one of the aims of this thesis to explore methods of devising condition-specific scenarios for use in valuation studies.

4.3.5 Development of descriptive systems and the role of patients

Section 4.3.2 discusses whose values should be used in the allocation of health care resources. Even if the values of the general public should be used, they would require a sufficient descriptive system. It would still be valid to use descriptions based on patient discussions, because whoever is valuing the state should be valuing as realistic a state as possible, and as far as is possible should be made to understand the full impact of the state. This applies equally to QALY and holistic methods. For the holistic method it would be a descriptive system for profiles rather than states.

In order to construct condition-specific health states and profiles in terms of descriptions of the illness or condition, it is necessary to conduct research into the condition and obtain enough information to describe the condition. If, for example, health states relating to IBS are to be valued, the descriptions of IBS in the states should enable the study sample to understand what it is like to have IBS.

There are four sources of information about the illness under study:

- Literary sources

- Health professionals
- Patients
- Patients' carers or kin

The first port of call should always be the literature. Not only will this provide information about the nature of the illness, but it will also determine whether somebody else has already done the planned research.

Health professionals may also be a good source of information about the condition, as they may have had a great deal of experience across patient groups. These may, for example, include GPs, hospital consultants, and nursing staff.

Patients are a very good source of information about the effects of the condition on their life in general and HRQoL in particular. They may be able to provide information that would not be available in the literature or from health professionals. For example, health professionals may be aware that varicose veins cause swelling in the legs. However, only the patient can know how much this affects them in their lives. The patient is able to give a subjective account of the symptoms, such as degrees of pain, discomfort, and depression. Non-sufferers such as health professionals cannot know all this first hand. The subjectivity of patients can be a bonus in obtaining descriptive information about the condition, because it is the subjective assets (*e.g.* level of pain) that make the condition matter. It is self-evident that if pain existed at a low level it would not have such a great effect as if it existed at high levels. Use of patients can help ensure that the descriptions have face validity (Bowling, 1997).

Ideally some form of qualitative research, *e.g.* focus groups, should be conducted to determine the effects of the illness and treatment on patients' lives, and effects on HRQoL. The state descriptions would then be written to include the main symptoms mentioned by patients, using the terminology of the patients. In some cases it might be necessary to use carers or kin as a proxy for the patient perspective. This might, for example, be so in the case of very young children.

4.3.6 Methods of construction of descriptive systems

The descriptive method chosen for the states and profiles in this thesis was of the holistic vignette type, in which one or more health states are described as mentioned in the previous section, in terms of the main symptoms. Another type of description is the health state classification system. Health states are described by a number of multi-level attributes, such as physical function, emotional function, *etc.* Each attribute has levels from best to worst. Each health state is defined by one level of each attribute. and there may be hundreds of possible combinations (Torrance, 1986). The generic Health Utility Indices (HUIs) are examples of health state classification systems. Condition-specific health state classification systems have also been devised (Frank *et al*. 1999). If health state classification systems had been designed for the studies contained in this thesis, the profiles would have consisted of one or more of the states derived from the classification systems. Even the simplest health state classification system in existence at present has too many attributes to be feasibly incorporated into health profile descriptions.

Once the health state and profile vignettes have been constructed, it is important to test them for how well they actually represent the condition (face and content validity) on a sub-sample of people who have the condition under study (Brazier and Deverill, 1999; Brazier *et al*, 1999). This sub-sample should express an opinion on how well the state and profile descriptions represent their illness. One aspect they could consider would be the accuracy of the symptoms described in the state, and whether all the important symptoms were included. Another aspect under consideration might be the wording and phrases used in the descriptions. The wording of the states should use the terminology of the sample population.

If the sub-sample found the descriptions sufficient, it might still be advantageous to conduct a pilot study. In the pilot, a sub-sample of the target population would be asked to value all the states and profiles to be used in the main study. The process of questionnaire interviewing to be used in the study proper would be put into practice. The difference between the pilot study and the initial test is that the test is much less formal, and might incorporate discussion, for example regarding and improvements required to improve the state and profile descriptions.

These vignettes were used for the valuation of health states (for the QALY calculations) and health profiles (for holistic valuations). The downside of the use of these types of vignettes is that they are not only condition-specific. but also may be study-specific. Thus if the vignettes differ between different studies of the same condition. the results of the studies may not be directly comparable.

4.3.7 Problems encountered with construction of health states/profiles

Although the ideal way to construct health state and profile description has just been described, it was not possible to follow this in all the studies outlined in this thesis. The IBS study described in Chapter 5 was conducted in collaboration with a pharmaceutical company, who approached the author for assistance in constructing health states and profiles and conducting a valuation survey. Researchers there had already decided what symptoms they wished to be included in the descriptive system to be used in the health states and scenarios, which was originally to be entered into an economic model. This author was not party to the precise reasons for their choosing these particular symptoms, but the choice of symptoms related to the clinical trial of a drug. The author worked on the development of a health state and health profile description system containing these symptoms. There was discussion with IBS patients during the construction of the final descriptive system. The second IBS study described in Chapter 6 used the same descriptive system for the sake of comparability. This issue related equally to the QALY and holistic methods.

The states to be valued in Chapter 8 consisted of descriptions of morbidity which might occur if treatment for large AAAs was unsuccessful, in particular states associated with chronic renal failure and stroke. The scenarios were meant to portray the greater risks of entering these states associated with patients with large AAAs. It would, in theory, have been possible to construct the health states with the aid of people who had suffered stroke or chronic renal failure. However, in practice the logistics of forming these different focus groups would have been beyond the resources of this study. The health states and scenarios were therefore constructed using information from the available literature. Thus, as for the two IBS studies, the short-comings in the development of the health states and scenarios in this study related equally to the QALY and holistic methods of valuation.

The varicose veins study described in Chapter 7 was the one that most closely followed the ideal methods of developing descriptive systems for health states and profiles. The literature was consulted, and questionnaires were sent to health professionals in order to obtain their suggestions for what should be included in the descriptions. Patient focus groups were held, and the affects of varicose veins on the life of sufferers were discussed at length. From these three sources health states and profiles were constructed, and a pilot study was conducted. The method was applied equally to QALY and holistic methods of valuation.

4.3.8 Clinical versus functional

An interesting point which is worthy of discussion is the method of choosing which material to include in the final health states, and which material to leave by the wayside in the construction of these health states. Several points came up repeatedly in the varicose veins focus groups, such as worry over deep vein thrombosis (DVT) during flying, worry about the appearance of the legs, and worry about getting an ulcer. However, it is unclear whether all or any of these factors should be included in the state descriptions.

The "clinical" approach suggests that we should merely describe the health state and allow the respondent to read the state and come to his own conclusions about how he would feel about it (Torrance, 1986). Thus a visual description of the appearance of the veins would be given in order to incorporate this aspect, which is clearly important to many patients. They would read the description and decide how much distress they would feel due to the appearance described. Their loss of utility due to the appearance of the veins would therefore be accounted for in their valuation of the health state. Similarly, the probability of getting DVT or a leg ulcer would be described in the health state, and respondents would incorporate this in their valuation of the state.

The "functional" approach is that, if negative feeling such as worry over DVT or distress over appearances is shown to be a relevant factor of having varicose veins in the focus groups, they should be included in the health state descriptions (Torrance, 1986). Thus if most patients worry about their appearances, the state of *worry* should be described in the health state description. For conditions such as depression, it would be ridiculous to exclude descriptions from the emotional domain. However, the nature of depression is that it affects the emotional domain primarily, leading to possible secondary effects on the physical domain (*e.g.* loss of appetite). Varicose veins, however, has a primary effect upon the physical domain with the possibility of secondary effects upon the emotional domain (*e.g.* depression as a result of appearances).

Generic descriptive systems have been designed for both clinical and functional points of view. The HUI takes a rather more clinical approach, favouring descriptive health states. The EQ-5D, however, is based on a more functional approach (Drummond *et al.* 1997).

Another way to describe instruments designed to measure HRQoL is the "within skin"/"out of skin" approach. The World Health Organisation has defined ill health into the three categories of organic impairment, disability, and handicap (Nord, 1997). According to this classification, organic impairments are "within the skin" mental and physical dysfunctions; a disability is the inability to perform specific functions due to the impairment; a handicap is the limitation in performing roles that are the societal norm. The HUI instruments are largely concerned with "within the skin" impairments, with some concern in the domains of disability. The EQ-5D has a wider spread of interest, covering impairment, disability, and handicap (Nord, 1997; Hawthorne *et al*, 2000).

The studies in Chapters 5 to 8 of this thesis took a mixed approach. To a large extent the clinical descriptive method was used, but in terms of relevant items and phrases. However, functional items were also included when they had been expressed as particular concerns, such as "you may worry about the possibility of getting an ulcer", or "you may find that you are organising your life around your symptoms". Because the health profiles were based upon the health state descriptions, this issue relates equally to the QALY and holistic valuation methods.

4.3.9 Perspective and framing

Another consideration in the construction of health states and profiles is perspective. For example, symptoms may be described using 1st, 2nd, or 3rd person. Llewellyn-Thomas *et al* (1984) asked respondents to value five vignettes, each of which was described twice using two methods. One set of vignettes used a 3rd person perspective and a very stylised descriptive method. The other set of vignettes used the 1st person perspective, and was much more personal. Llewellyn-Thomas *et al* found that respondents consistently rated the vignettes in 3rd person higher than the vignettes in 1st person. However, these findings were not supported by Gerard *et al* (1993), who found that valuations of vignettes did not differ when they were framed as 1st, 2nd, or 3rd person or when they contained disease labels such as cancer.

The 1st and 3rd person perspective seem more personal than the 2nd person perspective, and it was desirable that respondents imagine that the health state and profile descriptions applied to them. However, there do not seem to be any definitive findings as to which of the 1st and 3rd person perspectives produce valuations which greater reflect respondent preferences. In the lack of evidence either way, a 3rd person perspective was chosen for the studies in this thesis, and respondents were asked to imagine themselves in hypothetical health states and profiles.

Whether vignettes are framed in terms of success or failure may affect valuations. McNeil *et al* (1982) described outcomes of surgery for lung cancer in terms of mortality ("10 die during the postoperative period ... and 66 die by the end of 5 years") and survival ("90 live through the postoperative period ... and 34 are alive at the end of 5 years"). They found that framing statistics in terms of mortality caused respondents to view surgery less favourably than when the statistics were framed in terms of survival.

There were three forms of risk involved in the studies of this thesis. The first was the SG procedure used for the IBS studies in Chapters 5 and 6, which traditionally relies on respondents stating the value of p (the probability of success) at which they would be indifferent between taking the risky option and opting for the certainty. By stating p, a value of $1-p$ (probability of failure) is implicit. The SG layout in this thesis described both probabilities adjacently, so respondents could choose whether to make a decision based on probability of success, probability of failure, or both probabilities together. This is the *ex post* perspective. The SG method was used to value health states and profiles, and therefore this *ex post* risk format applied equally to the QALY and holistic valuation methods.

The second type of risk was that associated with the AAA scenarios in chapter 8. The scenario descriptions incorporated variable *ex ante* probabilities of different pleasant and unpleasant outcomes. These were listed in terms of the level of probability, so respondents could see all the possible outcomes and their associated probabilities and make their choices accordingly. However, it was only during the holistic valuations that respondents were able to take *ex ante* risk into account. For the QALY values the respondents valued the health states, and then the QALYs were calculated by multiplying the value for each health state by the probability of its occurrence (*ex post*). QALY values were then adjusted for individual risk attitude, which was measured by a series of certainty equivalent questions with an *ex ante* perspective.

The third type of risk was used in Chapter 7 to describe risks of recurrence after treatment for varicose veins, and a risk of mortality during the surgical procedure (*ex* *ante*). These risks were described in terms of failure. It would have been unfeasible to have two versions of each question incorporating risk, so that respondents could value each scenario using a "success" and a "failure" description of probabilities. There were several valuation questions, and it was necessary to ask for the minimum number of valuations possible so as not to over burden respondents. Also the risks were included specifically to assess how inclusion of negative risks affected valuations. Valuations of the risky profiles were compared to corresponding valuations of the same scenarios minus the risk attributes. Again, these *ex ante* valuations of risk were only used for the holistic values. For the QALY values the *ex post* probabilities of recurrence were multiplied by the value of the health states involved.

4.3.10 Presentation

Different descriptive methods were used for health states and profiles as the thesis developed. For example, Chapters 5, 6 and 7 deal with comparisons between different health states and profiles within the same illness (IBS and varicose veins). Health states and profiles describing different levels of illness were ranked, and then valued separately. To aid respondents in comparing the different states and profiles, symbols were inserted beside the symptoms to indicate at a glance whether they were present or absent. In Chapters 5 and 6 ticks were used to indicate the absence of symptoms and crosses to indicate the presence of symptoms. Thus ticks meant a good, and crosses meant a bad. However, this confused some respondents.

Chapter 7 used sad face icons to indicate the presence of symptoms. During the drafting process a set of states was devised containing sad faces to indicate the presence of symptoms and smiley faces to indicate the absence of symptoms. However, this produced too much crowding on the page, and the faces were not easy to tell apart at a glance. By just using sad faces, the bad items of the state were emphasised, and it was relatively easy to compare the different state descriptions.

In contrast to Chapters 5, 6 and 7, the study described in Chapter 8 valued health states across different conditions. Thus respondents were asked to rank a state describing chronic renal failure with another state describing stroke. Unlike in the case of the IBS and varicose veins health states, these states did not contain the same dimensions. It was not therefore felt that the comparison would have been aided by the use of iconic symbols.

The icons used in Chapters 5, 6, and 7 related to both QALY and holistic methods of valuation, because the health profiles incorporated the states containing the icons.

4.3.11 Valuation techniques for health states

This section discusses issues relating to methods used to value health states. This relates more to the elicitation of health state values for the calculation of QALYs than to holistic valuations. However, the contents of this section is also relevant to the valuation of health profiles using the holistic approach, which is discussed in the next section.

There has been considerable debate in the literature about the relative merits of alternative methods to value HRQoL. The three main techniques of eliciting valuations for health states are visual analogue scale (VAS), standard gamble (SG), and time trade-off (TTO). These were described in Chapter 2. VAS is considered by most health economists to be theoretically inferior to SG and TTO because of its choiceless context (Dolan, 1998b, 2000).

SG is favoured by many economists and decision theorists (Dolan *et al*, 1996a) because it requires people to make decisions under uncertainty and it has a foundation in EUT. Supporters of SG believe that this makes it superior to TTO. However, the life and death risks described in SG are not necessarily those faced in real-life situations. On the other hand, TTO requires people to make decisions about whether to trade life years at the end of life for improvements in health earlier in life. This is, in some cases, a more realistic kind of choice to make. Both SG and TTO have been widely used in the health state valuation field, and there are strong arguments in support of both methods of evaluation. The arguments have been clearly set out in Dolan (2000).

As Dolan (2000) points out, one way of attempting to make a choice of which is better between SG and TTO is to test their feasibility, reliability, and validity. Both instruments appear to be feasible, because they have high response rates and a high level of completed responses. In terms of split-test reliability, Torrance (1976) found that both methods had correlation coefficients between 0.80 and 0.90. In terms of test-retest reliability, Reed *et al* (1993) found correlations of 0.82 and 0.74 for SG and TTO respectively. However, the differences were not statistically significant. Dolan *et al* (1996a) found that TTO, specifically the board-based version, performed best with a correlation coefficient of 0.81 compared to 0.71 for the self-administered SG. This difference was not statistically significant.

One measure of convergent validity would be to determine how closely rankings of QALY valuations of health profiles as calculated by TTO and SG correlated with direct rankings of these profiles (Dolan, 2000). Bleichrodt and Johannesson (1997) found that the correlation between the two methods with direct ranking varied with the discount rate applied. With no discounting the TTO showed significantly greater correlations with direct ranking. This difference in correlation between the two methods was reduced as the discount rate applied increased. The correlation between direct rankings and TTO was stronger up to a discount rate of 9%.

Thus according to convergent validity TTO might appear to be more closely correlated with direct rankings than SG. The greater convergent validity shown for TTO is based upon the assumption that a measure is more valid if it correlates closer with direct rankings of profiles. However, this is not necessarily the case. It is difficult to be certain that the direct rankings are a correct summary of respondents' preferences. Indeed, since the ranking exercise usually comes near the beginning of the questionnaire, it also acts as a familiarisation process by which respondents become familiar with the concept of valuing health profiles, and familiar with the profiles being valued in the study. It is quite feasible that respondents may change their mind about their ranking order as they go through the valuation process and reflect more on the health profiles. Thus, although an indicator of validity, direct rankings cannot be taken as the correct ranking for certain.

Bleichrodt (2002) argued that the reason for the differences between SG and TTO are that they are biased to differing degrees by several factors. TTO assumes that utility with respect to duration is linear. There is plenty of evidence to the effect that it is not linear (Cairns and van der Pol, 2000), and hence Bleichrodt suggests that TTO values would be biased downwards. Whereas bias with regard to utility curvature is not a problem with SG, utility is assumed to be linear with regard to probability. There is plenty of evidence that this is not the case (*e.g.* Miyamoto and Eraker, 1985), and therefore SG values would be biased upwards. Another source of bias would be loss aversion (Kahneman and Tversky, 1979), which would have the effect of causing an upward bias in both measures. A fourth area for possible bias is known as scale compatibility. This occurs when a respondent gives greater weight to an aspect of the question the greater its compatibility with the scale of measurement. For example, for TTO the scale of measurement is duration, and therefore a respondent might give greater weight to the duration of the state than the quality of the state. In the case of SG the scale is in terms of probability. An SG exercise involves three probabilities: 1, *p*, and 1-*p*. According to Bleichrodt, if the respondent focused on 1, the treatment state would seem less attractive, thus biasing SG upwards. If the respondent focussed on *p*, the treatment option becomes more attractive than non-treatment, and indifference would be restored by lowering *p*, thus suggesting a downward bias to SG values. If the respondent chose to focus on 1-*p*, the treatment option becomes less attractive and indifference is restored by raising *p*, thus biasing SG values upwards. Thus the effects of scale compatibility are ambiguous for SG, but would bias TTO values upward. Bleichrodt argues that, whereas SG will normally overestimate values, TTO values will give the appearance of being unbiased because they are actually biased both upwards and downwards by the different factors. Bleichrodt suggests that these effects might more or less cancel out, producing a seemingly unbiased estimate by TTO. It should be noted that this section has so far only discussed SG and TTO in the context of valuing health states, not health profiles, for which there is little information.

4.3.12 Holistic valuations

This section follows on from the previous section, and relates specifically to valuation of health profiles using the holistic approach.

The two-stage HYE measurement technique devised by Mehrez and Gafni (1989, 1991) is not the only way to value health profiles directly. As discussed in Chapter 2, it is believed by many to be equivalent to the TTO. As suggested by Drummond *et al* (1997), a one-stage TTO could be used to value a health profile. Ried (1998) referred to this as the generalised TTO. Health profiles can also be valued directly using a one-stage SG (Drummond *et al*, 1997). As already mentioned, TTO and SG perform similarly in terms of reliability and feasibility for health states. There has been little research into reliability and feasibility with regard to valuing health profiles. The choice of which one should be used is left to the context of each valuation study. One of the aims of this thesis was to explore the use of the holistic valuation approach using different valuation methods.

It has been argued that SG is more suitable for a study on varicose veins patients, because their potential treatments involve a gamble rather than a trade-off of life

(Michaels, 2001). However, some of the varicose veins profiles in Chapter 7 included descriptions of risk. Including probability in the health state description in the form of a risk-risk valuation would increase the cognitive load for respondents, which could lead to invalidity of responses (Jones-Lee *et al*, 1995). One solution might have been for the questions to use the SG task except for those profiles which contained risk. which would use TTO. However, this would have introduced an unnecessary cognitive burden to respondents, who would have had to comprehend and use both the SG and the TTO methods. It was finally decided that the main valuation technique should be TTO, because the disadvantages of introducing additional cognitive burdens to respondents would have outweighed the possible advantages of using SG. The TTO method was used in the same way as for the health states. For valuations of the health profiles the number of years in full health equivalent to 20 years with the profile was taken as the valuation – *i.e.* simply the indifference value written in the questionnaire.

Similar reasoning lay behind the decision to use TTO in the aneurysm study described in Chapter 8. Valuations of *ex ante* profiles containing risk is an extension of the original HYE model (Johannesson, 1995a, 1995b; Wakker, 1996; Drummond *et al*, 1997). A TTO-based version of this method was used in Chapter 8 in valuations of AAAs. The TTO was used in a certainty equivalent question. Profiles consisted of cumulative probabilities of ending up in particular health states, and the respondent was requested to state the number of years in current health for certain that he would feel was equivalent to the probabilities.

There were no *ex ante* risks incorporated into the IBS profiles of Chapters 5 and 6. Treatments for IBS may involve a risk rather than a shortening of life. It therefore seemed appropriate to use the one-stage SG as the valuation method to value the IBS profiles. The value of each profile was the value assigned to *p*.

4.3.13 Variants of the valuation techniques

This section goes through the valuations techniques available, methods of use, and the choices made for each study in this thesis.

<u>Choice of anchor</u>

Many health states would be regarded as comparatively mild, and many respondents may be unwilling to accept even a tiny chance of dying should the treatment fail. Consequently, the worst treatment outcome of death for such states in the SG question may be replaced by a non-fatal health state that is worse than any of the other relatively mild states being valued. According to expected utility theory, it is possible to 'chain' values elicited from such a gamble onto a scale of death to full health (which is required for cost-per-QALY analysis), provided that the treatment failure health state has itself been valued in a gamble with death. This chaining procedure should substantially increase the sensitivity of the scale, because it has the effect of magnifying the part of the scale between the failure health state and full health. However, the valuations for states obtained via chaining are often different from valuations of the same states elicited by the non-chained SG.

Valuations can be chained by adjusting either the state of success or the state of failure. Stalmeier (2002) found that chained valuations elicited by adjusting the failure state were higher than non-chained valuations, but the discrepancies between the non-chained and chained values were smaller for utilities close to the extremes of the scale.

Most of the IBS states being valued were relatively mild (in comparison to serious health states, *e.g.* stomach cancer), and it was felt that respondents may not wish to risk death in order to improve their health when faced with these hypothetical choices. IBS health states and profiles were therefore chained to death via the worst IBS health state. Since the utilities for the IBS states involved in these studies were close to top of the scale (*i.e.* close to 1), the advantages of chaining were considered to outweigh the discrepancies. This applied equally to the QALY and holistic valuation methods.

Titration versus ping-pong versions of standard gamble

The version of SG used in Chapters 5 and 6 for both the QALY and holistic methods of valuation was the "titration" method originally developed by Jones-Lee *et al* (1993). It lists values for the chances of success from 0 in 100 to 100 in 100, and respondents are asked to place a tick against all the probabilities of success at which they are confident they would choose the treatment and a cross against all the values where they would reject the treatment. The space between the ticks and crosses indicates the region of indifference. Where this region covers more than one probability value, a mid-point is taken to be the indifference value. For example, if a respondent ticked 99/100 but put a cross beside 98/100, a midpoint valuation of 98.5 was taken. This was divided by 100 to place the valuation on a scale of 0 - 1. Holistic valuations of the profiles were obtained in a similar way to the health states.

The other SG variant is the "ping-pong" method (Furlong *et al*. 1990; Brazier and Dolan, 2002). The interviewer begins at the top of the scale, asking the respondent whether he would choose the certain prospect or the uncertain prospect if the probability of the best outcome was set at 0.9. If the respondent chooses the uncertain prospect, he is asked to decide which prospect he would choose if the chance of success is 0.1. The interviewer continues "ping-ponging" up and down the probability scale until the indifference point is reached. This method employs a chance board is used, which displays the probabilities under consideration numerically and by pie chart. This method is interviewer administered, and is usually performed on a one-to-one basis.

The titration version of SG has been found to produce more consistent and reliable data than the more commonly used ping-pong method with props (Dolan *et al*, 1996a). The Jones-Lee *et al* titration variant also has the advantage that it lends itself better to group interviews than does the ping-pong method. The choice of which SG variant to use in this thesis was therefore pragmatic, and does not affect the QALY versus holistic comparisons.

Scales of probabilities of success and failure were listed in order to allow respondents to choose for themselves whether to frame the choice as a chance of success or a risk of failure. Scales were from 0 to 100. An interval of 5% was chosen for the scales, except for values between 95 and 100 for success. In this area of the scale an interval of 1% was used, with a corresponding interval for the scale of failure at this end of the scale. The reason for this was to attempt to increase sensitivity at this end of the scale, because it was thought that the IBS states and profiles were so mild that many respondents would not wish to take large risks of failure. The increased sensitivity allowed a more exact elicitation from those respondents who chose to take a small amount of risk with between 95% and 100% chance of success. Respondents were asked to imagine states lasting for the rest of their lives.

TTO variants

The TTO method used for QALY and holistic valuations in Chapters 7 and 8 was a self-completion method based on that suggested by Gudex (1994). She suggested a procedure whereby respondents are shown state cards and examples of the question design prior to completing the valuation exercise. Ten years was suggested as a suitable time horizon, with a scale interval of 6 months. Respondents would value a state called

"Life B" by indicating on the scale how many years in full health ("Life A") were equivalent to 10 years in a state less than full health ("Life B").

Gudex (1994) also suggested a method of valuing states rated as "worse than death". In this situation, respondents were presented with a choice between "Life A" and "Life B". "Life A" comprised the worse than death state for time *a* followed by full health for time *b*, where $a + b = 10$ years. "Life B" comprised immediate death. Respondents were asked to indicate the values of *a* and *b* which were equivalent to immediate death.

Dolan *et al* (1996a) indicated that the props version of TTO was more reliable in terms of fewer missing responses than the booklet version used here. However, the booklet versions facilitate interviewing in groups, which has certain advantages over one-to-one interviews (see below).

Life expectancy

There was the question of what life expectancy to use in the TTO questions used in the varicose veins study, for both the QALY and holistic methods of valuation. Gudex (1994) suggested eliciting valuations over a life expectancy of 10 years. However, for many of the respondents in the varicose veins study 10 years would be an unrealistically short life expectancy. The original intention was to use the life expectancy of each patient, as obtained from life table charts. This would have involved obtaining information from the team at the Vascular Institute on each patient invited in terms of age. This would have had the advantage of providing a realistic TTO question in which each patient faced realistic choices tailored to them. However, the disadvantages include the fact that the TTO scales would differ. Thus an 86-year-old woman would have had a life expectancy of 5.89 years, which could reasonably be rounded up to 6 years (Government Actuary's Department, 2001). Her TTO table would have been very short, expressing full health from 0 to 6 years. A 23-year-old woman, on the other hand, would have a life expectancy of 57.70 or 58 years. Her TTO table would have been from 0 to 58, and the intervals of the scales would have differed from those of the 86-year-old for practical reasons (*e.g.* to fit on one side of paper). These necessary differences in TTO scales between varicose veins patients could have lead to biases in the resulting data. According to the QALY model this should not have mattered because of the constant proportional trade-off axiom. However, this has been refuted in several previous studies (see for example Kirsch and McGuire, 2000; Stalmeier *et al*, 2001). In order to prevent such biases the same life expectancy of 20 years was used for each person. This was considered to be long enough to be reasonably realistic for most respondents, but short enough not to be outrageously unbelievable for the oldest respondents.

As explained in Chapter 8, a maximum life expectancy of 36 months was chosen for the AAA study as one of the aims was to explore valuations over a short life expectancy as might be experienced by someone with a large AAA.

Valuations of states and profiles worse than death

The study described in Chapter 8 followed the method suggested by Gudex (1994) for valuing states and scenarios worse than death, and therefore this section applies equally to the QALY and holistic methods. It is important that what is being measured should be made absolutely clear. A respondent may be asked to value state *h*. She is given a choice between Choices A and B. Choice A is to remain in state *h* for *t* years and then die. Choice B is to be in full health for the remainder of one's life and then die. The indifference point is *x*, the number of years in full health equivalent to *t* years in state *h*. The lower the value of *x*, the worse the individual finds state *h*. If $x = 0$, this means that the individual prefers immediate death to *t* years in the unabated state of *h*. She is then asked to complete a "worse than death" valuation. In this valuation she is asked to choose again between Choices A and B. This time, however, Choice A is a period *x* in full health followed by a period in state *h*, and Choice B is immediate death. The periods in full health and state *h* are varied until the individual is indifferent between Choices A and B. A higher value of *x* indicates that the person finds state *h* worse than if she chooses a lower value of *x*. This is because a greater number of months in full health is required to compensate for the period in state *h*. Gudex (1994) suggested that the ill health state should precede the full health state, because this might force the respondent to think directly about the effects of being in the ill state. The state might seem more remote if it would occur years in the future.

The only study in this thesis for which states or profiles were valued as worse than death is the AAA study (Chapter 8). It was felt in this study that full health was unlikely to occur after the health state of stroke described. Also the time horizon was 3 years rather than the 10-year time horizon suggested by Gudex (1994). The health state would not therefore appear so remote to respondents. This study used the ordering of full health and ill health as originally suggested by Torrance (1986).

Typically states considered worse than death are valued by equation (4.1). However, this expresses time in good health as a fraction of the time in bad health - thus a ratio of good:bad health states. This results in negative scores ranging from 0 to infinity. An alternative formula for "worse than death" states is given in (4.2). This alternative transforms utilities to an interval negative scale ranging from 0 to -1, for which states with scores closer to 0 are preferred to states with scores closer to –1 (Dolan *et al*, 1996b).

$$-x / (t - x) \qquad (4.1)$$

$$-x / t \qquad (4.2)$$

When valuing health states, zero could be taken to equal death, and 1 to equal full health. However, it is less clear to what state a negative score is equivalent. Although a negative score indicates that states are thought to be very bad, the degree of badness is not clear because of the unclear nature of negative valuations.

Bearing in mind the lack of clarity regarding negative values, there can be little advantage in allowing negative values to reach infinity. If a handful of respondents give high negative values to a state, this could drag the mean value of the state down significantly. Patrick *et al* (1994) suggest that it is preferable to transform worse than death values to an interval scale limited to –1, because otherwise the results would be highly skewed because of high negative values. However, it should be noted that –1 is still an arbitrary lower boundary.

Groups versus individuals

The fact that the valuation questionnaires used were self-completed meant that it was possible to save resources by conducting interviews in groups rather than on an individual basis. A possible disadvantage with conducting the interviews in groups is that it may be possible for dominant members of groups to lead valuations to a group norm. However, the group format has the advantage that it allows respondents to discuss the states and profiles if they wish during or prior to the valuation process. This may help them to construct and articulate their preferences. Although one of the axioms underlying EUT is that people have all their preferences well-formed and ready to be elicited (Drummond *et al*, 1997), there is overwhelming evidence that this is not the case (Shiell *et al*, 1997; Bernstein *et al*, 1999). It is probable that preferences are constructed during the thinking process of decision-making rather than being pre-formed and complete (Simon *et al*, 1972; Slovic, 1995).

Cabasés *et al* (2000) compared group and one-to-one interviews, in which PTO was used to obtain social values for seven EQ-5D states. They found that there were no statistically significant differences between values of each state obtained individually or in a group setting. In the study by Cabasés *et al* (2000), respondents in the group setting were allowed to discuss the exercises and change their initial values, but only a minority of respondents did so. The authors reported that the IQRs for health states from the group interviews were narrower than for individual interviews, and suggested that this might have been due to a reduction in outliers by the group process. There has been little research into comparisons between health state values obtained in group and one-to-one settings. The work of Cabasés *et al* (2000) supports the use of group interviews for health state valuations. In the absence of similar research into health profile valuations, there appears to be no intuitive reason to suppose the results would differ for profile valuations.

During discussions within the group it was possible for the interviewer to ascertain that the respondents understood what was required of them. Time was spent explaining the procedure to the group, and the interviewer used the first question as a demonstrative example of the procedure. By using this method it was possible to take steps to avoid the problem often encountered with postal questionnaires, where there is a mass of evidence to suggest that respondents often interpret questions differently than intended (Slovic, 1995).

It was the intention of the researcher to interview respondents in groups of between four and ten, but the interviewing schedule was necessarily dependent on the convenience of the respondents. This meant that the numbers in each group were highly variable. In fact, respondents were interviewed either on an individual basis or in groups ranging from two to ten. All interviews were conducted by the author, who introduced herself and the project to the respondents, explaining that all information divulged by the respondents during the interview would be treated as confidential. She explained all the tasks to be performed during the interview. She then took them through the first valuation task in the questionnaire, explaining the procedure. They completed the rest of the tasks on their own. However, the interviewer remained in the room, and respondents had the opportunity to request further explanations as required.

4.3.14 Power calculations

It is essential in valuation studies to ensure that the sample is large enough to detect a meaningful difference between valuations for different health states, and equally so for health profiles. However, there is no gold standard for what is a meaningful difference. According to O'Brien (1996) and Drummond *et al* (1997) further research should be done to determine what the minimum economically important difference (MEID) is for evaluations. Essentially the MEID depends on the threshold cost-effectiveness ratio and all other parameters in the economic model being used.

An MEID in utility of 0.05 is commonly adopted with 80% power and 5% alpha. This was used to calculate sample sizes for the studies, using the equation (4.3) (Walters, 1999), where the mean difference is the MEID. In support of this the first study conducted in this thesis, the IBS study reported in Chapter 5, was able to detect a maximum mean difference between profile valuations of around 0.05. The standard deviation (SD) obtained in the IBS study was 0.13. From this it was calculated that a sample size of 56 would be sufficient to detect a difference of 0.05 in utility between the different profiles with 80% power and 95% chance of detecting this difference. The sample size for the first IBS study was pre-set by GW.

$$n = 2 + \{8 / (\text{mean difference} / \text{SD of difference})^2\} \qquad (4.3)$$

It should, of course, be noted that the mean difference and SD of difference which were used to calculate sample sizes were obtained from the first IBS study, which used SG. They were then applied to the other three studies, two of which were TTO-based. It would have been preferable to use an SD of the difference obtained from a TTO study to apply to sample size calculations for TTO studies, rather than using an SD of the difference from an SG study. However, it should be noted that the use of a formal sample size calculation appears to be very rare in this field.

4.4 Analysis

4.4.1 Overview

QALY and holistic methods of valuing health profiles will be compared in terms of feasibility, reliability, convergent validity, and logical consistency (Dolan, 2000). The scores obtained from the two methods will also be compared to determine if they produce different results.

4.4.2 Feasibility

One way of assessing the feasibility of a valuation technique is to examine the proportion of completed questionnaires (Dolan, 2000). It is also possible to determine the percentage of completed questionnaires which had to be excluded from analysis because of unclear responses, which may indicate a lack of understanding. The feasibility of each study is discussed in Chapter 9.

4.4.3 Reliability

Due in part to limited resources and in part to the difficulties in obtaining respondents, it was impossible to conduct a test-retest analysis of the questionnaires. However, a split-test of reliability was conducted in the second IBS study (Chapter 6). The health profile in the final valuation exercise in this study was equal to the failure outcome state of the gamble. As such, respondents should have been willing to gamble until there was a 0% chance of success, because life could not get any worse.

It was felt that adding a similar test to the varicose veins and AAA questionnaires would have increased the cognitive load to respondents, thus reducing the feasibility of the questionnaire.

4.4.4 Convergent validity with rankings and logical consistency

Respondents ranked profiles in three ways. Initially there was a direct ranking exercise. This was followed by valuation tasks using the composite QALY approach and a holistic method of valuing the health profiles. Each valuation method produced a ranking order of the profiles in terms of the different values given to each profile.

For the IBS and varicose veins studies described in Chapters 5, 6 and 7, the profiles could be ranked in a logical order. Thus there was one profile that would logically be considered to be the most favourable, one that was second most favourable, and so on. If the ranking implied by a valuation method follows this logical ordering, there is said to be logical consistency for the valuation method. If the profiles that are logically more favourable are consistently ranked higher than those that are less favourable, this is said to be *strong* logical consistency. However, if some of the profiles that should logically be ranked higher are ranked equally to the less favourable profiles, this is said to be *weak* logical consistency. If less favourable profiles are ranked higher than profiles which are logically more favourable, this is said to be logically inconsistent.

Convergent validity occurred when rankings obtained by either holistic or QALY valuations were the same as the original rank orders stated by respondents for the health profiles. As for logical consistency, there are degrees of convergent validity. Thus if the rankings by the holistic or QALY valuation methods followed the original rankings given by the respondents exactly, this is said to be strong convergent validity. However, if the valuation method ranked two profiles as equal that were ranked one over the other originally, this is said to be weak convergent validity. To be non-convergent, respondents had to provide valuations which were either in the opposite rank order to that stated, or were not equal to states where equality of preferences was stated in the ranking procedure.

Logical consistency in the orderings of the profile valuations was looked upon as a more reliable indication of valuation method accuracy than convergent validity in these studies. Thus it was thought that the valuation method which adhered more closely to logical consistency would be more likely to reflect the preferences of the individual, because it indicated an understanding of the decision-making process.

The profiles in the AAA study described in Chapter 8 did not have a logical ordering, and so only convergent validity was used. The implied ranking of the scores for each profile were compared with the original ordinal ranking of the profiles. This was done for both QALY and holistic valuations. Respondents were not excluded from further analysis on the basis of non-convergency between valuation scores and the original ranking of the profiles. The tasks in the questionnaire were relatively difficult to complete, and the ranking exercises were seen as warm-up exercises which gave respondents the opportunity to become accustomed to state and scenario descriptions. Because of the issues of learning, it may not be the case that the method giving results closer to the original ranking is the more accurate method. Thus this is not a foolproof test, but it acts as an indication of validity.

4.4.5 Comparison of results from QALY and holistic methods

The results of QALY and holistic valuation scores for the health profiles were compared for each study using appropriate statistical tests. The data was negatively skewed, because of the tendency for scores to be near the top end of the scale. The Wilcoxon-sign test was therefore used to compare each person's scores for each profile as obtained by the QALY and holistic methods since it does not assume a normal distribution. The Wilcoxon-sign test compares the distributions of the two variables. However, the paired t-test was also used since it compares the means, which are more often used in economic evaluation. It can also be used to estimate a 95% confidence interval. Parametric tests are generally considered to be stronger than non-parametric tests (Hicks, 1990), and are thought to be legitimate tests in samples of over around 50 (Lewis and Traill, 1998).

The null hypothesis was that mean QALY and holistic values would not differ significantly. In the AAA study described in Chapter 8, an attempt was made to measure respondents' attitude to risk and time preferences. QALY values were adjusted for these values and again compared to holistic values for scenarios.

4.5 Conclusions

This thesis sets out to test for violations of the axioms underlying the QALY model. Comparisons are made between valuations of health profiles obtained by the QALY algorithm and those obtained from holistic health profiles. The extent of the differences are investigated, and possible reasons for any such differences are discussed. Valuations are obtained for states and profiles describing IBS, varicose veins, and scenarios relating to AAAs.

The studies in this thesis take the perspective of the patient. Condition-specific health states and profiles were constructed for each study after consultation of the relevant literature, patients and health professionals. The self-completion non-props methods of SG and TTO designed by Jones-Lee *et al* (1993) and Gudex (1994) respectively were used to elicit valuations for states and profiles. These have an advantage over methods involving props in that they can be administered in a group setting. Conducting the interviews in a group setting allowed the economical use of limited resources. A group setting also has the advantage of allowing discussion, which may aid respondents in constructing their preferences. The elicitation methods were tested for feasibility, reliability, and validity.

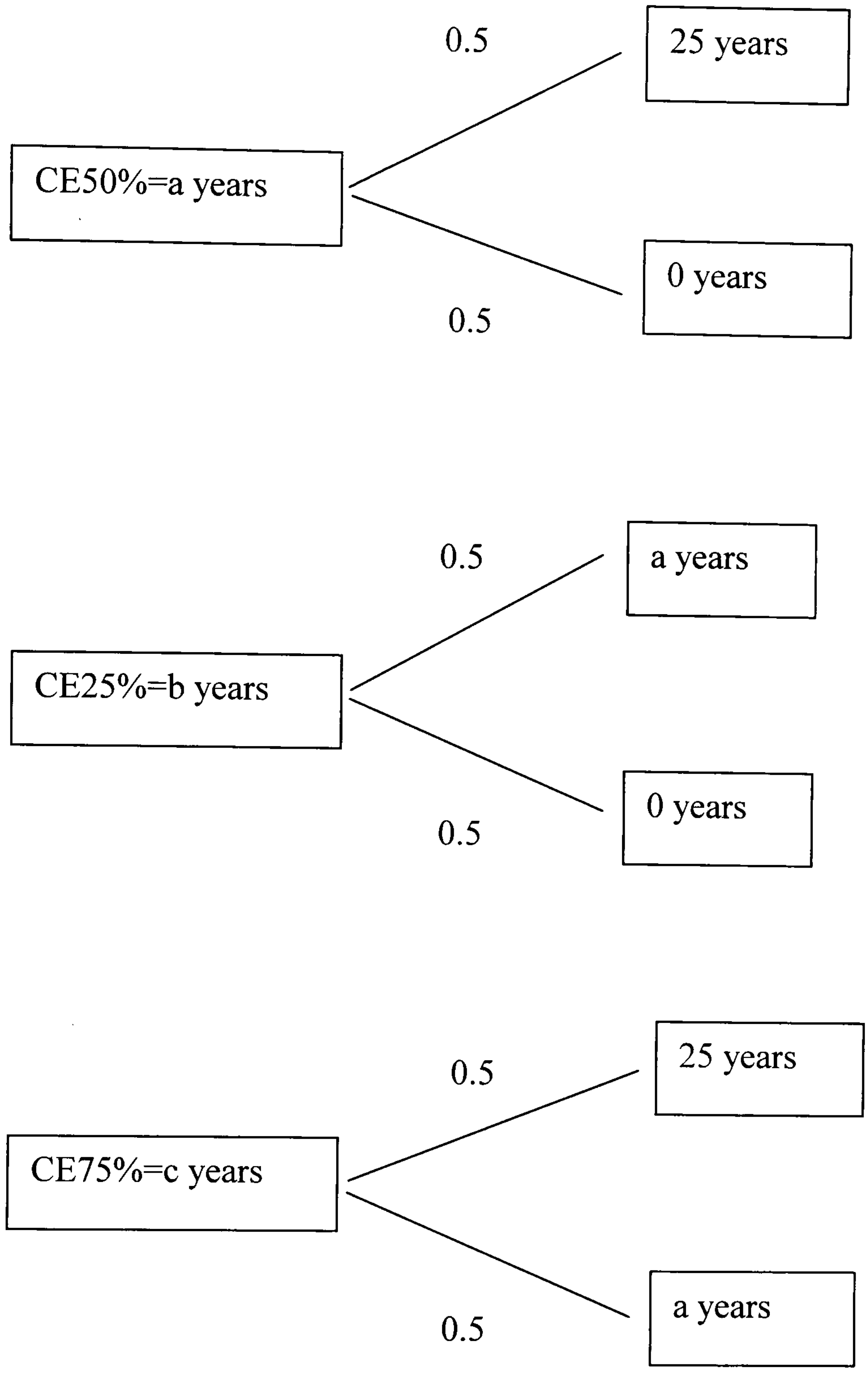


Figure 4.1 The method of determining certainty equivalents from McNeil *et al* (1978).

Table 4.1 QALY assumptions dealt with in each of the studies.

QALY assumption	*Ch.5*	*Ch.6*	*Ch.7*	*Ch.8*
Constant proportional trade-off				
Duration				
Zero time preference				✓
Quantity effect				
Mutual independence of life quality and quantity				
Constancy of risk attitude to survival duration				✓
Risk neutrality under all health states			✓	✓
Additive utility function				
Valuations of health state independent of succeeding/preceding health states				
Small duration effects and process are negligible to the patient			✓	
Valuations of profile proportionate to time in each state	✓	✓		

Chapter 5

Study 1: A test of additive utility over differing proportions of time in irritable bowel syndrome health states

5.1 Background

The background to this study was the development by the pharmaceutical company GlaxoWellcome (GW) of an economic model to determine the cost-effectiveness of a new treatment for irritable bowel syndrome (IBS). It was anticipated that the main benefits of the new treatment would be improvements in the HRQoL of IBS sufferers. The economic model would combine clinical data with resource use data. However, in order to estimate the number of QALYs associated with the new treatment as compared to existing interventions, it is necessary to have single index values for the main health states associated with IBS and its treatment. This study aimed to assist the GW economic model by providing valuations of health states and profiles related to IBS. It also compared QALY valuations of profiles with holistic valuations of the same profiles. The profiles were framed in terms of differing frequencies of symptoms.

Functional gastro-intestinal disorders account for a significant proportion of primary care and hospital outpatient visits. IBS is the most common of these disorders. IBS sufferers typically present with a wide range of symptoms, and there is growing evidence that they experience a level of HRQoL significantly below that of the general population (Wells *et al*, 1997; Akehurst *et al*, 2002). An important feature of IBS is that its symptoms tend to fluctuate considerably over relatively short periods of time. This has important implications both for the assessment of HRQoL in IBS patients and for the extent to which the QALY approach can represent the benefits associated with their treatment.

5.2 Objectives

A primary objective of this study was to obtain valuations for a number of IBS-related health states and profiles for use in the economic model devised by GW. There are several ways in which these valuations could be generated and this study sought to compare valuations from different approaches.

First, valuations for a number of IBS-related health states were estimated indirectly from responses to the SF-36 questionnaire from a large sample of IBS patients recruited to one of GW's clinical trials.

Second, valuations were elicited from a smaller sample of patients for those same health states. Thirdly, holistic valuations of profiles of health states were compared with the implied number of QALYs generated for the same profiles by using the health state valuations elicited from the same sample. This was in order to determine the extent to which assumptions of the QALY model were violated.

A fourth possibility would have been to estimate values for these states for the smaller study sample using the rating of their own health on the EQ-5D and SF-36 rather than SG or TTO methods. However, the number of respondents in each of the IBS states would have to be much larger than that required for the valuation study.

5.3 Methods

There were four main phases of work in this study. Firstly, it was necessary to develop the health states and health profiles associated with IBS.

Secondly, values for these IBS-related health states were obtained from data on the SF-36.

Thirdly, the study to elicit valuations directly from patients needed to be conducted. This involved designing the valuation exercise, selecting the sample, and conducting the interviews.

Fourthly, the data from the study required analysis.

5.3.1 Description of IBS-related health states and profiles

The health states were selected to be compatible with data collected in two clinical trials recently conducted by GW. In these trials, patients were asked whether or not they had experienced abdominal pain and discomfort (P) in the last seven days, and for the number of those days on which they felt a sense of urgency (U). "Urgency" was defined as having an urgent need to empty the bowels. For the purposes of the economic model, the urgency data were translated by GW into a dichotomous variable of more than three days and three days or less. Patients could also report whether or not they had suffered from constipation (C). These three key symptoms form eight possible health states (see Table 5.1). The presence of a symptom will henceforth be indicated by a '+' sign and the absence of the symptom by a '-' sign; thus P-U-C- is the best IBS-related health state and P+U+C+ the worst state.

The valuation questionnaire survey was clearly going to be quite lengthy and complex. and there was a desire to refrain from causing cognitive overloading of respondents. For this reason only six of these possible states were used in the study. Table 5.2 shows the number of patients in each health state in weeks 1 and 12 of the two clinical trials conducted by GW. Only 0.67% of their trial samples had health states P-U+C+ and only 1.25% had P-U-C+. The two states P-U+C+ and P-U-C+ were excluded to avoid overloading the respondents in this study. The states of P+U-C+ and P+U+C+ were included even though each of these states were only experienced by 1.67% and 2.08% of the GW sample respectively. Although the original intention of GW was to exclude patients who suffered from constipation, a small number of their sample did have constipation. States with constipation were included, because this was regarded as an important complication of treatment.

Derivation of the profiles was less straightforward than the states. Previous work constructing health profiles has dealt mostly with scenarios where patients are expected to go through a clear sequence of health states. However, IBS is characterised by considerable fluctuations in HRQoL over relatively short periods. In addition, the patterns of symptom fluctuations vary widely across patients. One way of summarising health experience of patients using the six IBS health states from Table 5.1 is in terms of the proportion of time spent in each state over a set period. In the case of IBS, a treatment may not be expected to entirely eliminate the patient's experience of a given state. Rather, treatments may be judged according to their ability to reduce the amount of time spent in that state.

An economic evaluation should ideally compare data for patients receiving the new treatment with data for those receiving the next best alternative. At the time this study was designed, there were only data relating to how the new treatment compared to a placebo control. There was also an issue with GW relating to professional confidentiality, and they were keen to provide as little information regarding their clinical trials as possible. Although the availability of further information might have been beneficial to their economic model, this was not so important to the objectives of the study as regards this thesis,which were to test the assumptions of the QALY model in IBS and explore issues relating to the use of holistic valuation methods. The use of actual trial data was not so essential for these purposes, and the best was made of the information available for the construction of health profiles. which in the end were based upon the possible range of profiles likely to be experienced by IBS patients with or without the benefit of the new drug.

Four profiles were chosen, based around the prevalence of symptoms at weeks one and twelve in the placebo controlled trials. In the absence of adequate information from GW, the following assumptions were made about the trial data in order to construct the profiles:

1) At Week 1 there was no effect of treatment. Week 1 was baseline.

2) At Week 12 the full effect of treatment was being experienced.

3) The percent of patients in a given health state was equal to the proportion of time spent in that state.

4) All benefits were accrued in the treatment group, and the placebo group underwent no change.

It seems safe to assume that any effect of the trial drug so close to first administration would have been minimal, thus there is no problem with assumption 1. Whether the full effect of treatment was being experienced at Week 12 is impossible to be certain of with the limited information available. However, assumption 2 is reasonable. Perhaps the most controversial of these assumptions is number 3. The data relating to people in each health state at Weeks 1 and 12 of the trials is presented in Table 5.2. These data were used to construct health profiles, such that the most favourable profile in terms of proportion of time in each health state was based on Week 12 and the least favourable health profile was based on the data in Week 1. The model for the construction of the health profiles is described in detail below. As can be seen in Table 5.2, 41.72% of patients were in the state P+U+C- in Week 1. This was used to construct a 12-week profile in which each individual patient was assumed to be in this state for 41.72% of the time over the 12 weeks. Thus the percent of patients in each state during Week 1 has been used to estimate the proportion of time spent in each state per patient over a 12-week period. In other words, the frequency of each health state during Week 1 has been translated to frequency of that health state per patient over 12 weeks. Although it is acknowledged that this is an assumption rather than a statement of fact based on trial data, it seems reasonable in the absence of more information from the trials. The fourth assumption is a simplifying assumption. For the purposes of the construction of profiles for testing the QALY and exploring issues relating to holistic valuations within this thesis, this is a reasonable assumption. Whether it was a reasonable assumption to apply to the economic model used by GW for their trial data is another question, which is impossible to answer with the level of information available.

The profiles were described in terms of the numbers of weeks in which a patient would be in each IBS state over a 12-week period, assuming that the overall length of time in each state was divided randomly over the 12-week period. Thus the states would occur in a random order, and the total time in a given health state might not occur in one block, but may be spread randomly over the 12 weeks. The 12-week period would then repeat itself for the rest of the person's life, each period having the states randomly distributed. This was made clear during explanations to respondents during the interview process.

The number of weeks in each state for the pre-treatment baseline profile was calculated using the following method. In Week 1, 41.72% suffer from P+U+C- (Table 5.2). It is assumed that every IBS patient in the sample will be in the state P+U+C- for 41.72% of the time over a period of 12 weeks. Thus 5.01 weeks out of 12 would be spent in the state P+U+C-. According to Table 5.2, 17.99% were in the state P-U-C-. Thus 2.16 weeks would be spent in this state in the pre-treatment baseline profile. A total of 2.41 weeks would be spent in the state P+U-C-, and 1.75 weeks in P-U+C-. These values summed to 11.33 weeks rather than 12 weeks. The reason for this is that a small proportion of the GW patients in Table 5.2 reported constipation, and these were excluded from the present study. The values were rounded up, and P+U+C- was rounded up to 6 weeks, giving Profile E (Table 5.3).

The calculations for the post-treatment profile were slightly more complex. It was assumed that the reduction in P+U+C- in Week 12 must have been within the treated group. Thus the 26.55% remaining in this state must be the placebo group plus the treated patients who did not improve. For the placebo group it was assumed that 41.72% of the time was spent in this state. The following formula was used to calculate the number of weeks in each health state for the post-treatment profile.

$$\{\%(\text{Week 1})+\%x\}/2=\%(\text{Week 12}) \qquad (5.1)$$

Thus for the treated group, x% of the time was spent in P+U+C-, and this can be calculated from equation (5.1) as {(26.55*2)-41.72}=11.38%. Thus 11.38% of 12 weeks, *i.e.* 1.37 weeks would be spent in P+U+C- in the post-treatment profile. The calculations were again rounded up to give Profile C in Table 5.3.

Profile D is an additional profile in between the two extremes of C and E. Profile I was designed to include two weeks of C+, which could occur at any point over the 12 weeks.

The two most notable changes in the trial data over the 12 weeks were a decrease in P+U+C- and a corresponding increase in P-U-C- (Table 5.2). It should be noted that the GW sample, and therefore the sample used in this present study (see Section 5.3.4), was not necessarily representative of IBS patients in general.

These profiles were in no way intended to reflect the effectiveness of the treatment, but are meant to cover the range of likely outcomes of treatment.

5.3.2 The SF-36 valuation of health states

In addition to collecting information on each individual patient's IBS-related health state, HRQoL was collected in the same GW clinical trials using the SF-36. This meant that it was possible to relate each patient's IBS-related state to their SF-36 responses. It is now possible to calculate a single index value for the SF-36 using an algorithm based on valuations elicited from the UK general population (Brazier *et al*, 1998; Brazier *et al*, 2002). It was therefore possible to calculate single index values for each IBS health state. These values were estimated separately for Week 1 and Week 12 data, and an overall average value was produced for each state.

5.3.4 The sample

The inclusion criteria for this present valuation study were women of 18 years of age or over who had been diagnosed with IBS based on the Rome criteria (Figure 5.1). They also had to have completed at least one of the relevant GW studies in the UK. The aim was to recruit a sample of 50 respondents from four ambulatory care centres in UK: Chorley, Crosby, Macclesfield, and Reading. This sample size was pre-set by GW, based on their estimation that it would be sufficient to detect a difference of about 0.1 (on a 0-1 scale) between different health states. There is no real consensus about what difference is to be considered meaningful but a difference of this kind is likely to be important in many contexts (O'Brien and Drummond, 1994).

Because this study involved patients, it was necessary to seek ethical approval. Prior to the commencement of the study, ethical approval was obtained by GW.

5.3.5 The interviews

The SG valuation method chosen for this study was that developed by Jones-Lee *et al* (1993). The reasons for this choice of valuation method are presented in Chapter 4.

A pilot study was conducted to ensure that the interview procedure was understood as intended. The questionnaire was piloted on the first five women to be interviewed, and these women were specifically asked for their comments on the state and profile descriptions and the SG process used. They considered the states and profile descriptions to be realistic, and the SG process to be fully comprehensible. It was not necessary to make any changes.

Patients were interviewed in small groups of 2 to 6 by a trained interviewer (the author). The health states and profiles and the SG procedure were considered too complex to send through the post for respondents to complete on their own. The interview consisted of a self-completed questionnaire (see Appendix 1). The questionnaire began by asking respondents for relatively non-intrusive background information relating to their age, occupation, age since they completed education, the length of time for which they had experienced IBS, and a general health question. They were then asked to indicate their current experience of the three IBS symptoms of pain, urgency and constipation. It was made clear to them that these questions formed the basis of the health states they would be asked to evaluate. At this stage, respondents had the opportunity to ask questions relating to how the descriptions should be interpreted.

Respondents were then asked to rank the states and profiles in order to familiarise themselves with the descriptions. They were presented with slips of paper upon which the health state descriptions were typed for the purpose of the ranking exercise, so that they were able to lay the states out before them and place them in the order of preference. Full health and death were included in the ranking exercise. They then completed the valuation exercises in the questionnaire. The questionnaire first asked them to evaluate five IBS-related states (the two mildest against a treatment failure state of P+U+ and the remaining three against death) (see Table 5.4). There was a written explanation of the SG health state valuation exercises. The interviewer read this aloud while the respondents had the opportunity to refer to the first valuation exercise over the page. The interviewer then further explained the procedure, using the first exercise as an example, until all respondents understood the task. Respondents completed the health state valuation exercises, with the interviewer present to assist if further explanations were required.

After completing the health state valuations, respondents were asked to rank the four health profiles, including full health and death. This was preceded by a written explanation of the health profile descriptions, which the interviewer read aloud. Each respondent was presented with an envelope containing written descriptions of the health profiles on pieces of paper. The interviewer proceeded to explain the health profiles, using these descriptions as a visual aid. It was verbally explained that the health states were those with which respondents had been dealing in the previous section of the questionnaire, and that in the profiles they were described in terms of frequency. Each health state would occur at random within the 12-week period with the stated frequency, and the 12-week block would repeat for the remainder of the respondent's life. Once each respondent understood the task, they ranked the profiles in order of preference.

The next task was the valuation of these health profiles using the SG procedure (see Table 5.4). Again, the interviewer explained the procedure, using the first valuation exercise as an example. Respondents then completed the four exercises, while the interviewer remained present to offer further explanations as required.

After completing the questionnaire, respondents were asked to rate their own health using the SF-36 (Ware and Sherbourne, 1991) and the EQ-5D (Brooks, 1996). When everyone in the group had completed the questionnaire, respondents were provided with the opportunity to comment on the exercise.

5.3.6 Analysis of the study data

Current health

Respondents' current health was described in terms of SF-36 rating of excellent, very good, good, fair and poor. Current health was also described on the EQ-5D dimensions scale. Scores for current health were then calculated for both methods. Current health was also described on the IBS classification system devised for this survey.

Logical consistency of responses

The first phase of the analysis concerned the quality of the data, particularly to determine whether it was necessary to exclude the valuations of certain respondents. In order to check the extent to which respondents were able to understand the valuation tasks, the logical consistency of their responses was examined. For some pairs of IBS health states, one state can be regarded as logically better than another if it is better on one or more of the symptoms and no worse on any symptom. The state that is logically better should not be valued as lower. Where there is evidence of a high rate of inconsistency, there might be a case for excluding such respondents. There is no gold standard to guide what level of inconsistency should be deemed unacceptable. A low level of inconsistency may arise due to a certain level of arbitrariness or random error in use of the SG scale. For most of the scale there was an interval of 5%. It seemed reasonable to allow respondents to have a random error of up to one interval on either side of their true value, which in this case allows a random error of +/- 0.05. This is the MEID discussed in Chapter 4. Thus if valuations of states or profiles were within 0.05 of being ordered consistently, respondents should not be excluded.

Responses to the ranking questions were first examined. Respondents were asked to rank the health states and profiles in two separate exercises. For each of these exercises the responses were examined to determine the degree of inconsistencies.

The raw valuation data was then examined. Valuations for the health states were compared to explore the level of logical consistency in the implied ranking. The same was done for the health profiles.

Calculating values for health states and profiles

For the states and profiles that were not evaluated against death as the failure outcome, it was necessary to 'chain' the valuations for these states onto a scale of 0-1 (death-full health) (Torrance, 1986). This was done by using the valuation of the reference state from a second gamble where it had been valued against full health and death. In all five cases, the reference state was P+U+C-, which was valued in Question 3 (Appendix 1). If the health state valuation obtained in the first gamble is U_1 and the valuation of the reference state (P+U+C-) from the second gamble is U_2, the formula for chaining onto the 0-1 scale is given in equation (5.2).

$$U_3=U_1+(1-U_1)(U_2) \qquad (5.2)$$

To calculate the QALY profile valuations using the valuations for discrete health states, each health state valuation was multiplied by the overall proportion of time spent in that state. This was straightforward for profiles C to E, which did not contain periods of constipation. However, Profile I included two weeks of constipation, but it did not specify when this would occur. If the impact of constipation on health state valuations is not additive, then it matters which state(s) it is combined with. For the purposes of the analysis, it was assumed that constipation would occur for one week of state P+U+ and one week of P+U-, the most common combinations of symptoms which include constipation according to Table 5.2.

A statistical analysis was conducted, which included the calculation of the descriptive statistics for the valuations of the IBS states and profiles. Given the skewed nature of some of the data, the results were confirmed by non-parametric methods. All analyses were undertaken using SPSS.

Convergent validity

The test of convergent validity described in Chapter 4 was carried out, in which the degree of agreement between ranking order and implied ranking by holistic valuations of profiles and QALY valuations was examined. There were four profiles involved (profiles C, D, E, and I), and therefore three pairs of profiles in the ordering. For example, if a respondent ranked the profiles in the order C $\succ$ D, D $\succ$ E, E $\succ$ I, and then went on to give values to the profiles using the QALY and holistic valuation methods such that the profiles were also rated in this order by these valuation methods, there would be complete agreement between the ordering of the three pairs for the original ranking each valuation method. If one of the valuation methods resulted in the respondent giving values such that the only difference was E $\succ$ D, there would be one pair out of the three pairs that was non-convergent. There were four possible levels of convergency for these three pairs for each valuation method: 3/3 pairs convergent, 2/3 pairs convergent, 1/3 pairs convergent, and 0/3 pairs convergent. The responses for each member of the sample were examined to determine their level of convergence between the original ranking order of the profiles and that implied by each of the QALY and holistic valuation methods. Convergency was examined both at the level of strong convergency (*i.e.* a profile ranked higher would be given a higher value, ranked equally would be given an equal value, and ranked lower would be given a lower value) and weak convergency (*i.e.* equality would be allowed where profiles had been ranked ordinally).

Logical consistency

Because there was a logical ordering to the profiles in this study, a test to determine which method gave results that were closer to this logical order was considered more valid than a test for convergent validity. As suggested by a mass of literature (see Chapter 3) and the discussion in Chapter 4, people may not have a complete set of preferences. Rather, they form their preferences during the elicitation process. Thus the original ranking order need not be seen as a true reflection of respondents' underlying preferences. A test of logical consistency was also therefore carried out.

Logical consistency was examined by pairwise comparisons. There were three profiles to be compared: C, D and E. Thus there were three possible comparisons: C versus D, D versus E, and C versus E. The chained valuations for profile C was compared to profile D, profile D was compared to profile E, and profile C was compared to profile E for both the QALY and holistic methods. Another way of looking at logical consistency would have been to compare profiles C, D and E all at once. However, it was thought that this would have been less informative, and would not have given much information about the degree of logical consistency for each method.

Respondents' comments

At the end of the interview respondents were given the opportunity to comment on any aspect of the interview. The final page of the questionnaire was dedicated to their comments. These were examined and categorised.

5.4 Results

5.4.1 Estimation of SF-6D values for health states

Table 5.5 shows the mean health state values calculated by applying the SF-6D preference-based algorithm which was derived from the SF-36 (Brazier *et al*, 2002) to the clinical trial data. There are small to moderate differences in value between P-U-C- and the other states without constipation but these differences are statistically significant. The numbers were too small to be able to estimate the impact of constipation, but there was no evidence of any substantial effect on health state values from combining the different states with constipation into a single state, C+.

5.4.2 Background characteristics

In total, 49 respondents took part in this study. The data was highly complete; there was only one missing SG valuation. The interviews took between 30 minutes and one hour to complete, with larger groups taking longer. The mean age of the sample was 46.6 (median 48.0) years and the average number of years with IBS was 12.1 (median 10.0). The mean age at completion of full-time education was 16.6 (median 16.0) years. A total of 37 (75.5%) of the women were in full-time or part-time paid employment. The background characteristics are presented in Tables 5.A.1 and 5.A.2 in Appendix 1.

Six interviews took place in Macclesfield, compared with 11 in Reading, 15 in Crosby and 17 in Chorley. There were no significant differences in any characteristics between the respondents from the four sites.

5.4.3 Current health

One (2.0%) respondent rated her current health as excellent; 15 (30.6%) rated current health as very good; 26 (53.1%) rated current health as good; six (12.2%) rated current health as fair; and one (2.0%) rated their current health as poor. In terms of the IBS symptoms, 32 (65.3%) reported that they did not have adequate relief of abdominal pain and discomfort over the last seven days. Thirty-six (73.5%) felt a sense of urgency for more than three days over the last seven days, and 18 (36.7%) experienced constipation over the last seven days.

Table 5.6 shows the percentages of respondents who reported any degree of problems in each of the EQ-5D dimensions by IBS-related health state. Even those respondents reporting current health state as P-U-C- reported moderate to extreme levels on the pain and mood dimensions. The mean EQ-5D tariff score for the sample was 0.66. This is lower than the average EQ-5D score obtained from a general population sample of 0.83 (Kind *et al*, 1998).

Figure 5.2 compares the mean SF-36 scores for the sample with those obtained by Garratt *et al* (1993) for their sample of 542 of the general population of Aberdeen. The Aberdeen sample ranged in age from 18 to 91 years (mean 47.9 years), and 53.9% were women. The means for the IBS sample range from 48.0 for energy/vitality to 77.7 for physical functioning. The range of values for the Aberdeen sample is 61.2 for energy/vitality to 79.2 for physical functioning. The SF-36 values are consistently lower for the IBS sample. This confirms the findings of Akehurst *et al* (2002), who compared quality of life questionnaires between 161 IBS patients and 213 members of a control group who did not suffer from IBS. Both groups were drawn from GP lists in Sheffield. Akehurst *et al* found that IBS patients scored lower for all SF-36 and EQ-5D dimensions than the control sample.

5.4.4 Logical consistency

The logical conditions for health states and profiles were examined both for the ranking tasks which respondents completed prior to the SG valuations, and the SG valuations themselves.

Ranking exercise

The number of violations for each logical consistency condition for the health states are shown in Table 5.7. All the violations involving P-U-C- are from one respondent. The number of violations for each condition is low. A total of 14 (29.2%) respondents demonstrated at least one violation of logical consistency in the health state ranking exercise.

Logical consistency conditions for the three health profiles C, D and E and the numbers of violations are shown in Table 5.8. One respondent showed one violation. Thus for this ranking exercise the sample was reasonably consistent overall.

Valuations

The logical consistency conditions in the valuation exercises are shown in Table 5.9, together with the number of violations. Table 5.9 refers to raw data. In other words, no chaining had taken place, and the QALY values for the profiles were not included. This exercise was simply to assess the level of understanding of respondents for the tasks.

There are very few violations of the consistency conditions relating to the health states and these rates are comparable with other studies (Dolan and Kind, 1996). It is noteworthy that there are more violations of consistency in the profile valuations. However, many of the inconsistencies were such that the valuation of the logically worse state or profile was within 0.05 of the valuation of the logically better state or profile. As previously discussed (see Section 5.3.6), 0.05 was considered to be the MEID. In addition, the inconsistencies were unrelated to respondent characteristics and no one respondent was inconsistent throughout. Furthermore, the mean valuations for each state and profile, and more importantly the differences between means, did not differ significantly when the inconsistent responses were excluded. For these reasons, it was decided that all the data should be included in the subsequent analysis.

5.4.5 Health state valuations

Table 5.10 shows the mean and median valuations given to the health states. The first two states, P-U+C- and P+U-C-, were evaluated against P-U-C- and P+U+C- (see Table 5.4). They therefore needed to be chained to death using equation (5.2). Both chained and non-chained values are shown. Since one respondent had missing data for health state P+U+C-, results were only produced for 48 respondents. It can be seen that the overall valuations follow a logical ordering. All mean and median valuations are significantly below the valuation for P-U-C- ($p < 0.01$) by both the paired t-test and Wilcoxon-sign test. Although the non-chained mean and median values for the states P-U+C- and P+U-C- suggest that the latter is valued higher, these differences are not significant by either the paired t-test or the Wilcoxon-sign test.

There is a noticeable variation in the spread of values around the means and medians for the different health states (Table 5.10). The mean non-chained valuations for the states of P-U+C- and P+U-C- have relatively large IQRs (ranges of 0.300 and 0.400 respectively) and SDs (0.23 and 0.21 respectively). This effect is hidden once the chaining procedure has been carried out. The IQRs for the chained health states have a range of approximately 0.01, and the SD for chained P+U-C- is 0.03 and for chained P-U+C- is 0.04. The health states P+U+C-, P+U-C+, and P+U+C+ did not require chaining because they were valued against death as a failure outcome (see Table 5.4). The IQRs for these states have ranges of 0.02, 0.027, and 0.037 respectively; the SDs for these states are 0.10, 0.11, and 0.17. These data indicate that there was a high level of variability in the valuations of the two states which were valued against P+U+C- as a failure outcome. The level of variability in the three states which were valued against death as the failure outcome was considerably lower.

5.4.6 Valuations of health profiles

Table 5.11 shows the non-chained valuations for profiles C, D and E. Profile I is excluded, because it was evaluated against death and was therefore not to the same scale. For both mean and median values, there is a decline in score from profile C through to profile E. This follows the logical ranking order for these profiles. However, the holistic values for profiles C, D, and E were not significantly different from each other with the exception of profiles C and E (p=0.049 according to the paired t-test, and p=0.02 according to the Wilcoxon-sign test). The paired t-test comparison between the QALY values of the profiles could not be performed, because the standard error of the difference was zero (see below). However, the results were significantly different according to the Wilcoxon-sign test for all comparisons between the three profiles (p=0.000).

The standard deviations around the means were far larger for the holistic method than the QALY method, as were the interquartile ranges around the medians (Table 5.11). The SDs and IQRs for the non-chained holistic profile valuations are comparable to those of the non-chained health states (Table 5.10), indicating a relatively large variability of responses. The SDs for the QALY valuations of the profiles are all 0.06, and the ranges of IQR are approximately 0.11 (Table 5.11). The fact that the SDs for the QALY valuations are so small could explain the finding that the standard error of the differences were zero when attempting to perform parametric statistical comparisons between the QALY valuations of the different profiles (see above) (Rowntree, 1991). However, the question arises as to why, when the variability in responses to the health state valuations are so large (Table 5.10), the apparent variability in the QALY valuations of the profiles are so small (Table 5.11).

A possible explanation for this finding would relate to individual valuations of the health states which were entered into the algorithm to obtain QALY values for the health profiles. For the non-chained QALY valuations of the health profiles, the valuations of the two states of P-U+C- and P+U-C- were entered into the QALY algorithm (the state P+U+C-, against which these two states were valued as a failure outcome, was given a value of zero). There are four possible ways an individual might have valued these two health states:

1) high values to both
2) low values to both
3) high value to P-U+C- and low value to P+U-C-
4) low value to P-U+C- and high value to P+U-C-

The large IQRs and SDs in Table 5.10 have already shown that the variability of responses for both health states was relatively high. As already stated, QALY values for the profiles were calculated from individual values for the health states. Two weeks were spent in each of the two health states P+U-C- and P-U+C- over a 12-week period. For respondents in categories (1) and (2), QALY values for the profiles would have been high or low respectively. However, QALY profile values would have been more similar for respondents in categories (3) or (4), and similar numbers in each category may have concealed the variability of their health state values in the QALY values for the profiles. If there were a high proportion of respondents in categories (3) and (4), this would explain the low IQRs and SDs in Table 5.11. The definitions of "high" and "low" have been set at greater than 0.5 and less than 0.5 respectively. A total of 13 (27%) respondents fitted into categories (3) or (4), with four having higher values for P-U+C-, and eight having higher values for P+U-C-.

Table 5.12 and Figure 5.3 compare the chained data for the holistically elicited profile valuations (including profile I) with the number of QALYs associated with those same profiles, as calculated using the valuations for the discrete health states.

Whereas the mean QALY results for profiles C, D and E decrease linearly, the mean holistic results are similar for all the profiles. These results are reflected by the median valuations. This suggests that, when valuing profiles by the holistic method, the proportion of time in each health state for each profile is not a predictor of the valuation outcomes.

The QALY valuation of C is higher than the holistic valuation. However, this is not statistically significant. Both methods provide similar values for profile D. Unsurprisingly, the t-test finds no significant differences between valuations from the two methods. However, the Wilcoxon-sign test finds the difference to be highly significant (p = 0.008). For profile E the holistic valuation is significantly higher than the QALY valuation according to both statistical tests.

Profile I is not included in Figure 5.3 because it includes two weeks of constipation. Apart from for the inclusion of constipation it is identical to profile C. According to the QALY method the value of profile I is 0.982, around halfway between the valuations for profiles C and D. However, the holistic valuation for profile I is 0.937. This is a great deal lower than the holistic valuations for profiles C, D, and E. This finding indicates that the presence of constipation may have a significant detrimental affect on

HRQoL. The differences between QALY and holistic valuations for profile I are significant according to both the t-test and the Wilcoxon-sign test.

It is notable that the results of the t-test and Wilcoxon signed-rank tests differ widely for profiles C and D. The t-test looks for differences in the means, whereas the Wilcoxon signed-rank test looks for differences in the distribution, taking into account both the sign of differences (*i.e.* whether positive or negative) and the magnitude of such differences. Thus the differences in Table 5.12 could be due to the differences between the two statistical tests. Histograms of the distributions of the chained holistic and QALY valuations are shown in Figures 5.A.1 to 5.A.8 in Appendix 1. The distributions for the QALY and holistic valuations of profiles C and D appear no more different from each other than for the other profiles.

The data is extremely skewed because of the nature of the chaining process, which compresses the data into the top part of the 0-1 scale. To discount this factor, the natural logarithm was applied to the data. The t-test and Wilcoxon signed-rank test were applied to these versions of the data. However, they still gave different results. Another way of looking at the data is to use Bland-Allman plots, which indicate how much the difference between means differs from zero. The holistic values are subtracted from QALY values on the y axis, and if there is no difference the value would be zero. The average value from the two methods is plotted on the x axis. Bland Allman plots are shown for the chained valuations of the four profiles in Figures 5.A.9 to 5.A.12. The results reflect the figures in Table 5.12, with many of the values clustering around zero for profiles C and D, and more QALY values below zero for profile E. QALY values tend to be above zero for profile I.

A similar situation arose in Chapter 7, with the results of t-tests and Wilcoxon-sign tests differing. The roots of these differences are explored in some detail in Section 7.10.9, and that discussion will not be repeated in full here. However, in summary the differences appear to be due to the two tests measuring different things and examining different aspects of the data (Altman, 1991). In samples consisting of over 30 individuals it is appropriate to use parametric tests even though the data may be non-normal in distribution. The t-test is a stronger test than the Wilcoxon-sign (Lowry, 1999-2005), and it produces 95% confidence intervals. These factors make it the preferable of the two tests.

5.4.7 Convergent validity

The level of convergency across the sample for the two valuation methods is reported in Table 5.13, which shows the number of pairs of profiles out of the three possible pairs showing convergency. For the QALY valuations at the level of strong convergency, most of the results were distributed across the 2/3 and 3/3 levels of convergency. A total of 29 (60.4%) respondents showed convergency for 2/3 pairs, and a further 18 (37.5%) showed convergency for 3/3 pairs of profiles. One (2.1%) respondent showed convergency for just 1/3 pairs, and there were no respondents with no convergency across pairs of profiles for the QALY method. At the level of weak convergency the results were the same for the QALY.

The distribution of degree of convergency is slightly different for the holistic method of valuation at the strong level. Most respondents were distributed across the 1/3 and 2/3 levels of convergency (see Table 5.13). A total of 25 (52.1%) respondents showed convergency for 1/3 pairs of profiles, and 22 (45.8%) showed convergency for 2/3 pairs of profiles. One (2.1%) respondent showed no convergency at all across the profiles (0/3 pairs), and no respondents showed full convergency (3/3 pairs). However, these figures improved at the weak level of convergency, for which only eight (16.7%) respondents showed convergency for 1/3 pairs, 38 (79.2%) showed convergency for 2/3 pairs, and one respondent (2.1%) showed convergency for all three pairs.

These results demonstrate a higher level of convergent validity for the QALY method than the holistic method overall.

5.4.8 Logical consistency

All pairwise comparisons are strongly consistent for the QALY method (Table 5.14). The following discussion will therefore pertain to the results for the holistic method. Overall there was a low level of strong consistency for the health profiles, with just 10 (20.8%) respondents showing strong consistency over all three profiles. A further seven (14.6%) valued all the profiles equally. Table 5.14 shows the levels of logical consistency for the QALY and holistic comparisons between pairs of profiles. For the comparison between profiles C and D, 21 (43.8%) were strongly consistent ($C \succ D$). However, this rose to 39 (81.3%) when tested for weaker consistency ($C \geq D$). A total of 9 (18.8%) respondents showed inconsistency in their orderings of the profiles ($C \prec D$). For the comparisons between profiles D and E, a total of 21 (43.8%) respondents were strongly consistent ($D \succ E$) for their holistic valuations. However, this figure rose to 36 (75%) for weaker consistency (C $\geq$ D). A total of 12 (25%) were inconsistent in their holistic valuations (D $\prec$ E). As for comparisons between profiles C and E, 27 (56.3%) were strongly consistent (C $\succ$ E), rising to 37 (77.1%) for weak consistency (C $\geq$ E). A total of 11 (22.9%) were inconsistent (C $\prec$ E).

Despite the above evidence that the QALY provides greater consistency in terms of logical ordering than the holistic method, upon examination of the responses it was found that many of the responses were inconsistent by a very small amount. The degree of inconsistency for holistic valuations ranged from 0.001 to 0.118. Nine responses were inconsistent by less than 0.01. Three respondents were inconsistent by less than 0.05. Thus most of the inconsistencies were within the MEID.

5.4.9 Respondents' comments

Patients were given the opportunity to write their comments at the end of the interview. These are reproduced in Table 5.A.3. Fourteen (29.2%) provided comments. These could be categorised as follows:

a) Problems with death as failure state (5)

b) Good descriptions of IBS (6)

c) Response may depend on current health or state of mind (4)

d) Not all relevant symptoms were represented (1)

e) Gamble approach difficult to understand (2)

5.5 Discussion

This study used IBS-specific health states to obtain QALY values for IBS profiles, and explore possible weaknesses in the QALY model for dealing with this kind of health profile. This study also explored the issues surrounding the construction of IBS health profiles for holistic valuation.

QALY valuations tend to be *ex post* rather than *ex ante*. The profile is described in terms of health states, which the health economist then builds up into health profiles. If health profiles are valued holistically, they can be valued eith *ex post* or *ex ante*. The health profiles in this study were valued in the *ex post* by the QALY, and *ex ante* by the holistic method. The holistic method asked respondents to value profiles of health in which the sequence of states was unknown, and only the proportion of time in each state was known. According to the traditional QALY model, the sequence does not matter. Because of the *ex ante* nature of the health profiles as valued holistically, much is left to the imagination of the respondent in the detail of each profile. For this reason, it is possible that each respondent could imagine the profiles to play out quite differently to the other respondents. However, this is realistic, because the profiles are likely to differ between IBS patients in real life.

5.5.1 Strengths and weakness of the study

Chapter 4 discusses the ideal way of setting up a valuation study. The construction of health states and profiles should ideally be based upon discussions and focus groups with health professionals and patients. However, in the case of this study the symptoms to be included had been previously decided by GW. It is not known upon what their decisions were based. However, the pilot group of patients found the descriptions used in the health states and profiles realistic and typical of IBS.

The IBS patients used in this study may not have been representative of IBS patients. One of the inclusion criteria was that respondents must have taken part in at least one of the GW trials. The method of obtaining patients for their trials was not divulged by GW. However, it was stated that they attempted to exclude constipation sufferers from their trials. Although this attempt was not altogether successful, this is evidence that these patients were a select group. The amount of information from GW about their reasons for choosing such a select group of patients was limited, for obvious reasons of commercial confidentiality. One reason may have been due to the nature of the drug they were testing in their trials. It may have been that the drug was targeted at sufferers who suffered certain symptoms and who did not have significant levels of constipation, in which case they may have been justified in choosing a sample who were representative of the patients they were aiming to treat rather than representative of IBS sufferers more generally. The degree to which the lack of representativeness of the sample is of importance depends on the uses to which the resulting valuation data were to be put. The objectives of the study described here were to test the assumptions of the QALY algorithm, explore methods of constructing and obtaining values for holistic profiles for IBS, and provide valuations for the economic model being used by GW. The fact that the respondents may not have been representative of IBS patients does not matter to the first and second objectives, because this study was methodological in nature (although a more representative group may have given different responses). It is impossible to comment on the degree to which it matters to the third objective without further information about the purposes for which they intended to use the data.

In the latter part of the questionnaire the health profiles that had a logical ranking order were presented in the order of preferability. Thus the best was presented first and the worst last. Ordering effects are well-known. The way this was done introduced the possibility of framing effects, such that individuals may have been more likely to rate later profiles lower due to seeing them as a loss (Bernstein *et al*, 1997, 1999; Kahneman and Tversky, 1979, 1982; Tversky and Kahneman, 1981, 1986). Ideally, a larger sample would have been recruited and divided into sub-samples, each of which would have been provided with a different ordering. In this way it would have been possible to analyse the effects of ordering on valuations of profiles. However, as discussed in Chapter 4, difficulties were encountered throughout the PhD in obtaining large samples. In the case of this particular study, GW was responsible for patient recruitment, and they specifically requested a sample of around 50.

The ordering of profiles from good to bad may have given rise to the possibility of leading respondents' in their valuations by framing effects. However, it allowed the differences between the profiles to be clearly seen by respondents. According to Nord (1992) and Shiell *et al* (1997), being very open to respondents about the implications of their responses and enabling them to understand the implications of the task as clearly as possible may actually assist researchers to elicit true preferences. This could include allowing respondents to clearly see the connections between different valuation questions, as was done in this study.

The SG tasks did not have death as one of the reference states for five out of the nine valuations, so these health states and profiles had to be chained to death via a reference health state. The valuation task needed to be realistic for the respondents to take them seriously, and since death is not a significant factor in IBS it would not have seemed realistic or relevant to gamble against death as a failure state for most of the states and profiles. Chaining through the reference states of P-U- and P+U+ had a second advantage of increasing the sensitivity of the scale. Given that this study addressed issues around the derivation of QALYs, where zero represents states regarded as equivalent to death and one states in good health, it was necessary to chain the values using the valuation of P+U+. This may be criticised due to inconsistency between chained and non-chained valuations (Ubel, 1999). However, the analysis was also conducted using unchained values and the same relationship was found between holistic and QALY profile values.

Another possible weakness of the study design was in the way constipation was valued. One profile (I) incorporated constipation. Profile I was the same as profile C in that it included 6 weeks in P-U- and 2 weeks in P+U+. However, profile I stated that two weeks would be spent in the state of constipation. This might occur with any of the states of P+U+, P+U-, P-U+, or P-U-. This did not present a problem to the holistic valuation of this profile. However, in order to apply the QALY algorithm to profile I, it was necessary to assume additivity in assigning the two weeks of constipation. The two states of P+U-C+ and P+U+C+ were evaluated in the earlier part of the questionnaire, and it was assumed that profile I would contain these states. Thus for the QALY application it was assumed that profile I comprised the 5 weeks in P+U+C-, 1 week in P+U+C+, 2 weeks in P-U+C-, 1 week in P+U-C-, 1 week in P+U-C+, and 6 weeks in P-U-C-. If the assumption of additivity is incorrect, it may matter with which symptoms constipation occurs. An alternative method of dealing with constipation would have been to add constipation to a specific health state in the profiles, so that when valued holistically they were valued as closely to the QALY as possible.

The opportunity for respondents to make their comments provided an insight into the degree of challenge of the cognitive task of completing the valuation exercises. Detailed comments are recorded in Table 5.A.3 in Appendix 1. There is no sense that the tasks proved too cognitively difficult for the majority of this sample. Only two respondents commented that they had difficulty. One respondent state that she found “the gamble approach difficult to understand”, and another respondent said that it “... would be hard to totally understand what was being asked just by reading the questions as they all seem to read very similar...” and that it “...was helpful to have someone with us to help and guide us”. Only 14 (29%) wrote comments, so this cannot be taken as representative of the opinions of the whole sample. The same proportion (29%) of the sample made a minimum of one violation of logical consistency in the health state ranking exercise. It is possible that this could indicate cognitive difficulty with the health states, but it is also possible that respondents used the ranking exercise to familarise themselves with the health states.

5.5.2 Health state valuations

The results suggest that the health states associated with IBS result in a small yet significant decrement in HRQoL as compared to the symptom-free health state. Table 5.10 shows that the IBS health states in this study were valued lower than the P-U-C- state by 0.014 to 0.068. When compared to the SF-6D values in Table 5.5, the states containing urgency are below the symptom-free state by a similar amount in both Tables 5.5 and 5.10 (differences of P-U+C- from P-U-C- of 0.022 for the SF-6D values and 0.018 for this valuation study, and differences of P+U+C- from P-U-C- of 0.041 for the SF-6D values and 0.048 for this valuation study). However, the presence of pain and discomfort is seen to be worse when using the SF-6D index (differences of P+U-C- from P-U-C- of 0.031 versus 0.014 for the valuation study). For constipation, the SF-6D values were based on comparatively small numbers and did not show a significant impact on the IBS health states. In contrast, it appears from the SG valuations that constipation had an important effect on HRQoL. When constipation is added to P+C-, the valuations differ from P-U-C- by 0.049 as oppose to 0.014 (Table 5.10). When constipation is added to the state of P+U+, mean valuations decline by 0.02.

It is also worth noting that the IBS symptom-free state has an SF-6D score of 0.896 (Table 5.5), thus indicating that P-U-C- is somewhat below the state of full health. This suggests that the IBS health states are not picking up all of the factors of ill health (which may or may not be IBS-related) amongst these patients. As demonstrated in Table 5.6, there are reductions in health as defined by the dimensions of the EQ-5D. For each IBS health state (including P-U-C-), a significant number of patients report moderate or extreme problems on the pain and mood dimensions. As discussed in Section 5.5.1, the author was not party to the decision process by which GW chose which symptoms to include in their model. However, the pilot study indicated that the patients felt that the health state descriptions were accurate.

5.5.3 QALYs versus holistic valuations

This study offers an important insight into the way people value conditions such as IBS, where the condition does not follow a clear course of progression. Previous work in the economics literature on valuing time profiles of health has tended to focus on sequences of states. IBS is characterised by two distinct features, which have not been examined in previous work on valuing profiles. The first is that IBS is subject to considerable and unpredictable fluctuations in its symptoms, and hence HRQoL varies over relatively short periods. Yet at the same time it is a condition that can last for many years.

A method was developed in this study for summarising the patient's health experience using IBS health states in terms of the proportion of time spent in each state over a period of 12 weeks, which would be repeated for the remainder of the respondent's life. The time spent in each health state was presented as being distributed in an unpredictable way over each period of 12 weeks. This method allows the outcomes of treatment to be compared in terms of the proportion of time spent in each state. IBS is not the only condition with these features and the results presented in this chapter have implications for other conditions (such as depression, for example).

Whereas the QALY valuations of Profiles C, D and E showed a steady decline in valuations as the number of weeks in P+U+C- increased, there was no such decline according to the holistic valuations of these profiles (Table 5.12, Figure 5.3). These results indicate that the holistic valuations of IBS health profiles differ significantly from those implied by the QALY assumption of linearity over time. The main reason for this seemed to be that the holistic valuation did not vary as the proportion of time spent in the worst IBS state (P+U+) changed from two to six weeks (out of 12). Thus rather than "duration neglect" (Ariely, 1998), the patients in this sample were demonstrating "proportion of time neglect".

Although this result could have been an artefact of the high levels of inconsistency in the holistic valuations of the profiles, there was a clear break during the valuation task between the valuation of states and profiles, at which the interviewer pointed out the nature of the differences between the profiles and the respondents were asked to rank them. This allowed respondents the opportunity to become familiar with the health profiles. As shown in Table 5.8, there was only one violation of logical consistency in the ranking exercise for health profiles. This is very different from the results of the tests of logical consistency of the valuations (Tables 5.9 and 5.14). One possible explanation for the high rates of logical inconsistency of valuations could be a lack of ability to deal with the task cognitively. The direct ranking of the health profiles rated very well in terms of logical consistency (Table 5.8) compared to that of the health states (Table 5.7). In order to rank the health states, respondents had to compare different combinations of symptoms and declare their preferences over these combinations. The health profiles ranking task may have been less demanding cognitively, despite the apparent complexity of the profiles. The profiles were all the same in terms of the combinations of symptoms they contained. What differed between the profiles was the proportion of time in each health state. This was clearly set out in the profile descriptions, and respondents could read the descriptions at the same time and make direct comparisons. However, it is possible that the standard gamble task in combination with the complexity of these health profiles proved too difficult for the respondents. Other possible explanations are discussed in the following paragraphs.

The findings of insensitivity to scope have echoes in the contingent valuation (CV) literature (Baron, 1997). Most of the explanations for this finding in the CV literature are concerned with public goods rather than potential private benefits such as one`s own health. Another possible explanation could be a threshold effect, whereby after some proportion of time spent in a particular state further increases have little additional impact (Propper, 1990). As discussed above, it is also possible that this insensitivity to proportion of time in P+U+ is due to the profiles being too difficult to visualize. Thus patients may have used heuristics to judge them, such as "I will feel bad most of the time" or "I will feel good most of the time".

There are also possible explanations from the psychological literature. Ariely and Zauberman (2000) addressed issues around valuing a whole profile versus valuing various segments of that profile. They showed that the overall rating of an entire profile of continuous annoying sounds was based on the what the authors called "gestalt" characteristics, which were specific characteristics of the sequence such as peak intensity, final intensity, *etc.* However, when the profile was broken into segments of sound which were rated separately, the mean value of each segment became more important in the overall valuation of the profile. A similar experiment involving simulated stock performance over time indicated that getting respondents to provide momentary valuations throughout a continuous sequence acts to segment the profile so that the overall evaluation provided at the completion of the sequence is based on the mean value rather than gestalt characteristics. This work could go some way towards explaining the difference between QALY and holistic methods of valuing IBS health profiles. The QALY values are obtained by segmenting the profiles into their constituent health states whereas the holistic valuations may be based on gestalt characteristics (which for the IBS profile could, for example, be the worst health state rather than the proportion of time they could expect to be in the state). For valuations by the QALY method a good segment would be valued highly, and a bad segmented valued lower.

5.5.4 *Implications of this research*

It has been argued by some commentators that the holistic valuation of profiles is superior to the QALY algorithm since it does not assume people's preferences are linear over time (Mehrez and Gafni, 1993). This argument takes no account of the cognitive difficulty associated with the valuation of profiles compared to health states. The psychology literature suggests that respondents in this circumstance may use simplifying heuristics to overcome the cognitive burden, such as concentrating on the state rather than considering the time spent in the state. For the purpose of valuing benefits of health care, this has important implications for economic evaluation and the valuation of benefits for conditions such as IBS where symptoms are subject to considerable fluctuation over time. The QALY approach may provide a more sensitive method for valuing the benefits of treatments for this type of condition.

As indicated above, the QALY values were a direct reflection of the proportion of time in each health state for each profile of health. The holistic valuations of the profiles did not reflect the proportion of time in each health state. This study has not allowed a definitive conclusion that one method is better than the other. However, the results from the two methods could lead to very different conclusions. If the holistic results are a true reflection of preferences, this could mean that the QALY underestimates HRQoL in IBS. However, if the reverse is true the HRQoL associated with IBS could be overestimated.

5.5.5 *Further research*

Figure 5.3 plots holistic and QALY valuations of the profiles. For the QALY valuations the plot incorporates the entire range of time in P+U+ from 0 to 12 weeks. However, for the holistic valuations only a small part of the plot is included, namely P+U+ over 2, 4 and 6 weeks. The proportion of time neglect demonstrated by the results of the holistic valuations was an unexpected result. It is unclear what the holistic results would have been for P+U+ between 6 and 12 weeks. A follow-up to this study is reported in the next chapter, which will seek to determine whether there is indeed a threshold effect operating by obtaining valuations for profiles in which P+U+ occurs in totals ranging from two to 12 weeks. It would be expected that 12 weeks in a profile containing P+U+ would be of equal value to the health state P+U+ occurring over 12 weeks. One possible scenario is that the plot would follow that in Figure 5.3 until the profile P+U+ was compared to the state P+U+, and then the non-chained value would drop to zero.

The follow-up study in Chapter 6 will comprise an empirical exploration of holistic valuations with P+U+ occurring over the whole range of 0 to 12 weeks. However, if the results from the follow-up study are similar to this study in terms of the lack of sensitivity of the holistic method to proportion of time in each health state, additional research will be required into the underlying reasons for the differences between holistic valuations and QALY valuations of the same IBS health profiles. Such research would probably require the interviewing of a small sample of IBS patients, using qualitative techniques to draw out possible explanations for the findings.

5.6 Conclusions

This study succeeded in its objective to generate a set of useable IBS-related health state valuations relating to health states containing varying combinations of abdominal pain and discomfort, urgency, and constipation. These symptoms were prevalent in the study sample.

The study compared different approaches to obtaining health profile valuations, and found that valuations obtained using the QALY algorithm differed from valuations obtained using a holistic method. These findings indicate that the additive utility function may not apply in the case of IBS. The results of this suggest that HRQoL of IBS sufferers may not be so low as suggested by the QALY algorithm. However, it is possible that the holistic valuations were too cognitively demanding for respondents. who demonstrated a "proportion of time" neglect. The QALY algorithm produced a higher level of logical consistency.

The next chapter describes a study to determine how profiles with P+U+ occurring over the entire range of 0 to 12 weeks are valued by the holistic method.

Symptoms (Rome criteria)

1) Abdominal pain or discomfort, relieved with defecation, or associated with a change in frequency or consistency of stools
2) Irregular (varying) pattern of defecation at least 25% of the time (three or more of):
    - Altered stool frequency
    - Altered stool form (hard or loose/watery stool)
    - Altered stool passage (straining or urgency, feeling of incomplete evacuation)
    - Passage of mucous

Bloating or feeling of abdominal distension

Figure 5.1 The Rome criteria for diagnosis of IBS.

Figure 5.2 Mean SF-36 scores for the IBS sample and a UK sample.

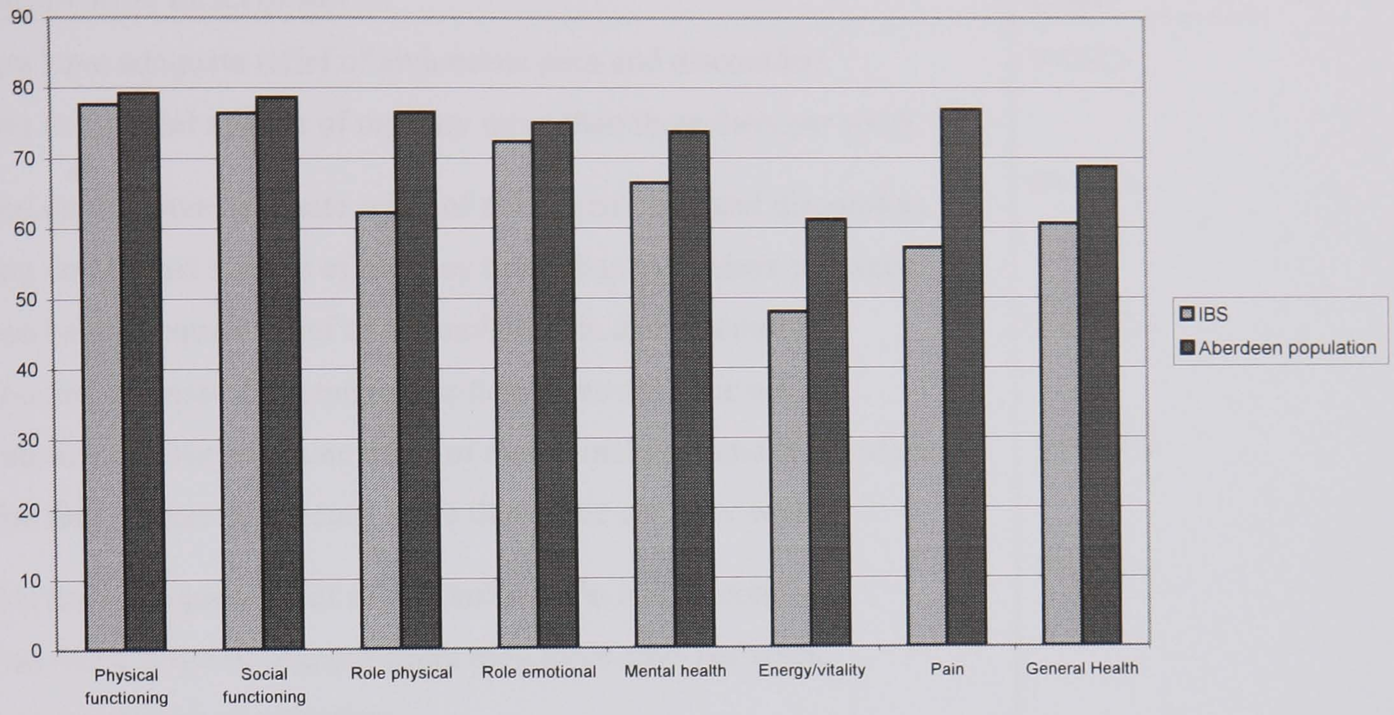


Figure 5.3: Comparison of mean chained holistic and QALY profile values

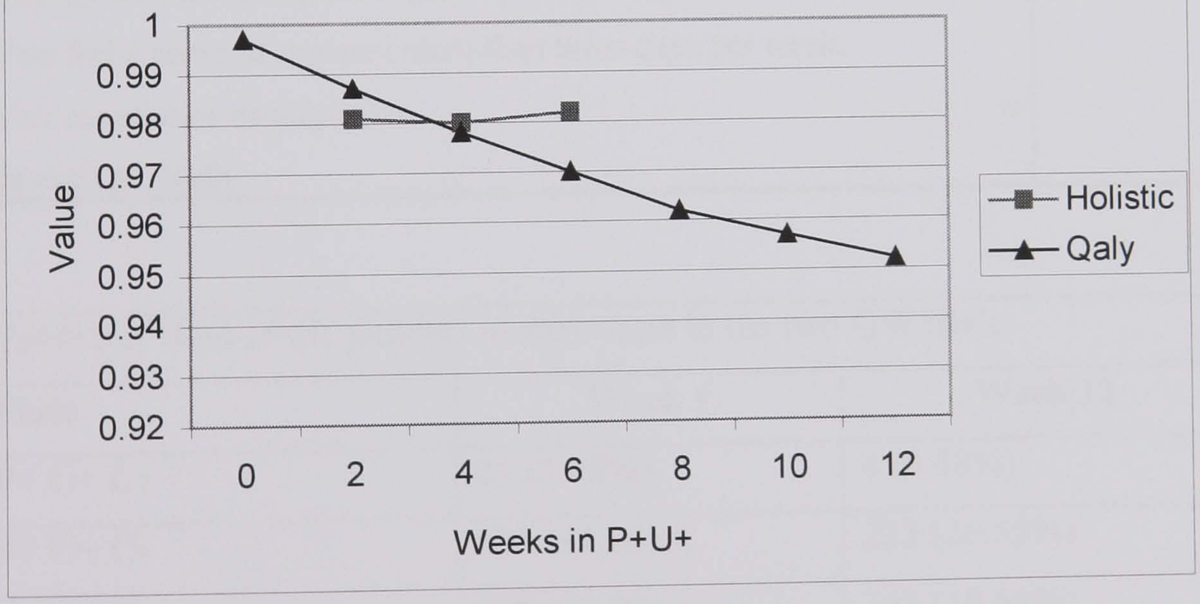


Table 5.1: The IBS health states

Health state description	*Code*
You have adequate relief of abdominal pain and discomfort. You do not feel a sense of urgency more than three days per week.	P-U-C-
You do not have adequate relief of abdominal pain and discomfort. You do not feel a sense of urgency more than three days per week.	P+U-C-
You have adequate relief of abdominal pain and discomfort. You feel a sense of urgency more than three days per week.	P-U+C-
You do not have adequate relief of abdominal pain and discomfort. You feel a sense of urgency more than three days per week.	P+U+C-
You have adequate relief of abdominal pain and discomfort. You feel a sense of urgency more than three days per week. You experience constipation.	P-U+C+
You have adequate relief of abdominal pain and discomfort. You do not feel a sense of urgency more than three days per week. You experience constipation.	P-U-C+
You do not have adequate relief of abdominal pain and discomfort. You do not feel a sense of urgency more than three days per week. You experience constipation.	P+U-C+
You do not have adequate relief of abdominal pain and discomfort. You feel a sense of urgency more than three days per week. You experience constipation.	P+U+C+
Immediate death	

Table 5.2 Sum of the patients in each state in the two GW trials.

State	**Week 1**	**Week 12**
P+ U+ C+	25 (2.08%)	4 (0.48%)
P+ U+ C-	501 (41.72%)	223 (26.55%)
P+ U- C-	241 (20.07%)	165 (19.64%)
P+ U- C+	20 (1.67%)	5 (0.60%)
P- U- C+	15 (1.25%)	10 (1.19%)
P- U+ C-	175 (14.57%)	114 (13.57%)
P- U+ C+	8 (0.67%)	0 (0%)
P- U- C-	216 (17.99%)	319 (37.98%)
Total	**1201 (100%)**	**840 (100%)**

Table 5.3: The health profiles

	Number of weeks (out of 12) spent in each state in profile:			
Health states	C	D	E	I
P+U+C-	2	4	6	2
P+U-C-	2	2	2	2
P-U+C-	2	2	2	2
P-U-C-	6	4	2	6
Constipation	-	-	-	2

Table 5.4: The standard gamble questions

Gamble	State/Profile	Success outcome	Failure outcome
1	P-U+C-	P-U-C-	P+U+C-
2	P+U-C-	P-U-C-	P+U+C-
3	P+U+C-	P-U-C-	Death
4	P+U-C+	P-U-C-	Death
5	P+U+C+	P-U-C-	Death
6	Profile C	P-U-C-	P+U+C-
7	Profile D	P-U-C-	P+U+C-
8	Profile E	P-U-C-	P+U+C-
9	Profile I	P-U-C-	Death

Table 5.5 SF-6D index values.

Health state	N	Mean value	SD	Mean difference from P-U-C-
P-U-C-	247	0.896	0.073	
P-U+C-	133	0.874	0.094	0.022*
P+U-C-	209	0.865	0.091	0.031*
P+U+C-	366	0.855	0.094	0.041*
C+	43	0.875	0.731	0.021

* difference is significantly different from the value of P-U-C- at p<0.05

Table 5.6 Percentage of respondents reporting levels 2 or 3 on the EQ-5D by current IBS health state

States	N	Mobility	Self-care	Usual activities	Pain/ discomfort	Mood
P-U-C-	6	0.0	0.0	0.0	66.7	33.3
P-U+C-	7	14.3	0.0	0.0	71.4	28.6
P+U-C-	3	0.0	0.0	0.0	66.7	33.3
P-U-C+	1	0.0	0.0	100.0	0.0	0.0
P+U+C-	15	20.0	0.0	40.0	93.3	33.3
P+U-C+	3	0.0	0.0	0.0	100.0	66.7
P-U+C+	3	33.3	33.3	66.7	100.0	66.7
P+U+C+	11	54.5	18.2	63.6	100.0	63.6

Table 5.7 Number of violations for the health state ranking exercise.

Consistency condition	No. of violations
P-U-C- > P+U-C-	1
P-U-C- > P-U+C-	1
P-U-C- > P+U-C+	1
P-U-C- > P+U+C-	1
P-U-C- > P+U+C+	1
P+U-C- > P+U+C-	3
P-U+C- > P+U+C-	3
P+U-C- > P+U-C+	4
P+U+C- > P+U+C+	4
P+U-C- > P+U+C+	1
P-U+C- > P+U+C+	2
P+U-C+ > P+U+C+	1

Table 5.8 Numbers of violations for the health profile ranking exercise.

Consistency condition	No. of violations
Profile C > Profile D	1
Profile C > Profile E	0
Profile D > Profile E	0

Table 5.9 Logical inconsistencies of valuations

Consistency condition	No. (%) of violations
P+U+C- > P+U+C+	5 (10)
P+U-C+ > P+U+C+	4 (8)
Profile C > Profile D	9 (18)
Profile C > Profile E	11 (23)
Profile D > Profile E	12 (25)

Table 5.10 Health state valuations. States P-U+C- and P+U-C- were chained to the 0 to 1 scale, where 0 is death and 1 is P-U-C-. Both non-chained and chained values are shown.

Health state	N	Median (IQR)	Mean (SD)	Mean difference from P-U-C-
P-U+C-	48	0.997 (0.987-0.999)	0.982 (0.04)	0.018 **
P+U-C-	48	0.997 (0.989-0.999)	0.986 (0.03)	0.014 **
P+U+C-	48	0.995 (0.975-0.995)	0.952 (0.10)	0.048 **
P+U-C+	48	0.995 (0.968-0.995)	0.951 (0.11)	0.049 **
P+U+C+	48	0.985 (0.958-0.995)	0.932 (0.17)	0.068 **
Non-chained values				
P-U+C-	48	0.675 (0.475-0.775)	0.629 (0.23)	0.371 **
P+U-C-	48	0.700 (0.475-0.875)	0.659 (0.21)	0.341 **

** Significant by both t-test and Wilcoxon-sign test at $p < 0.01$.

Table 5.11 The unchained holistic and QALY valuations for profiles C, D, and E.

Profile	N	Median (IQR)		Mean (SD)	
		Holistic	QALY	Holistic	QALY
C	48	0.775 (0.525-0.875)	0.711 (0.660-0.767)	0.719 (0.20)	0.715 (0.06)
D	48	0.725 (0.475-0.875)	0.544 (0.494-0.600)	0.689 (0.21)	0.548 (0.06)
E	48	0.700 (0.475-0.825)	0.378 (0.327-0.433)	0.663 (0.21)	0.381 (0.06)

Table 5.12 Holistic profile valuations and implied number of QALYs chained through P+U+C-.

Profile	N	Median (IQR)		Mean (SD)		Mean holistic – QALY (95% CIs)	Mean difference from P-U-C-		Holistic versus QALY	
		Holistic	QALY	Holistic	QALY		Holistic	QALY	*p* (t-test)	*p* (Wilcoxon)
C	48	0.997 (0.99-0.999)	0.998 (0.992-0.999)	0.981 (0.05)	0.987 (0.03)	-0.005 +/- 0.006 (-0.011 to 0.001)	0.019	0.013	0.076	0.928
D	48	0.997 (0.992-0.999)	0.997 (0.987-0.998)	0.980 (0.05)	0.978 (0.04)	0.001 +/- 0.003 (-0.002 to 0.004)	0.020	0.022	0.513	0.008
E	48	0.997 (0.991-0.999)	0.996 (0.983-0.997)	0.982 (0.04)	0.970 (0.06)	0.012 +/- 0.007 (0.005 to 0.019)	0.018	0.030	0.002	0.000
I	48	0.985 (0.975-0.995)	0.997 (0.990-0.998)	0.937 (0.16)	0.982 (0.04)	-0.045 +/- 0.038 (-0.083 to –0.007)	0.063	0.018	0.021	0.000

Table 5.13 Degree of convergence between the QALY and holistic valuations and the original ranking of the profiles C, D, E, and I. The number (%) of pairs of profiles showing strong and weak convergency is shown.

	Strong				Weak			
	QALY		Holistic		QALY		Holistic	
	n	%	n	%	n	%	n	%
0/3 pairs	0	0	1	2.1	0	0	1	2.1
1/3 pairs	1	2.1	25	52.1	1	2.1	8	16.7
2/3 pairs	29	60.4	22	45.8	29	60.4	38	79.2
3/3 pairs	18	37.5	0	0	18	37.5	1	2.1

Table 5.14 Logical consistency for chained QALY and holistic valuations of health profiles C, D and E.

		D		E	
		QALY	Holistic	QALY	Holistic
C	$\succ$	48 (100.0%)	21 (43.8%)	48 (100.0%)	27 (56.3%)
	=	0 (0.0%)	18 (37.5%)	0 (0.0%)	10 (20.8%)
	$\prec$	0 (0.0%)	9 (18.8%)	0 (0.0%)	11 (22.9%)
D	$\succ$			48 (100.0%)	21 (43.8%)
	=			0 (0.0%)	15 (31.3%)
	$\prec$			0 (0.0%)	12 (25.0%)

Chapter 6

Study 2: A follow-up test of additive utility in health profiles relating to irritable bowel syndrome

6.1 Background

If, as implied by the QALY algorithm, utility is a linear function of time spent in each health state, we should expect there to be a linear relationship between the number of weeks spent in P+U+ and the overall valuations of the entire time period. However, in the previous study this was not found to be the case for the holistic valuations. The study found that holistic valuations of these profiles did not decline in utility as the number of weeks with symptoms increased. Rather, utility remained set at around 0.98.

In Study 1, the durations of 2, 4 and 6 weeks in P+U+ were not chosen to answer methodological issues. Rather they were chosen as described in Section 5.3.1, to represent the range of likely outcomes of treatment based upon the drug versus placebo data obtained from the GW clinical trials. The aim of this second study was to extend the duration spent in P+U+ over the whole range of durations from 0 to 12 weeks in order to examine the relationship between duration and utility over a higher frequency of P+U+, and to compare holistic and QALY valuations of these profiles. The expected outcome is that a profile consisting entirely of 12 weeks in P+U+ would be valued equally to the worse outcome of the gamble, the state of P+U+.

6.2 Methods

6.2.1 Construction of health states and profiles

The health profiles comprised the top four health states shown in Table 5.1, which were based on the presence or absence of pain (P) and urgency (U) as in Study 1. States with constipation were excluded from this follow-up study for the purposes of direct comparison with the results from Chapter 5.

Health states were valued by the same standard gamble procedures used in Chapter 5 in order to enter their valuations into the QALY algorithm to obtain QALY valuations for the profiles. States P+U- and P-U+ were valued against P+U+ as a worse state. State P+U+ was then valued against a worse state of immediate death.

Table 6.1 shows the possible combinations of these states into health profiles. The shaded profile (P-U- for 12 weeks) was not valued, because this is the anchor state and will therefore be given a value of 1. Due to the dichotomous nature of the urgency variable, people in P-U- may have some degree of urgency (three days per week or less). It is also possible that the health state descriptions do not capture every aspect of IBS, or that there are other comorbid conditions. As shown in Table 5.5, SF-6D data from 247 IBS sufferers in the state P-U-C- indicated a mean utility of 0.896. Thus it cannot be assumed that P-U- is equal to full health.

All the health profiles were rated against the states P-U- and P+U+ using the same standard gamble procedure as used in Chapter 5. The final profile presented in Table 6.1 (P+U+ for 12 weeks) was included as a split-test of reliability to test the understanding of the respondents for the valuation tasks. If the standard gamble procedure was fully understood, respondents should have rated this profile equal to the reference state P+U+ against which it was being valued.

Profiles C, D and E are the same as in Study 1.

6.2.2 The questionnaire

The questionnaire design was the same as that of Study 1. The version of the standard gamble devised by Jones-Lee *et al* (1993) was used. The questionnaire is presented in Appendix 2. It consisted of questions on:

- Background characteristics (age, sex, occupation, level of education, health rating, length of time with IBS)
- EQ-5D ratings
- Health status in terms of presence or absence of the symptoms which make up the states in Table 5.1, and also constipation
- A health state ranking procedure
- Standard gamble valuations of the health states
- A health profile ranking procedure
- Standard gamble valuations of the health profiles

- A section for patients' comments

6.2.3 The pilot

The questionnaire was piloted on three female IBS sufferers. Two were housewives aged 62 and 28. The other was a retired woman aged 69 years. They rated their health as "good", "fair", and "very good" respectively. Their highest levels of education were primary school, A-level, and secondary school respectively. They had all suffered from IBS for four to six years.

The respondents completed the questionnaire successfully. However, in contrast to Study 1, two of the respondents felt that the questionnaire did not fully describe IBS symptoms. For example, they thought that the pain/discomfort item was not described satisfactorily.

6.2.4 Recruitment and interviews

As was pointed out in Chapter 5, the sample obtained in the previous study may not have been representative of IBS patients in general. One of the aims of this study was to obtain valuations from a more representative sample of IBS sufferers. Previous research into IBS at the Institute of General Practice and Primary Care, the University of Sheffield, had involved surveys of IBS patients from general practices. The Institute had therefore forged contacts with general practices whose lists contained patients with IBS. The author approached staff at the Institute in order to discuss recruitment of IBS patients for this study using the contacts already available to the Institute.

A research protocol for the study was drawn up and submitted to the North Sheffield Local Research Ethics Committee. The author attended the meeting of the ethics committee, and ethical approval was obtained for this study.

Through the assistance of staff at the Institute of General Practice and Primary Care, details were obtained of seven GP practices in Sheffield, all of which had taken part in previous studies on IBS with the Institute. These practices covered a wide range of areas within Sheffield, such as Woodhouse, Meersbrook, Hillsborough, Highfield, Shirecliffe and other areas, allowing the possibility of access to patients from a range of different backgrounds.

The author visited each practice to obtain the permission of the practice manager to interview the IBS patients on their lists, but only those who had previously taken part in a study with the Institute of General Practice and Primary Care. Letters of invitation were sent from the practice manager on behalf of the partnership to the patients. The letters contained an information sheet, a consent form to be returned to the author so that she could contact the patients, and a statement that patients who agreed to be interviewed would receive a gift voucher. This was for £10. The contents of the letter are shown in Appendix 2.

A total of 183 patients were invited. Respondents were interviewed in groups of one to seven by a trained and experienced interviewer. Since the questionnaire did not require respondents to divulge personal or embarrassing information about themselves to the rest of the group, it was not considered necessary either to conduct one-to-one interviews or to segregate into groups of men and women.

6.2.5 Analysis

Background characteristics

The statistics for background characteristics were described in terms of age, duration of IBS, sex, employment, educational attainment.

Current health

Generic values for current health were calculated for the sample using EQ-5D and SF-36 scores.

Current health was also examined in terms of the IBS symptoms of abdominal pain or discomfort, degree of urgency, and constipation present in each member of the sample.

Exclusion criteria

A complete set of valuation data was required for the anlaysis. Respondents with missing valuation data were excluded from further analysis.

Logical consistency of ranking for health states and profiles

Each respondent was presented with two ranking exercises. In the first exercise they were requested to place a set of health state descriptions relating to IBS in order from most preferred to least preferred. There were conditions of logical consistency for the health states (Table 6.5). For example, it would seem illogical to prefer a state containing pain to a state containing no symptoms whatever. Responses to the health state ranking exercise were examined for logical consistency.

In the second ranking exercise they were required to place a set of IBS health profile descriptions in order of preference. The profiles all contained the same health states, but were different in terms of the frequency of each health state within the profile. There was a logical order of preference for the profiles, such that it would seem logical to rank them so that the one with the highest frequency of best health is preferred to the one with the highest frequency of worst health. Responses to the health profile ranking exercise were examined for logical consistency.

Health state valuations

Health state values were obtained for the health states of P+U- and P-U+ by the standard gamble method. These two states were valued against P-U- as the best outcome of the gamble and the worse state P+U+ as the failure outcome. The state P+U+ was then valued against immediate death as the failure outcome. In order to obtain values for states P+U- and P-U+ on a scale including death, the first two states were chained through the value for P+U+ using equation (5.2).

Health profile valuations

Health profiles were valued against P-U- as a success outcome and P+U+ as a failure outcome of the gamble. All the profile values had to be chained through the value for health state P+U+ in order to transfer the valuations to a scale ranging from death to no symptoms. The same applied also to the QALY valuations. The states of P+U- and P-U+ were chained through P+U+ using equation (5.2). Equation (5.2) was also used to chain health profile valuations through P+U+ to immediate death. The QALY algorithm was applied for each profile by multiplying the valuations for each discrete health state by the overall proportion of time spent in that state.

QALY and holistic health profile valuations were compared using the paired-sample t-test and Wilcoxon-sign test in order to determine whether there were significant differences in health profile values between the two methods. These two statistical tests were also used to determine whether values for each health profile were significantly different from other profile values for each valuation method.

Convergent validity

The degree of agreement between the original ranking order of the health profiles and the implied ranking of the health profiles from the QALY valuations was compared for each member of the sample at the level of strong and weak convergency. The same comparison was also carried out for the holistic valuations. In this study there were a total of eight health profiles, and therefore a total of seven pairwise comparisons between the profiles. Convergent validity was scored according to the number of pairwise comparisons out of a possible seven which were ranked the same between the original ranking exercise and the order implied by the valuation method. Any individual in the sample could have a level of convergency of zero out of seven pairwise comparisons to seven out of seven pairwise comparisons.

Logical consistency

A comparison of the extent to which holistic and QALY valuation rankings coincided with the logical rank orders of the health profiles was conducted in order to determine which method demonstrated the greater level of logical consistency. Logically, the higher the frequency of P+U+ and the lower the frequency of P-U-, the worse the health profile is. Pairwise comparisons were carried out over profiles A to H as described in Chapter 5.

Unwillingness to gamble

The valuation data were examined by individual respondent to determine the extent of unwillingness to gamble for health states and profiles. The above analysis was repeated with the exclusion of those respondents who were unwilling to gamble.

Respondents' comments

At the end of the questionnaire, respondents were given the opportunity to write their comments about any aspects of the issues arising during the interview. The comments were examined and categorised.

6.3 Results

6.3.1 Background characteristics

There were a total of 56 respondents to the main questionnaire study. a response rate of 30.6%. This low response rate was in part due to difficulties in arranging interviews. For example, some people worked during the day and were unable to come to the interviews during the evenings due to family commitments. It would have been useful to compare respondents with non-responders. Such comparisons could have been based on general demographics (*e.g.* gender, age, employment). It would also have been useful to determine what differences, if any, existed in terms of IBS symptoms between respondents and non-responders. However, such comparisons were impossible due to ethical constraints. The author was allowed access only to those patients who chose to respond and participate in the study.

The mean age of the sample was 49 years (median 50 years) with a range of 21 to 67 years (Table 6.A.1, Appendix 2). The mean age of the sample in Study 1 was 46.6 years (median 48). This sample had suffered from IBS for a mean of 13.9 years (median 11 years) with a range of two to 40 years (Table 6.A.1), compared to 12.1 (median 10) for Study 1. A total of 5 (8.9%) were men compared to 51 (91.1%) women.

A total of 33 (58.9%) were in paid employment. Details of occupational status are in Table 6.A.2 in Appendix 2. Seventeen (30.4%) were housewives/mothers. A further five (8.9%) were retired, and one (1.8%) was unemployed. For the previous study 75.5% were in full- or part-time employment.

A total of three (5.4%) reported their highest level of education as primary school level, 32 (57.1%) reported it as secondary, four (7.1%) reported it as A-level, six (10.7%) reported a university education, and 11 (19.6%) reported their highest level of education as "other" (*e.g.* nursing, M.A., college, *etc.*).

6.3.2 General health

Table 6.2 shows the responses to the EQ-5D questions. The overall mean EQ-5D score for the sample was 0.766 (median 0.859) with a range of –0.100 to 0.929. The EQ-5D score in Study 1 was 0.66. However, the score for Study 2 was still lower than the general population score of 0.83 found by Kind *et al* (1998). There was a contradiction in that general health appeared to be better in Study 1 according to the SF-36 general health questions (see Table 6.3).

6.3.3 Current IBS symptoms

Forty (71.4%) respondents reported that they had had adequate relief of abdominal pain and discomfort in the last seven days. This compares to 34.7% in Study 1. Twenty-five (44.6%) reported that they had felt a sense of urgency more than three days during the past week, compared to 73.5% in Study 1. Twenty-five (44.6%) reported that they had experienced constipation to the extent that they would wish to seek relief over the last seven days, compared to 36.7% in Study 1.

Table 6.4 demonstrates the extent to which these symptoms were coexistent or non-coexistent for each respondent across the two studies. A higher proportion of Study 2 respondents were in state P-U-C- (25% compared to 12.2% in Study 1). A higher proportion of Study 2 respondents also had only one symptom (39.3% compared to 22.4% in Study 1). Only 7.1% of Study 2 respondents had all three symptoms compared to 22.4% of respondents in Study 1. These findings support the EQ-5D analysis in finding that the Study 2 sample appeared to be on average healthier than the Study 1 sample.

6.3.4 Exclusions

Seven respondents were excluded on the basis of missing valuation data. Details of these respondents are provided in Table 6.A.3 in Appendix 2. This left a total sample of 49 respondents in the remainder of the analysis.

6.3.5 Logical consistency of ranking

Health states

Each respondent was presented with six health state descriptions to rank. These were the four health states shown at the top of Table 5.1 plus current health and immediate death. The logical ranking order for health states and the number of violations are shown in Table 6.5. There were eight (16.3%) respondents who gave inconsistent responses to the health state ranking exercise. However, two respondents gave two inconsistencies, and one respondent gave four inconsistencies. One person did not answer the ranking exercise. Forty (81.6%) respondents ranked the states in a logically consistent manner. The inconsistency is lower than in Study 1, in which 29.2% violated the consistency conditions for health states. However, the comparisons were different (Table 5.7) as the health states P+U-C+ and P+U+C+ were included in Study 1 but not in Study 2.

Table 5.9 shows the number of violations of logical consistency of health state rankings obtained in the SG ratings of states in Study 1. It was not necessary to present a corresponding table in this study because all the consistency conditions involve reference states for the gambles (Table 6.5). Both P+U- and P-U+ were given values between P-U- and P+U+ by all respondents.

Health profiles

Each respondent was presented with nine profiles to rank (see Table 6.1). The greater the number of weeks in P+U+, the fewer the number of weeks in P-U-, and therefore the less preferable the profile is logically. Thus the logical order of preference is A $\succ$ B $\succ$ C $\succ$ D $\succ$ E $\succ$ F $\succ$ G $\succ$ H.

A total of 28 (57.1%) ranked the profiles in a logically consistent manner. Nineteen (38.8%) ranked the profiles in an illogical order. One gave questionable ranking responses, ranking one state equally with the state logically below in order. One respondent had missing data for the profile ranking exercise. The relatively high rate of illogical ranking orders may have been due to the process of familiarization with the complex profiles.

The responses to the ranking exercise were examined at an individual level in order to determine the extent of illogicalness for each respondent. The respondent who ranked two profiles as equal did so for only two of the profiles. The 19 illogical respondents mis-ranked the profiles to varying degrees of illogicalness, as shown in Table 6.6. The maximum number of violations per individual is 36. The distance of error is defined as the distance from correctness that a profile is ranked. Thus if A is ranked below C, this is an error with a distance of two. The maximum distance of error is eight.

It can be seen from Table 6.6 that almost half of the respondents who did not meet the consistency conditions for the ranking of profiles only had one violation which involved swapping the position of one profile for the profile that should have been ranked immediately below it.

The degree of inconsistency of SG ratings for the profiles will be examined in Section 6.3.9 in a comparison with QALY valuations.

6.3.6 Health state valuations

The mean non-chained and chained valuations for P-U+, P+U-, and P+U+ are shown in Table 6.7. For the non-chained valuations, P-U+ and P+U- were valued against P+U+ as a lower reference point. P+U+ was valued against death. For the chained valuations P-U+ and P+U- were chained through P+U+ to a scale of death to P-U-. There was, of course, no need to chain P+U+.

As can be seen from Table 6.7, there is virtually no mean preference indicated between P-U+ and P+U-. The mean is lower than the median for both states, both chained and non-chained, indicating a negative skew. The ranges and IQRs indicate that some respondents dragged the mean down by giving particularly low values to the states.

The mean valuations for P-U+ and P+U- (both chained) and P+U+ in Study 2 are lower than in Study 1 (0.960 v. 0.982, 0.960 v. 0.986, and 0.894 v. 0.952 respectively). However, the medians for P-U+ and P+U- are 0.997 for both studies, but the IQRs are wider for Study 2. The median valuation for P+U+ in Study 2 is 0.985 compared to 0.995 in Study 1. These findings indicate that the health state valuation data has a wider dispersion for the Study 2 sample.

6.3.7 Health profile valuations

The results of the non-chained valuation exercises are shown in Table 6.8 and Figure 6.1. The y-axis ranges from zero (P+U+) to 1 (P-U-). There is a steady decline in the QALY valuations as would be expected. There is also a decline in the holistic valuations except for Profile B. However, the holistic valuations never reach a mean of zero, despite Profile H being equal to the worse state against which it was being valued.

The chained valuations are shown in Table 6.9 and Figure 6.2a and b. Figure 6.2a shows the data on a scale of death to P-U-, which puts the valuations from this study in context with the HRQoL scale as a whole. In Figure 6.2b the top part of the y axis is magnified in order to show the part of the scale covered by the valuation data from this study in more detail. The t-test and Wilcoxon-sign tests were used to test for significant differences between the results from the QALY and holistic methods. These two tests gave similar results for this study, and the differences between the results of the statistical tests described in Chapter 5 are not repeated here. There are significant differences between the results by the two valuation methods for all profiles except profile C. The lines on the graph (Figure 6.2) cross at profile C, so it is unsurprising that the values of profile C should be found to be similar for both methods.

The only profiles in Study 2 that were valued holistically in Study 1 were Profiles C. D and E. The results are similar to those of the health state valuations in that the Study 2 sample gave lower mean chained valuations for Profiles C. D and E than the study reported in Study 1 (Tables 5.12 and 6.9). These were 0.972 v. 0.981, 0.964 v. 0.980, and 0.960 v. 0.982 respectively. Again the median valuations were similar but the IQRs were wider for Study 2, as were the standard deviations. The same discrepancy between the SDs and IQRs that was seen between the health state valuations and the QALY valuations in Chapter 5 (Section 5.4.6) was repeated in this study. If the definitions of high and low values for the health state values are again set at above 0.5 and below 0.5 respectively, 30 (61.2%) respondents had high values for both health states, eight (16.3%) had low values for both health states, five (10.2%) had high values for P-U+ and low values for P+U-, and six (12.2%) had low values for P-U+ and high values for P+U-. This was similar to Study 1 and could explain the discrepancy.

Unlike for Study 1, there was a steady downward trend in both holistic and QALY valuations for the means of Profiles A to H (including for Profiles C, D and E). However, the downward slope was much steeper for the QALY method.

The paired t-test and Wilcoxon tests were used to determine whether there were significant differences between each valuation of each profile and the next profile for each method in the present study. For example, holistic chained Profiles A and B were compared, and B and C, *etc.* The same was done for the chained QALY valuations. The results are shown in Table 6.10. Each profile was significantly different from the next profile ($p < 0.01$) by the QALY method. For the holistic method, only the differences between C and D, G and H were significant according to the t-test. According to the Wilcoxon-sign test there were significant differences between profiles B and C, D and E, and G and H.

Since the results of the holistic method were not as clear-cut as the QALY results, they were examined further to see if any further insights could be obtained. The holistic results were examined to determine how marginal utility was affected by each increase by two weeks in P+U+. Marginal utility is defined as "the increase in total utility obtained by consuming one more unit of that good…" (Begg *et al*, 1994). Thus in this case an increase in marginal *dis*utility might be expected for each 2-week increase in P+U+. The differences between A and B, B and C, C and D, D and E, E and F, F and

G, and G and H were calculated. The results were all between 0.00 and 0.01. No trend was present. Thus adding two weeks in P+U+ did not affect marginal disutility.

It could be that the holistic results were related to unwillingness to gamble. This might fit the threshold effect theory. If an individual is unwilling to gamble then the value given to all the profiles should be 0.995 or above. Two respondents were unwilling to gamble, and the value given to all the health states and profiles was 0.995.

The analysis was repeated excluding the two respondents who were unwilling to gamble. The results are reported in Table 6.11 and shown graphically in Figure 6.3. The median values for the profiles are pretty much the same as in Table 6.9, but the IQRs are wider for the holistic valuations. Mean values reach lower values by both methods as the number of weeks in P+U+ increases, but only by a little (Profile H is 0.003 lower for the holistic method and 0.004 lower for the QALY method). The profiles were significantly different from each other for both methods to exactly the same degree as in Table 6.10. Excluding the two respondents who were unwilling to gamble made no difference to this aspect of the data.

6.3.8 Comparison of convergent validity between the QALY and holistic methods

The level of convergency between the ranking implied by each of the valuation methods and the original ranking order of the health profiles is shown in Table 6.12. One member of the sample had missing values for the profile ranking exercise, so the test of convergent validity was carried out for the remaining 48. The results are similar to that of Study 1 (Table 5.13) in that overall there is a higher level of convergent validity for the QALY method than the holistic method. For the QALY method, at the strong level of convergency, 27 (56.3%) respondents showed convergent validity with the original ranking for all seven pairwise comparisons between health profiles. Approximately one-third of the sample demonstrated convergent validity for five or six of the pairwise comparisons. Two (4.2) respondents were totally non-convergent for all pairwise comparisons. At the weak level of convergency, there were no respondents who achieved zero convergency, and a further three who achieved total convergency (all seven pairs).

For the holistic valuation method, at the strong level of convergency, only one (2.1%) showed convergence for all seven of the pairwise comparisons. A quarter of the sample showed zero convergence. The remainder of the sample showed convergent validity for between one and six of the pairwise comparisons. The results were more favourable at the weak level of convergency, at which 16 (33.3%) respondents achieved convergency for all seven pairwise comparisons, 11 (22.9%) achieved convergency for six pairs, seven (14.6%) achieved convergency for five pairs, 8 (16.7%) achieved convergency for four pairs, two achieved convergency for three pairs, three (6.3%) achieved convergency for two pairs, and one (2.1%) achieved convergency for zero pairs. The majority of the difference was because many respondents valued all profiles equally by the holistic approach.

6.3.9 Comparison of logical consistency between methods of valuations

By the holistic method, one (2.0%) respondent valued all profiles in the logical order (A ≻ B ≻ C ≻ D ≻ E ≻ F ≻ G ≻ H), eight (16.3%) valued all profiles equally, 18 (36.7%) valued the profiles such that some were equal to the one that should have been lower, and 22 (44.9%) ordered the profiles illogically. Thus 55.1% of the respondents showed strong or weak overall consistency for holistic valuations compared to 72.9% in Study 1.

By the QALY method, 22 (44.9%) respondents valued all the profiles in the logical order, two (4.1%) valued all profiles equally, and 25 (51.0%) valued the profiles such that some were valued equally to the one that should have been lower. No respondents valued any of the profiles in an illogical order by the QALY method in either this study or Study 1. However, in Study 1 the proportion showing strong consistency was higher (79.2%).

A detailed analysis of the extent of logical consistency and violations is shown in Table 6.13. The extent of violations is clearly much greater for the holistic method, but this method does not perform so badly in pairwaise comparisons between profiles as when all the profiles are compared. In comparisons between profiles which should have been ranked adjacently in the ranking order, the highest proportion of violations of logical consistency occur in comparisons between values for health profiles A and B, with a total of 11 (22.4%) giving A a lower value than B. Profiles A and B were the first two to be valued. In contrast, only four (8.2%) respondents gave H a lower value than G. A total of 19 (38.8%) rated G higher than H. Comparisons between profiles G and H showed the best performance for holistic values in terms of logical consistency. Despite the relatively high level of inconsistency from the holistic method, most of these responses were inconsistent by less than the MEID of 0.05. Only two responses were inconsistent by a greater amount of approximately 0.07.

Most of the pairs of profiles were valued equally by two respondents by the QALY method. As stated above, there were no violations of logical consistency for the QALY method in as much as that no profiles were rated lower than another profile that should logically have been rated lower (Table 6.13). A total of 20 (40.8%) rated profile A equal to profile B. The logical consistency for the QALY method was much higher for the other profile comparisons.

6.3.10 Patients' comments

Overall 26 (46.4%) of the total sample of 56 respondents gave comments after completing the questionnaire. These comments are presented in full in Table 6.A.4 in Appendix 2. However, the five main category of comments are as follows:

a) Need to know what type of treatment would be used (3)

b) Descriptions of IBS symptoms not adequate (5)

c) Descriptions of IBS symptoms were good (9)

d) Answers depend on how bad respondent feels while completing the questionnaire (5)

e) Confused by or did not like the gamble approach (8)

f) Thought the questionnaire was good (5)

6.4 Discussion

6.4.1 Strengths and weaknesses of the study

Some of the strengths and weaknesses of this study are dealt with in Chapter 5. This study was an extension of Study 1, and used the same basis for health state descriptions. The issue of how the descriptions of health states were designed and whether there was patient input applies equally to this study. Although in Study 1 the pilot group agreed that the descriptions were typical of IBS, in this study two of the three patients in the pilot group thought the descriptions suffered from incompleteness. Specifically, they said that the pain/discomfort item was unsatisfactory. The fact that this study was a direct follow-up of Study 1 was the reason for the inclusion of only the two symptoms of pain and urgency in the health states. As discussed in Chapter 5, the reason these symptoms were chosen by GW may have been because GW believed these symptoms to be affected by the drug under development by them at the time. The reason for using them in this study was to determine whether the apparent insensitivity of the holistic method to proportion of time in health states was repeated when a wider range of profiles was subjected to valuation. The use of the same type of health profiles was supported by the acceptance of them in the previous study. However, as already discussed, that may not have been a representative sample.

As stated above, this study was a follow-up of the study presented in Chapter 5. The objective of this study was to determine how extending the proportion of time in P+U+ over the full range of the 12-week period affected valuations. The profiles used were therefore extensions of profiles C, D and E from Study 1. As such, constipation was not evaluated in this study. The use of combinations of two symptoms to form states rather than three symptoms will have reduced the cognitive load for respondents. However, if respondents did not feel the descriptions were complete, this may have affected the meaningfulness of their valuations. A large proportion of the respondents reported that they were suffering from constipation at the time of completing the questionnaire. These were 36.7% in Study 1 and 44.6% in the current study (Tables 5.6 and 6.4).

By recruiting patients from GP practices, it was ensured that the sample used in this study was community-based. These patients may have been more representative of IBS patients than those who took part in Study 1. Broadly speaking, this study sample appeared to be healthier. Although, as pointed out in Section 5.5.1, using a non-representative sample did not detract from testing the QALY algorithm, by using a more representative sample in this study the results are more comparable with other IBS studies and perhaps used as reference valuations.

A total of eight (16.3%) respondents indicated in the comments section that they had problems with the approach to valuing health states and profiles. These problems related to the gamble approach and the descriptions of the states or profiles themselves. The ease of dealing with the method of valuation will have an obvious impact on the validity of studies such as this. Difficulty with the standard gamble approach is not relevant to this study alone, but to any studies which use this method. The standard gamble is commonly used in valuation studies. If it was the state or profile descriptions that caused problems in understanding then this would be a relevant and useful finding. One of the purposes of this research is to investigate the different ways of contructing health states and profiles. It is one of the strengths of this study that it allows such issues to come to light.

6.4.2 Comparisons with Study 1

The results of Study 2 are encouraging in that they show that IBS sufferers give lower valuations to profiles that contain greater frequencies of P+U+ by both the QALY algorithm and the holistic method. This is contrary to the findings of Study 1, which reported valuations over 2, 4 and 6 weeks in state P+U+ out of a 12-week period and found that the patient sample were insensitive to proportion of time in the health states within the profiles, thereby demonstrating "proportion of time" neglect.

However, the findings of Study 2 are quite different. The state of P+U+ was presented in profiles where it occurred over 0, 2, 4, 6, 8, 10, 12 weeks. The state of P-U- occurred in reverse frequencies in each profile. This patient sample did show sensitivity to proportion of time, although the differences between holistic valuations were in many cases not significant. The only profile for which there was no significant difference between valuation methods was for P+U+ at 2 weeks. This was the point at which the plots for the two methods crossed over (Figure 6.2). Thus this patient sample appears to have paid attention to the frequency of symptoms, and did not demonstrate a complete "proportion of time neglect". However, for most of the profiles the holistic valuations were significantly higher than the QALY valuations.

It is possible that the differences in sensitivity to proportion of time between the two samples in Study 1 and Study 2 were in part due to the greater number of profiles in the general practice sample. This may have given them a greater insight into the declining nature of the profiles. However, the GW sample had only three such profiles, and the way in which the profiles declined may not have been so clear.

6.4.3 Explanations from the field of psychology

One of the theories to explain the proportion of time neglect found in Study 1 was that there might be a threshold effect operating, such that respondents were not willing to give values of less than a certain amount to the health profiles. The results from Study 2 confirm the suggestion that there appears to be a threshold effect operating for the holistic valuations. Respondents do not appear to be willing to give health profiles valuations of below around 0.95. Another possible explanation suggested in Chapter 5 is that people use simplifying heuristics to evaluate cognitively demanding scenarios.

There is increasing evidence that people's preferences are not complete, and that they therefore form their preferences during the valuation exercises (Slovic, 1995; Shiell *et al*, 2000). Heuristics are used to ease the task. Heuristics are "cognitive strategies or shortcuts which operate to simplify the data processing requirements of many aspects of cognitive functioning" (Lloyd and Hutton, 2002). They may be used when the valuation tasks are complex. According to the fuzzy-trace theory formulated by Reyna and Brainerd (1991, 1995), individuals use heuristics to simulate numerical information by taking the gist of the numbers and mentally coding them. Reyna and Brainerd suggest that, for information relating to risk, people may code the numerical information as "high risk" or "medium risk", *etc.*

The health profiles in this and the previous study present information in terms of proportion of time spent in each health state. The fuzzy-trace theory may go some way towards explaining why the holistic valuations do not decline as a linear function of time in P+U+. The theory predicts that people may make fuzzy estimates of the magnitudes of differences, such as simply coding from highest to lowest. Respondents may have estimated the numerical information provided in the profile descriptions using a fuzzy-trace heuristic rather than mentally grasping the exact data provided. Thus the mean valuations of the profiles decline as number of weeks in P+U+ increases, but the decline in valuations does not reflect the magnitude of the increases in P+U+.

6.4.4 Tests of reliability and validity

As stated earlier in the chapter, Profile H was equal to P+U+, the worse state against which it was evaluated and the state through which it was chained to death. However, the mean valuations for this profile are significantly higher than the mean valuation for the state P+U+. It is possible that there were framing effects at work. The state of P+U+ against which it was evaluated was described simply as 12 weeks in P+U+. However, Profile H was described as 12 weeks in P+U+, 0 weeks in P+U-, 0 weeks in P-U+, and 0 weeks in P-U-. This would seem to emphasise the absence of better health states, and might persuade people to rate it as lower than the health state P+U+. However, the opposite occurred. The mean valuation of this profile was higher than the mean valuation of the state description. This may have been because the state P+U+ was referred to as the failure state, and was therefore framed as a loss regardless of how the state and profiles were described. Another possible explanation is that respondents were displaying duration neglect. The work of Ariely (1998) indicated that people may show duration neglect for retrospective evaluations of stimuli of constant intensity, whereas duration was more likely to be taken into account if intensity varied. However, since both profile H and the reference state consisted of a constant state of P+U+, the effects of framing seem the more likely explanation for the findings.

The comparison between the P+U+ profile and the state P+U+ could be taken as a test of reliability of the holistic valuation procedure. It is evident that the majority of the sample did not give the logical response of zero for the comparison. Only two (4.1%) did so. However, a further six (12.2%) crossed the choice of 0.00 chance of success and 1.00 chance of failure (indicating that they were unwilling to take this chance), and ticked the choice of 0.05 chance of success and 0.95 chance of failure (indicating that they were willing to take this chance). Thus an indifference value of 0.025 was taken for these respondents. It could be argued that this response was logical on the following grounds. The respondents may have realised that the certain outcome was equal to the failure outcome. However, they may have added the process of treatment to their thought processes. They may therefore have decided that they would not take the zero chance of success because they would then be taking a treatment that would not succeed, and they may have ascribed disutility to the process of taking the treatment.

If respondents who gave values of zero or an indifference value of 0.025 to this valuation procedure are accepted as logical, this still leaves a total of 41 (83.7%) who gave higher values. A likely explanation for this is that these respondents did not fully understand the valuation process or the scenarios. This is, of course, a damning finding with relation to the use of holistic measurement. The fact is that, even though respondents appeared to the interviewer to grasp the method, and their responses generally declined with the increase in the worst health state, they did not appear to realise that there was no difference between the scenario and health state in the last valuation exercise. The results of this reliability test cast serious doubts over the validity of the holistic valuations. However, before condemning holistic valuations, further research is required into the reasons for this apparent failure of the reliability test. What would have been required for respondents to give results that allowed the holistic method to pass this test? If it was a matter of framing, perhaps the exercise could be framed in a way that made it easier for respondents to see the descriptions presented clearly. Qualitative research into the reasons for the values given is called for in order to examine the issue in greater depth and gain a deeper understanding of the holistic values given to health profiles.

The test of convergent validity (Table 6.12) showed that the majority of the sample had a degree of convergence between ranking of QALY valuations and the original ranking of health profiles of five or more out of seven pairwise comparisons, with over half the sample achieving the full seven out of seven pairwise comparisons, *i.e.* complete convergence. The holistic method performed less well on the test of convergent validity at the strong level, with 25% of the sample achieving zero convergence out of the seven pairwise comparisons. Most of the sample achieved a level of convergence of one to five pairwise comparisons. Only one member of the sample achieved convergent validity for seven out of the seven pairwise comparisons. Of course, this measure assumes that the original ranking order of health profiles reflects respondents' true ordinal preferences. As previously discussed, respondents may have used the ranking procedure to assist in forming their preferences.

A test of logical consistency between the two valuation methods indicated that the QALY method also performed better than the holistic method in terms of logical ordering of values given to the health profiles (Table 6.13). However, the highest level of inconsistency in pairwise comparisons for the holistic method was 22.4% for comparisons between profiles A and B. For this comparison 53.1% of the sample rated A and B equally, and 24.5% rated A higher than B. This could be a reflection of how people make choices over scenarios containing detailed and complex information (*e.g.* cognitive heuristics or fuzzy-trace theory).

6.4.5 Implications for valuing health profiles

As discussed above, there could be several explanatory factors for the apparent "proportion of time" neglect detected in the holistic valuations. Without further research into the reasons underlying responses to holistic valuations, it is not yet possible to say which of the QALY and holistic methods more accurately reflects preferences. Although the tests of validity and reliability used in this study support the QALY, if the results of future research support the holistic valuation results, these results could indicate a possible violation of the QALY axiom of additive utility. According to the QALY algorithm it should be possible to obtain valuations for health profiles by multiplying time spent in a state by the utility of that state. However, the holistic valuations in these studies indicate that utility may not simply decline as a direct function of proportion of time in the worst health state, and therefore doubt is cast on the validity of the additive assumption.

6.4.6 The potential for further research

As indicated in Table 6.A.4, some respondents felt that the holistic descriptions of IBS health profiles used in this study were not adequate descriptions of their experiences of IBS. Interviews with IBS patients prior to constructing holistic health profile descriptions should be an important step in subsequent studies to value IBS health profiles holistically, in order to ensure that the profiles include all the important attributes of IBS.

Qualitative studies are required to explore the underlying reasons for the apparent failures in validity and reliability implied by the responses to the holistic profile valuations. The findings of this study are inconclusive without this further research.

6.5 Conclusions

The general practice sample of IBS patients recruited into this study was healthier in terms of average EQ-5D score and current IBS symptoms of pain, urgency and constipation than the sample of Study 1. There was a relatively low response rate of 30.6%, which was in part due to difficulties in arranging appointments that were convenient to the respondents.

Logical consistency was greater for QALY valuations of profiles than holistic valuations, suggesting that, when there are small to moderate changes in frequency of symptoms, the QALY method may be more sensitive than the holistic method to these changes. This is perhaps due to the additive nature of the QALY algorithm, and the use of heuristics in the holistic valuation process to aid judgements between cognitively demanding profiles that contain differences in numerical information. There was no evidence of the complete "proportion of time" neglect that appeared apparent in Study 1. However, the results of the holistic valuations, though declining with the increase in P+U+, never reached the low levels reached by the QALY valuations.

The majority of the sample did not give a value of zero to the profile consisting entirely of P+U+, even though it was valued against a failure state of P+U+. The fact that this test of the reliability of the holistic valuation method failed casts doubts upon the validity of the method, and lends support to the continued use of the QALY algorithm.

However, further research is required in order to more fully understand the responses to the holistic valuation exercises. In particular, it would be useful to carry out qualitative research into the reasons underlying individual responses.

Figure 6.1 Comparison of mean non-chained holistic and QALY profile values

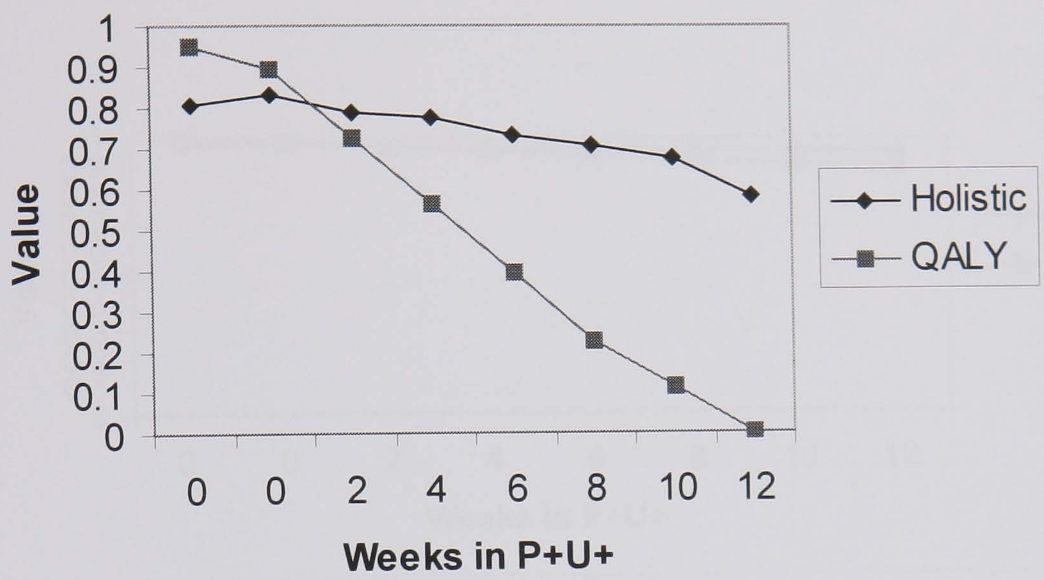


Figure 6: Comparison of mean chained holistic and QALY profile values (a) on a scale of 0 to 1 to put the data in context, and (b) with the upper part of the scale for increased detail.

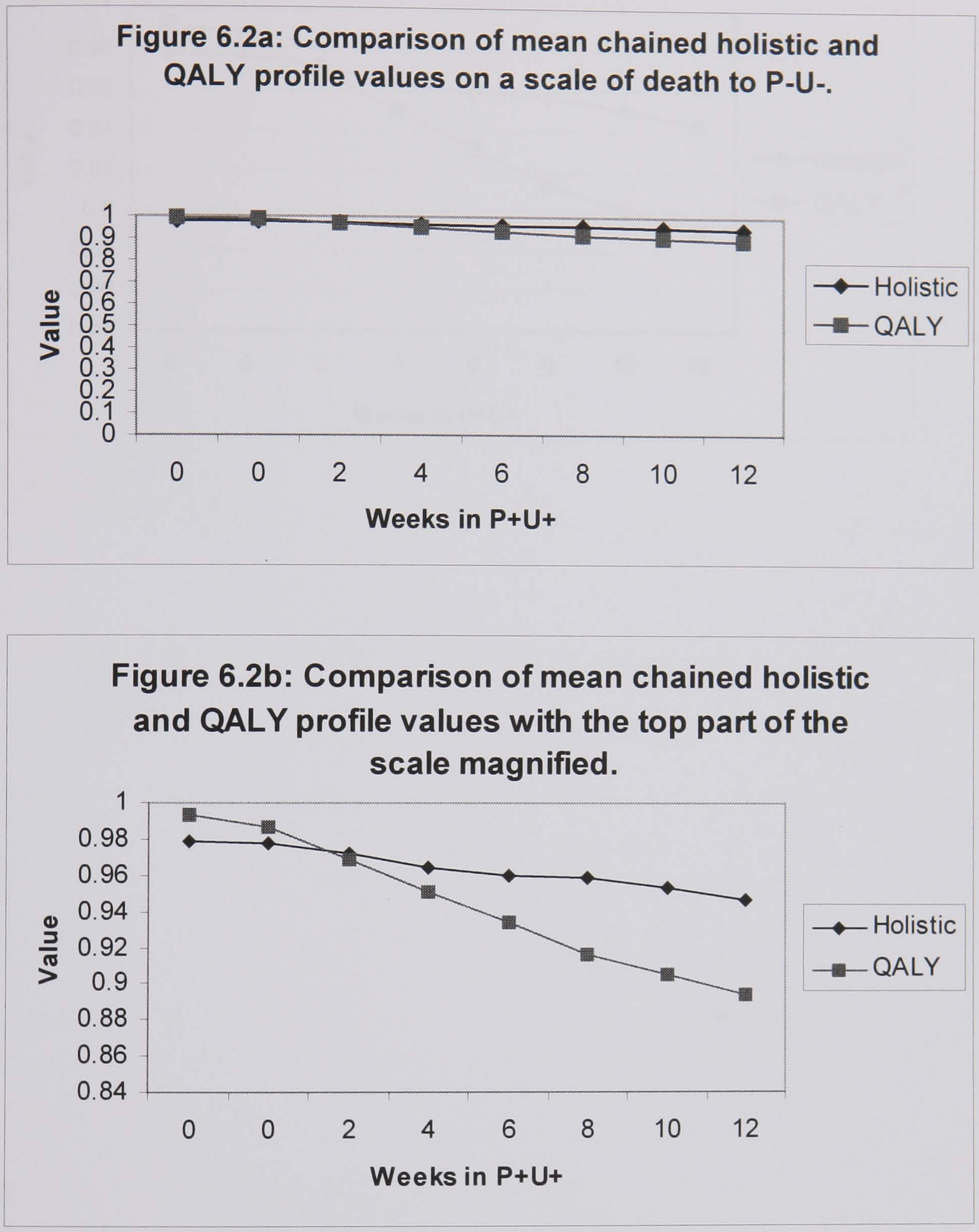


Figure 6.2a: Comparison of mean chained holistic and QALY profile values on a scale of death to P-U-.

Figure 6.2b: Comparison of mean chained holistic and QALY profile values with the top part of the scale magnified.

Figure 6.3 Holistic and QALY profiles excluding non-gamblers.

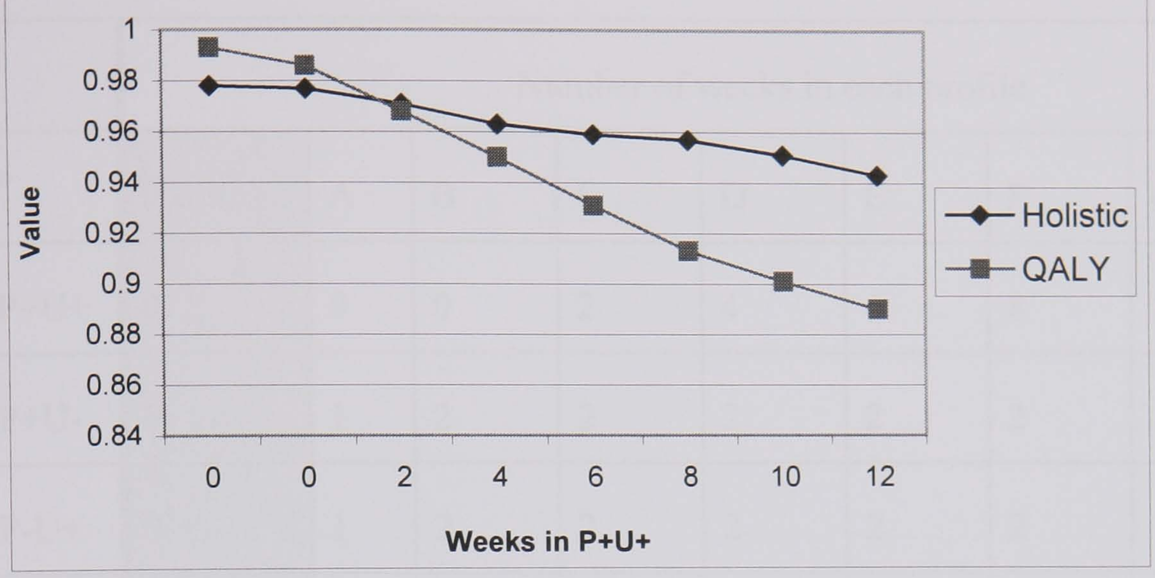


Table 6.1 Health profiles.

	Number of weeks in each profile								
	Profiles	A	B	C	D	E	F	G	H
P+U+	0	0	0	2	4	6	8	10	12
P+U-	0	1	2	2	2	2	2	1	0
P-U+	0	1	2	2	2	2	2	1	0
P-U-	12	10	8	6	4	2	0	0	0

Table 6.2 Percentage of respondents reporting levels 2 or 3 on the EQ-5D by current IBS health state.

States	N	Mobility	Self-care	Usual activities	Pain/discomfort	Mood
P-U-C-	14	14.3	0.0	14.3	50.0	42.9
P-U+C-	12	25.0	8.3	41.7	58.3	25.0
P+U-C-	3	0.0	0.0	33.3	66.7	33.3
P-U-C+	7	28.6	14.3	57.2	100.0	42.9
P+U+C-	2	50.0	50.0	100.0	100.0	50.0
P+U-C+	7	0.0	28.6	28.6	71.4	71.4
P-U+C+	7	28.6	14.3	14.3	57.1	28.6
P+U+C+	4	75.0	25.0	75.0	100.0	50.0

Table 6.3 A comparison of the general health item on SF-36 between the IBS patient samples from Studies 1 and 2. The number in each category is followed by the percentage of that study sample in brackets.

	Excellent	Very good	Good	Good/Fair	Fair	Fair Poor	Poor
Study 1	1 (2.0)	15 (30.6)	26 (53.1)	0.0	6 (12.2)	0.0	1 (2.0)
Study 2		16 (28.6)	16 (28.6)	1 (1.8)	15 (26.8)	2 (3.6)	6 (10.7)

Table 6.4 The current health states of the sample, where P = pain, U = urgency, and C = constipation.

Code	Study 2	Study 1
	N (%)	*N (%)*
P-U-C-	14 (25)	6 (12.2)
P+U-C-	3 (5.4)	3 (6.1)
P-U+C-	12 (21.4)	7 (14.3)
P+U+C-	2 (3.6)	15 (30.6)
P-U+C+	7 (12.5)	3 (6.1)
P-U-C+	7 (12.5)	1 (2.0)
P+U-C+	7 (12.5)	3 (6.1)
P+U+C+	4 (7.1)	11 (22.4)
Total	56 (100)	49 (100)

Table 6.5 Logical consistency conditions for health states.

Consistency condition	No. of violations in ranking exercise
P-U- > P+U-	3
P-U- > P-U+	4
P-U- > P+U+	1
P+U- > P+U+	4
P-U+ > P+U+	1

Table 6.6 Degree of mis-ranking for health profiles.

Number of violations	Number of respondents	Maximum distance of error
1	7	1
2	2	2
3	2	3
4	1	3
5	2	4
6	2	6
19	2	7
24	1	8

Table 6.7 Non-chained and chained valuations for health states.

		P-U+	P+U-	P+U+
Non-chained	Mean (SD)	0.680 (0.205)	0.691 (0.215)	0.894 (0.188)
	Median (IQR)	0.725 (0.475-0.875)	0.725 (0.475-0.925)	0.985 (0.875-0.995)
	Min	0.225	0.225	0.275
	Max	0.995	0.995	1.000
Chained	Mean (SD)	0.960 (0.077)	0.960 (0.082)	N/A
	Median (IQR)	0.997 (0.942-0.999)	0.997 (0.951-0.999)	N/A
	Min	0.646	0.646	N/A
	Max	1.000	1.000	N/A

Table 6.8 Non-chained health profile valuations.

Profile	Median (IQR)		Mean (SD)		Mean holistic-QALY (95% CIs)	t-test	Wilcoxon
	Holistic	QALY	Holistic	QALY		*p*	*p*
A	0.875	0.946	0.807	0.948	-0.141 +/- 0.062	0.000	0.000
	(0.775-0.975)	(0.923-0.975)	(0.234)	(0.032)	(-0.203 to –0.078)		
B	0.925	0.892	0.829	0.895	-0.066 +/- 0.045	0.005	0.018
	(0.775-0.975)	(0.846-0.950)	(0.192)	(0.063)	(-0.110 to –0.021)		
C	0.850	0.725	0.785	0.728	0.057 +/- 0.047	0.018	0.018
	(0.600-0.965)	(0.679-0.783)	(0.206)	(0.063)	(0.010 to 0.104)		
D	0.825	0.558	0.778	0.562	0.216 +/- 0.044	0.000	0.000
	(0.625-0.955)	(0.513-0.617)	(0.188)	(0.063)	(0.173 to 0.260)		
E	0.775	0.392	0.733	0.395	0.338 +/- 0.051	0.000	0.000
	(0.575-0.940)	(0.346-0.450	(0.214)	(0.063)	(0.287 to 0.389)		
F	0.725	0.225	0.707	0.228	0.479 +/- 0.063	0.000	0.000
	(0.500-0.960)	(0.179-0.283)	(0.254)	(0.063)	(0.416 to 0.542)		
G	0.725	0.113	0.674	0.114	0.559 +/- 0.076	0.000	0.000
	(0.425-0.970)	(0.090-0.142)	(0.277)	(0.032)	(0.484 to 0.635)		
H	0.625	0.000	0.579	0.000	0.579 +/- 0.102	0.000	0.000
	(0.300-0.925)	(0.000-0.000)	(0.355)	(0.000)	(0.477 to 0.681)		

Table 6.8 Non-chained health profile valuations.

Profile	Median (IQR)		Mean (SD)		Mean holistic-QALY (95% CIs)	t-test	Wilcoxon
	Holistic	QALY	Holistic	QALY		*p*	*p*
A	0.875	0.946	0.807	0.948	-0.141 +/- 0.062	0.000	0.000
	(0.775-0.975)	(0.923-0.975)	(0.234)	(0.032)	(-0.203 to –0.078)		
B	0.925	0.892	0.829	0.895	-0.066 +/- 0.045	0.005	0.018
	(0.775-0.975)	(0.846-0.950)	(0.192)	(0.063)	(-0.110 to –0.021)		
C	0.850	0.725	0.785	0.728	0.057 +/- 0.047	0.018	0.018
	(0.600-0.965)	(0.679-0.783)	(0.206)	(0.063)	(0.010 to 0.104)		
D	0.825	0.558	0.778	0.562	0.216 +/- 0.044	0.000	0.000
	(0.625-0.955)	(0.513-0.617)	(0.188)	(0.063)	(0.173 to 0.260)		
E	0.775	0.392	0.733	0.395	0.338 +/- 0.051	0.000	0.000
	(0.575-0.940)	(0.346-0.450	(0.214)	(0.063)	(0.287 to 0.389)		
F	0.725	0.225	0.707	0.228	0.479 +/- 0.063	0.000	0.000
	(0.500-0.960)	(0.179-0.283)	(0.254)	(0.063)	(0.416 to 0.542)		
G	0.725	0.113	0.674	0.114	0.559 +/- 0.076	0.000	0.000
	(0.425-0.970)	(0.090-0.142)	(0.277)	(0.032)	(0.484 to 0.635)		
H	0.625	0.000	0.579	0.000	0.579 +/- 0.102	0.000	0.000
	(0.300-0.925)	(0.000-0.000)	(0.355)	(0.000)	(0.477 to 0.681)		

Table 6.9 Chained health profile valuations.

Profile	Median (IQR)		Mean (SD)		Mean holistic-QALY (95% CIs)	t-test	Wilcoxon
	Holistic	QALY	Holistic	QALY		*p*	*p*
A	0.999	0.999	0.979	0.993	-0.015 +/- 0.011	0.006	0.001
	(0.994-1.000)	(0.992-1.000)	(0.046)	(0.013)	(-0.025 to –0.004)		
B	0.999	0.999	0.978	0.987	-0.009 +/- 0.006	0.012	0.018
	(0.966-1.000)	(0.983-1.000)	(0.039)	(0.026)	(-0.015 to –0.002)		
C	0.998	0.997	0.972	0.969	0.003 +/- 0.007	0.401	0.144
	(0.961-1.000)	(0.960-0.999)	(0.055)	(0.055)	(-0.004 to 0.010)		
D	0.999	0.995	0.964	0.951	0.013 +/- 0.008	0.003	0.000
	(0.959-1.000)	(0.939-0.998)	(0.073)	(0.086)	(0.005 to 0.021)		
E	0.997	0.992	0.960	0.934	0.026 +/- 0.015	0.000	0.000
	(0.948-1.000)	(0.918-0.997)	(0.075)	(0.117)	(0.012 to 0.041)		
F	0.997	0.990	0.959	0.916	0.042 +/- 0.023	0.000	0.000
	(0.944-1.000)	(0.897-0.996)	(0.080)	(0.148)	(0.020 to 0.065)		
G	0.997	0.987	0.953	0.905	0.048 +/- 0.025	0.000	0.000
	(0.944-1.000)	(0.886-0.996)	(0.098)	(0.168)	(0.023 to 0.073)		
H	0.995	0.985	0.946	0.894	0.051 +/- 0.029	0.001	0.000
	(0.942-1.000)	(0.875-0.995)	(0.116)	(0.188)	(0.022 to 0.080)		

Table 6.10 Results of paired t-test and Wilcoxon test to determine whether profiles means are significantly different from each other.

	Holistic			QALY		
Profiles	Mean difference (95% CIs)	t-test (*p*)	Wilcoxon (*p*)	Mean difference (95% CIs)	t-test (*p*)	Wilcoxon (*p*)
A to B	0.000 +/- 0.005 (-0.004 to 0.005)	0.839	0.480	0.007 +/- 0.004 (0.003 to 0.010)	0.001	0.000
B to C	0.006 +/- 0.007 (-0.001 to 0.013)	0.107	0.019	0.018 +/- 0.009 (0.009 to 0.027)	0.000	0.000
C to D	0.008 +/- 0.008 (0.000 to 0.016)	0.047	0.086	0.018 +/- 0.009 (0.009 to 0.027)	0.000	0.000
D to E	0.004 +/- 0.006 (-0.002 to 0.009)	0.153	0.019	0.018 +/- 0.009 (0.009 to 0.027)	0.000	0.000
E to F	0.002 +/- 0.004 (-0.002 to 0.005)	0.345	0.214	0.018 +/- 0.009 (0.009 to 0.027)	0.000	0.000
F to G	0.005 +/- 0.009 (-0.004 to 0.014)	0.231	0.186	0.011 +/- 0.006 (0.005 to 0.017)	0.001	0.000
G to H	0.008 +/- 0.008 (-0.000 to 0.015)	0.050	0.001	0.011 +/- (0.005 to 0.017)	0.001	0.000

Table 6.11 The profile values excluding the two respondents who were unwilling to gamble (n = 47).

Profile	Median (IQR)		Mean (SD)		t-test	Wilcoxon
	Holistic	QALY	Holistic	QALY	*p*	*p*
A	0.999 (0.978-1.000)	0.999 (0.992-1.000)	0.978 (0.05)	0.993 (0.01)	0.006	0.001
B	0.999 (0.961-1.000)	0.999 (0.983-1.000)	0.977 (0.04)	0.986 (0.03)	0.012	0.018
C	0.998 (0.961-1.000)	0.997 (0.959-0.999)	0.971 (0.06)	0.968 (0.06)	0.406	0.185
D	0.999 (0.957-0.999)	0.994 (0.939-0.998)	0.963 (0.07)	0.950 (0.09)	0.003	0.000
E	0.997 (0.947-1.000)	0.992 (0.918-0.997)	0.959 (0.08)	0.931 (0.12)	0.000	0.000
F	0.997 (0.942-1.000)	0.989 (0.897-0.996)	0.957 (0.08)	0.913 (0.15)	0.000	0.000
G	0.997 (0.938-1.000)	0.987 (0.886-0.996)	0.951 (0.10)	0.901 (0.17)	0.000	0.000
H	0.995 (0.934-0.999)	0.985 (0.875-0.995)	0.943 (0.12)	0.890 (0.19)	0.001	0.000

Table 6.12 Degree of convergence between the QALY and holistic valuations and the original ranking of the health profiles. The number (%) of pairs of profiles showing convergency is shown.

	Strong				Weak			
	QALY		Holistic		QALY		Holistic	
	n	%	n	%	n	%	n	%
0/7 pairs	2	4.2	12	25.0	0	0.00	1	2.1
1/7 pairs	0	0.0	4	8.3	0	0.00	0	0.0
2/7 pairs	0	0.0	5	10.4	0	0.00	3	6.3
3/7 pairs	2	4.2	8	16.7	2	4.17	2	4.17
4/7 pairs	1	2.1	6	12.5	1	2.01	8	16.7
5/7 pairs	5	10.4	9	18.8	5	10.42	7	14.6
6/7 pairs	11	22.9	3	6.3	10	20.8	11	22.9
7/7 pairs	27	56.3	1	2.1	30	62.5	16	33.3
Total	48	100.0	48	100.0	48	100.0	48	100.0

Table 6.13 Logical consistency for chained QALY and holistic valuations of the health profiles.

		B		C		D		E		F		G		H	
		QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic
A	≻	29 (59.2)	12 (24.5)	47 (95.9)	15 (30.6)	47 (95.9)	17 (34.7)	47 (95.9)	23 (46.9)	47 (95.9)	23 (46.9)	47 (95.9)	25 (51.0)	47 (95.9)	27 (55.
	=	20 (40.8)	26 (53.1)	2 (4.1)	24 (49.0)	2 (4.1)	21 (42.9)	2 (4.1)	18 (36.7)	2 (4.1)	16 (32.7)	2 (4.1)	12 (24.5)	2 (4.1)	11 (22.
	≺	0 (0.0)	11 (22.4)	0 (0.0)	10 (20.4)	0 (0.0)	11 (22.4)	0 (0.0)	8 (16.3)	0 (0.0)	10 (20.4)	0 (0.0)	12 (24.5)	0 (0.0)	11 (22.
B	≻			43 (87.8)	16 (32.7)	47 (95.9)	19 (38.8)	47 (95.9)	25 (51.0)	47 (95.9)	24 (49.0)	47 (95.9)	25 (51.0)	47 (95.9)	26 (53.
	=			6 (12.2)	28 (57.1)	2 (4.1)	22 (44.9)	2 (4.1)	18 (36.7)	2 (4.1)	18 (36.7)	2 (4.1)	15 (30.6)	2 (4.1)	16 (32.
	≺			0 (0.0)	5 (10.2)	0 (0.0)	8 (16.3)	0 (0.0)	6 (12.2)	0 (0.0)	7 (14.3)	0 (0.0)	9 (18.4)	0 (0.0)	7 (14.3
C	≻					47 (95.9)	11 (22.4)	47 (95.9)	20 (40.8)	47 (95.9)	22 (44.9)	47 (95.9)	24 (49.0)	47 (95.9)	24 (49.
	=					2 (4.1)	32 (65.3)	2 (4.1)	23 (46.9)	2 (4.1)	20 (40.8)	2 (4.1)	16 (32.7)	2 (4.1)	18 (36.
	≺					0 (0.0)	6 (12.2)	0 (0.0)	6 (12.2)	0 (0.0)	7 (14.3)	0 (0.0)	9 (18.4)	0 (0.0)	7 (14.3
D	≻							44 (89.8)	17 (34.7)	47 (95.9)	19 (38.8)	47 (95.9)	23 (46.9)	47 (95.9)	24 (49.
	=							5 (10.2)	27 (55.1)	2 (4.1)	23 (46.9)	2 (4.1)	18 (36.7)	2 (4.1)	19 (38.
	≺							0 (0.0)	5 (10.2)	0 (0.0)	7 (14.3)	0 (0.0)	8 (16.3)	0 (0.0)	6 (12.2
E	≻									40 (81.6)	14 (28.6)	47 (95.9)	21 (42.9)	47 (95.9)	23 (46.
	=									9 (18.4)	27 (55.1)	2 (4.1)	18 (36.7)	2 (4.1)	17 (34.
	≺									0 (0.0)	8 (16.3)	0 (0.0)	10 (20.4)	0 (0.0)	9 (18.4
F	≻											35 (71.4)	15 (30.6)	45 (91.8)	21 (42.
	=											14 (28.6)	28 (57.1)	4 (8.2)	22 (44.
	≺											0 (0.0)	6 (12.2)	0 (0.0)	6 (12.2
G	≻													42 (85.7)	19 (38.
	=													7 (14.3)	26 (53.
	≺													0 (0.0)	4 (8.2)