University of Southern Denmark · Syddansk Universitet

An Empirical Assessment of the Person Trade-Off: Valuation of Health, Framing Effects, and Estimation of Weights for Fairness

Kim U. Wittrup-Jensen[1,3] *& Kjeld M. Pedersen*[2]

(1): Bayer HealthCare AG, kim.wittrup-jensen@bayerhealthcare.com

(2): Institute of Public Health – Health Economics, University of Southern Denmark, kmp@sam.sdu.dk

(3): The study was done when Kim U. Wittrup-Jensen was a PhD student at University of Southern Denmark

Health EconomicsPapers
2008:9

UNIVERSITY OF SOUTHERN DENMARK

Abstract

Background. The Person Trade-Off (PTO) method is a preference-based evaluation technique that can be used to elicit individuals' preferences for health states to obtain an index number on a cardinal scale. The technique is similar to the Time Trade-Off (TTO) method, but focuses on persons instead of time as the trade-off unit. This is the first study using randomly drawn respondents from the general population. Hopefully this will add new and important evidence about the PTO method.

Objectives. The study has three main aims: (1) to conduct the first PTO study in a randomly drawn sample of the general population estimating cardinal values for health states and to test whether the framing of the questions influences the valuations of health states, (2) to estimate fairness weights (severity-of-illness and potential-for-health), and (3) to investigate the general population's preferences for health (societal preferences).

Data and methods. The study was based on face-to-face interviews in the respondents' own homes. In total interviews from 580 randomly drawn respondents in the Danish adult population were obtained. For health states, the EuroQol (EQ-5D) classification system and specially designed boards to make the trade-off between health states less difficult for the respondents, were used. Four different frames: PTO-1 (prevention), PTO-2 (prevention), PTO-1 (treatment), and PTO-2 (treatment) were tested and a non-parametric *t*-test to test for framing effects was applied.

Results. The mean valuations of the EQ-5D health states were clustered toward the upper end of the 0 to 1 scale. The EQ-5D health state 33333 was valued at around 0.6. The results indicate that using a different point of reference (gain versus loss) did not make a difference, which conflicts with the concept of Prospect Theory. However, evidence was found that using either a prevention or a treatment scenario resulted in the majority of EQ-5D health states being significantly different. The study proved that it is possible to estimate weights for fairness and that the results are consistent. Respondents tend to divide scarce health care resources equally, which is supported by previous findings, that is, pursue an equity strategy rather than a simple maximisation strategy.

Conclusions. As shown in previous studies, the mean PTO valuations of EQ-5D health states were higher than corresponding valuations when applying either the TTO or Standard gamble (SG) scaling techniques. However, this was not only due to framing effects, but also to the structure of our study. The message is that it would be wrong to conclude that given that the PTO technique is able to reflect societal values, valuations elicited using either the TTO or the SG techniques are inappropriate just because the (mean) valuations, in general, are lower. It is too soon to say anything conclusive about whether the PTO technique reflects societal preferences. Nevertheless, the PTO is an appealing possibility, since asking PTO questions is more related to how allocation of scarce health care resources is practised in the 'real' world.

Introduction

Decision-makers are increasingly faced with more explicit choices between devoting scarce resources for health care to one group of patients at the cost of care to different groups of patients. Hence questions concerning priority setting have led to a more open debate about how society ought to allocate scarce health care resources. When comparing outcomes and costs of different diseases, one normally uses the Quality Adjusted Life-Year (QALY) approach. This method is appropriate if the focus is on maximizing health gains. However, during the past decade criticism of the QALY method has increased. The reason is that it has become more widely accepted that social decisions concerning the treatment of different patient groups may be shaped by issues other than just the magnitude of health benefits in the form of the number of QALYs gained [Green 2001].

It may be argued that society, in general, places importance on issues other than the maximisation of QALYs, for example, age (preferring to treat the young patient to the elderly patient), the severity of a patient's pre-treatment condition, or the presence of a life saving opportunity, as opposed to a life saving one [Williams 1997; Ubel 1999; Hadorn 1991; Gold *et al.* 1996; Nord 1992; Dolan & Cookson 2001]. The implications would be that decision-makers when allocating scarce resources have to apply techniques that are capable of assessing some of these concerns. Some authors have pointed out that the usual techniques such as the Rating Scale (RS), Standard Gamble (SG), and Time Trade-Off (TTO) contain important problems [Eddy 1991; Nord 1993; Nord *et al.* 1993; Richardson 1994; Nord 1992; Cohen 1995; Menzel 1990].

According to Prades (1997), one of the main problems with the afore-mentioned scaling techniques is that they all ask questions about how people value their *own* health, which may not necessarily provide any guidelines on how society wishes to balance health benefits for different groups of the population. In other words, these techniques fail to capture the preferences of individuals with respect to health gains that affect other individuals.

According to Nord (1995), the PTO technique is a way of estimating the *societal values* of different health care interventions. Basically the PTO technique consists in asking individuals how many outcomes of one kind they consider equivalent in social value to *X* outcomes of another kind, expressed by number of persons. The technique was first introduced by Patrick et al. (1973) as the 'equivalence of numbers technique'. The label 'person trade-off' was later introduced by Nord (1992). The attractiveness of the PTO method is that it intends to capture the preferences of individuals relative to collective choices that do not directly affect the health status of the individual whose preferences are being elicited [Cabasés *et al.* 2000]. The PTO technique places the respondent in the position of a decision-maker with a limited budget who has to choose among a series of alternative health care interventions. The goal is to obtain the individual's valuation of given health states through a trade-off procedure similar to the TTO technique. The trade-off explicitly implies that some people could benefit from the intervention and that others, if this intervention is chosen, are consequently being denied a different intervention.

The PTO technique presents the individual as a decision-maker with two variables: 1) the number of people that may benefit from a health care intervention, and 2) the health improvement to be brought about by the implementation of this health care intervention. The trade-off implies that the individual must choose between the combinations of the two variables that she/he finds to be equivalent using persons as the equilibrating mechanism. To date, PTO has been applied in a limited number of studies.[1] All of these are characterised as being of a more experimental character rather than trying to make random and representative valuations within the general population [Olsen 1994; Nord *et al.* 1993; Nord *et al.* 1995; Ubel *et al.* 1998; Dolan & Green 1998; Cabasés *et al.* 1999].

There is a general agreement that QALYs are a measure of the volume of health output, however, the QALY approach has been widely criticised for not incorporating distributive preferences [Weinstein & Stason 1977; Nord 1989; Mooney & Olsen 1991; Olsen 2000]. As stated by Nord *et al.* (1999) *"... society's overall valuation of health output is a function not only of total output, but also of the distribution of health output across individuals."* This implies that society must be prepared to make some sacrifices in the total production of health in order to secure that health care is allocated in a (more) fair and equitable way. In the literature there have been several proposals to assign so-called *equity weights* to the QALY approach in order to account for distributive preferences [Culyer 1989; Broome 1991; Williams 1997; Nord *et al.* 1999].

Nord *et al.* (1999) introduced the term *Health-Related Societal Value* (HRSV), which can be used to describe the overall value that society in general assigns to different health outcomes and interventions, when both concerns for equity and efficiency are taken into account in the allocation of health care resources. All things being equal, equity weighted QALYs are measures of HRSV.

How should such equity weights look and how can they be incorporated into the QALY approach? Two proposals have been made: the F*air Innings* argument proposed by Williams (1997), and C*ost-Value Analysis* proposed by Nord (1999). The former approach reflects the feeling that everyone is entitled to some 'normal' life span of health (usually expressed in life years) and anyone failing to achieve this has been cheated, whilst anyone getting more than this is 'living on borrowed time'.[2] The latter approach addresses two concerns for fairness: 1) severity of illness, and 2) limitations in potential for health.

[1] However, one large study has been conducted by the WHO, where the PTO method was used to estimate Disability Adjusted Life-Years (DALYs), cf. Lauer (2000).

[2] For a more in-depth discussion of the fair innings argument, refer to Williams (1997).

Objectives

The main objective of this study is to test empirically the PTO technique in measuring health state valuations by involving a randomised, yet not representative, sample of the general Danish population. Thus far PTO studies have been limited to 'classroom' samples, which are neither randomised nor representative. This study will provide more robust empirical results on how respondents react when asked trade-off questions within the PTO scenario. A second objective is to test whether different frames for the PTO questions result in different values. This is based on the argument that one of the critical characteristics of the PTO method (and all other methods) be that it asks *the right question* [Prades 1997]. However, asking the right question may not be enough, given that the literature shows that preference elicitation methods can be subject to important biases caused by framing effects [Kahnemann *et al.* 1982; McCord & Neufville 1986; Hershey *et al.* 1991]. As noted by Fischhoff (1991), such problems are especially important if asking questions with which the respondents are unfamiliar. As documented by Nord (1995), there may be biases when using different frames. Hence, four different frames will be tested to check whether the numerical values of the health states depend on the way the PTO question is framed.

According to Prospect Theory preferences are reference-point dependent, which suggests that framing of the PTO questions matters, and evidently will result in different health state valuations [Kahnemann & Tversky 1979]. It will be investigated whether the foundations of prospect theory hold.

The third objective is to elicit weights for *severity-of-illness* and *potential for health* by asking respondents PTO questions about their preferences concerning to whom to allocate health care resources. The procedure has been proposed by Nord *et al.* (1999). This study largely adopts these proposals, and investigate whether they can be applied in a broader empirical setting.

To date, two other studies have tried to estimate weights for severity-of-illness and potential for health. However, both studies were based on small and non-randomised samples [Nord 1995; Cabasés *et al.* 1999]. So, finally, in order to test for preferences for equity, we asked the respondents more directly how they would allocate health care resources when having to choose between two alternatives: the largest health gain, or favouring people who are worse off.

Data

Study design

This study employed an interview-based computer-administered questionnaire including a prior telephone screening asking individuals whether they wished to participate. The study was performed by trained interviewers employed by AC Nielsen AIM and took place in the months of September and October 2001. Due to a limited budget, the sample could not be made representative of the Danish population. Nevertheless, we aimed to make the study randomised according to gender, age, and geographic domicile across Denmark. AC Nielsen AIM used the Danish National Register to draw addresses. The study contained four sub-studies due to different framings of the trade-off questions. Consequently, four samples (randomly drawn and stratified on age, gender and geographic domicile) were derived from the general Danish adult population of individuals 18 years old and over.

The respondent characteristics of the four samples are illustrated in Table 1. In total, addresses of 3,678 respondents were drawn of which 2,876 were used for the telephone screening. 96 addresses were dropped due to severe illness, problems with the Danish language etc., 23 addresses were not private addresses and in 1,301 cases the specific individual was not at home. This resulted in a total net sample of 1,456 individuals of whom 876 declined to participate in the study. Thus 580 individuals agreed to participate, a response rate of 39.8 per cent (580/1,456) or 20.2 per cent (580/2,876), depending on how one interprets the response rate to be calculated. Since from the outset we were aware that the study could not be representative of the Danish population, the response rate itself was of less importance. Instead, we focused on the absolute number of 580 interviewed individuals, which was as many as our budget allowed us to interview.

Table 1. Respondent characteristics.

	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Total**
Used addresses	703	1,439	711	825	3,678
- *unused addresses*	78	444	180	100	802
= Gross sample	625	995	531	725	2,876
Omitted due to disease, language etc.	28	31	28	9	96
Not a private address	12	2	8	1	23
Person not at home	263	511	191	336	1,301
= Net sample	322	451	304	379	1,456
- *Refused to participate in interview*[a]	202	281	184	209	876
= Acc. to participate in interview	102	170	120	170	580

[a]Stated in the telephone screening.

The majority of non-respondents did not give reasons for their refusal to participate in the study. To preserve the anonymity of the respondents no sample information was kept after the interviews were conducted. Thus a strict dropout analysis could not be conducted. As illustrated in Table 2, the distribution of the respondents in the sample with respect to gender differed from the distribution of the general Danish population, since our sample had an over-weighting of females (under-weighting of males). We chose not to weight the sample, even though a difference between females and males was present, which one should be aware of when performing analyses where the gender factor is significant. Weighting gender would not have had a significant impact on the results, and could have had adverse side effects such as an 'over-weighting' of cases, with errors in measurement or other inaccuracies. Weighting was a possibility, but for operational reasons we chose not to undertake such an exercise. There were no large differences according to age distribution.

Table 2. Distribution in the PTO study by gender and age distributions. Per cent.

	General population (N = 4,127,847) (January 1^{t} 2000)	**PTO personal interview survey (N = 580) (Spring 2001)**
Gender		
Male	48.9 %	40.7 %
Female	51.1 %	59.3 %
Age		
18 – 29 years	21.1 %	22.8 %
30 – 59 years	54.1 %	56.0 %
≥ 60 years	24.8 %	21.2 %

Methods

The study was face-to-face interview-based and split into seven independent exercises: 1) the EQ-5D profile questionnaire, 2) a rank ordering exercise, 3) a valuation exercise, 4) self-reported health status on a Visual Analogue Scale (VAS), 5) a PTO exercise, 6) a weighting exercise, and 7) a distribution exercise and, finally, some relevant background questions. Exercises 1) to 3) were so-called 'warm-up' exercises, where the idea was to familiarise respondents with the way of thinking when valuing health states using the EQ-5D descriptive system as well as each health state. The results from these exercises are reported in this section.

Health states

We applied the EuroQol (EQ-5D) as descriptor for the health states employed [Brooks *et al.* 1996]. All respondents were confronted with the same EQ-5D health states. Since there is evidence that individuals can cope with only a limited number of health states, we presented each respondent with 18 EQ-5D health states including the states 'unconscious' and 'immediate death' [Dolan *et al.* 1995]. The standard EQ-5D questionnaire encompasses only 15 health states. Hence, we added a further 3 EQ-

5D health states in order to make the sample of health states as representative of the EQ-5D classification system as possible. Table 3 shows the EQ-5D health states used in this study.

Table 3. EQ-5D health states used in this study and their representation.

EQ-5D health states	**Commentary**
11211 11111 21232 11122 11121 22233 33333 33321 21111 Unconscious 12111 11112 32211 22323 Immediate death	Health states encompassed in the standard EQ-5D valuation exercise
11113 22331 22222	The health states added

The EQ-5D profile

All respondents filled out the EQ-5D profile to describe their own state of health. The results are illustrated in Table 4. Around 67 per cent of the respondents valued their own health as perfect (health state 11111). Around 30 per cent indicated that they had a moderate problem (defined as all respondents who did not indicate the health state 11111) and around 3 per cent had a severe problem (defined as those respondents who indicated having at least one extreme problem, shown by at least one '3').

Table 4. Distribution on EQ-5D health states based on results from the EQ-5D profile. (n = 580).

Health states	Number of respondents	Per cent	Health states	Number of Respondents	Per cent
11111	393	67.2	21111	3	0.5
11112	11	1.9	21112	2	0.9
11121	65	11.2	21121	19	3.3
11122	6	1.0	21122	3	0.5
11123	2	0.3	21211	5	0.9
11131	2	0.3	21221	19	3.3
11211	4	0.7	21222	4	0.7
11221	13	2.2	21223	1	0.2
11222	8	1.4	21231	2	0.3
11232	1	0.0	21232	2	0.3
11322	1	0.0	21311	1	0.2
11323	1	0.2	21321	2	0.3
12111	2	0.3	22211	1	0.2
12121	1	0.2	22221	3	0.5
12211	1	0.2	22222	1	0.2
12221	1	0.2	Total	580	100.0

The rank ordering exercise

In the rank ordering exercise the respondents had all health states (n = 18 including 'Unconscious (UNC)' and 'Immediate Death' (ID)) placed in front of them and were asked to rank them so that the best health states were placed at the top and the worst health states at the bottom. The respondents were told that they had to picture themselves in each health state for a period that would last *10 years*, after which they would die.

Table 5 shows the respondents' mutual ranking of the 18 health states. As can be seen from the table, judged by median value (of ranking), respondents in all four samples, ranked the health state 11111 (perfect health) as the best health state, which followed the method in the EQ-5D classification system. Further, the table illustrates that the worse the dimensions became, the lower the health state was ranked, compared to the remaining health states. The health state 33333 was, in all four samples, placed at the bottom, followed by the states ID and UNC as the second worst ranking, i.e. the health state 33333 was judged to be worse than death.

Table 5. Rank ordering of health states. Judged by median value. (n = 580).

Rank	Sample 1	Sample 2	Sample 3	Sample 4
1	11111	11111	11111	11111
2				
3	11211,12111,21111	11211,11121,21111	11211,12111,21111	11211,12111,21111,11121
4	11121	12111	11121	
5	11112	11112	11112	11112
6				
7	11122	11122	11122	11122
8	22222	22222	22222	22222
9	11113			
10	32211,21232	21232, 11113	32211, 11113	21232, 11113
11	22331	32211,22331	21232,22331	32211, 22331
12	22323	22323	22323	22323
13	22233	22233		22233
14	33321	33321	33321,22233	33321
15				
16	ID, UNC	ID, UNC	ID,UNC	ID,UNC
17	33333	33333	33333	33333
18				

Note: ID = Immediate Death, UNC = UNConscious.

The valuation exercise

In the valuation exercise respondents were asked to value the same health states as in the ranking exercise, with the same time frame, on a scale ranging from 0 (worst) to 100 (best). The results are shown in table 6, where it is seen that the valuations of the health states followed the same pattern as in the ranking exercise: the more serious the health state was, the lower the value. In all four samples, and in total, the health state 11111 was assessed to a value close to 100. At the other end of the scale, for all four samples and in total, the health state 33333 was assessed to have a value close to 0. In all four samples and in total, the health state ID was assessed to be better than the health state 33333, indicating this latter health state to be worse than death.

Table 6. Mean value for EQ-5D health states in the valuation exercise. (n = 580).

Health states	Sample 1	Sample 2	Sample 3	Sample 4	Total
11111	99.4	98.6	99.9	99.1	99.2
11211	82.8	83.2	81.2	82.0	82.4
21111	81.9	82.3	79.7	80.5	81.2
12111	80.1	81.2	78.7	79.2	79.9
11121	78.6	80.6	77.8	80.2	79.5
11112	71.7	74.1	71.1	74.5	73.1
11122	61.1	63.8	63.0	64.9	63.4
22222	46.6	48.4	47.4	46.6	47.3
11113	37.3	38.2	36.5	38.2	37.7
21232	34.2	34.5	34.1	30.9	33.3
32211	32.3	31.1	32.7	30.6	31.5
22331	29.0	26.2	28.7	26.3	27.3
22323	24.8	25.3	24.1	24.9	24.8
22233	20.5	20.8	20.1	19.1	20.1
33321	17.1	19.1	21.2	19.4	19.2
ID	14.7	14.0	19.2	15.7	15.7
UNC	11.0	11.3	17.4	11.6	12.6
33333	5.3	5.6	4.6	4.9	5.1

Note: ID = Immediate Death, UN = UNConscious.

Self-reported health status using the Visual Analogue Scale (VAS)

The VAS exercise is, like the EQ-5D profile, an instrument that assesses subjective health status, however, it is not limited to a discrete number of attributes. In the VAS exercise, respondents were asked to indicate their own health on a ruler, with end-points of 0 (worst imaginable health state) and 100 (best imaginable health state). The results are presented as mean score, standard deviation, median and minimum/maximum values for all samples, and in total.

The PTO exercise

Since framing effects are a known cause of bias, we designed four different frames - one for each sample - in order to test whether framing of the question has an effect on health state valuations. In total we operated with two times two different representations: PTO-1 (prevention) versus PTO-2 (prevention) and PTO-1 (treatment) versus PTO-2 (treatment). The design of PTO-1 and PTO-2 is based on preliminary work developed by Prades [1999], whereas the scenario for using prevention versus treatment is new. In all four samples respondents were confronted with 16 EQ-5D health states excluding the states ID and UNC. The framing of the four samples was si-milar, but then again also different. The framing of the PTO questions took the following forms.

PTO-1 (prevention) (sample 1, n = 120). The respondents were asked to imagine two *prevention pro-grammes*. However, only one of these prevention programmes could be implemented. Prevention *programme B* could prevent 10 named persons being brought into a life-threatening health state. A life- threatening health state meant that the person would die immediately when entering this health state. Prevention *programme A* could prevent *X* persons being brought into a given health state. The objective was to locate *X*, which consequently was the number of persons that made the respondent indifferent in choosing between the two prevention programmes.

PTO-1 (treatment) (sample 2, n = 170). The respondents were asked to imagine two *treatment programmes*. However, only one of the treatment programmes could be implemented. Treatment *programme B* could treat 10 named persons who were in a life-threatening health state. A life-threatening health state meant that the person would die immediately without treatment. If they did receive treatment they would fully recover and be able to live a normal life. Treatment *programme A* could treat *X* persons who were in a given health state. If these persons did not receive any treatment they would probably spend their remaining lifetime in this health state. However, if treated they would be fully recovered and able to live a normal life. The objective was to locate *X*, which consequently was the number of persons that the respondent judged equivalent in choosing between the two treatment programmes.

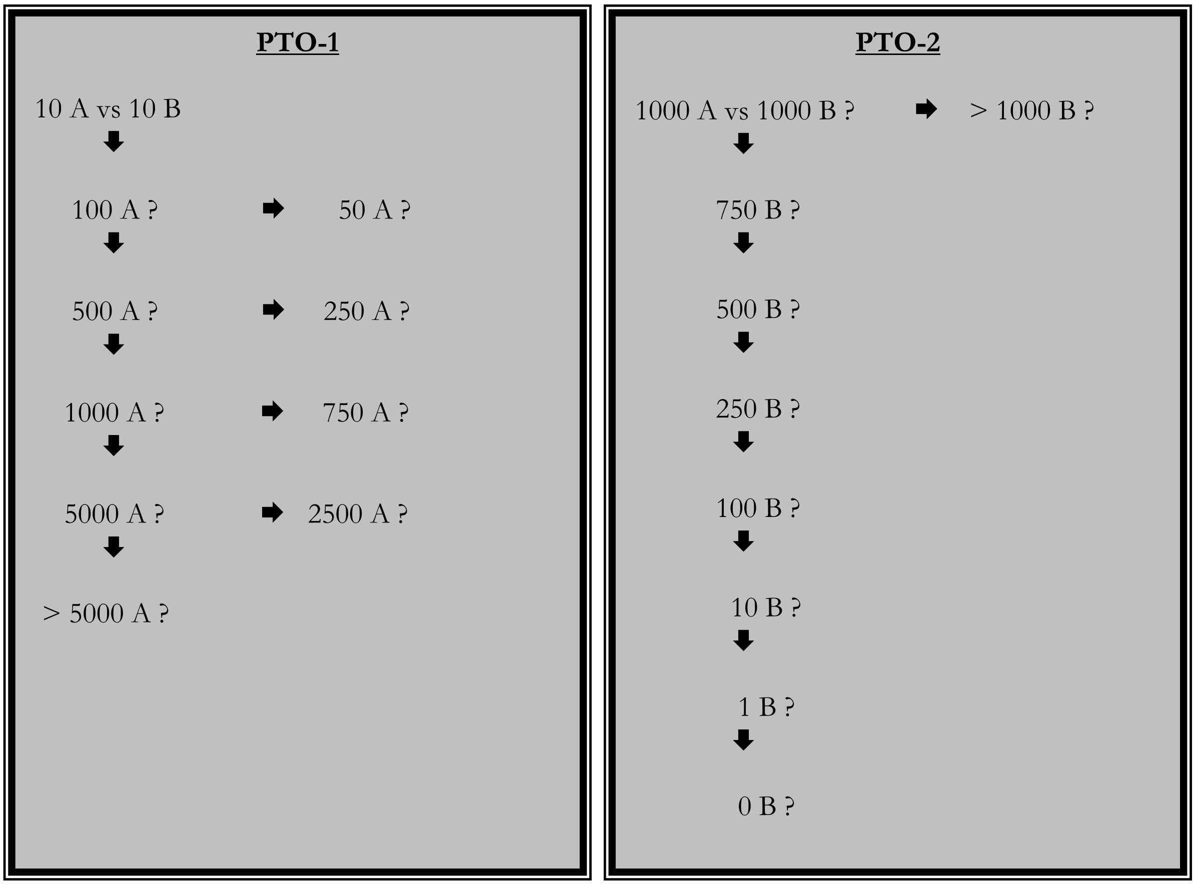
PTO-2 (prevention) (sample 3, n = 120). The respondents were asked to imagine two *prevention pro-grammes*. However, only one of these prevention programmes could be implemented. Prevention *programme B* could prevent *X* persons being brought into a given health state. Prevention *programme A* could prevent 1000 persons being brought into a life-threatening health state. A life-threatening health state meant that the person would die immediately when entering this health state. The objective was to locate *X*, the number of persons that made the respondent indifferent in choosing between the two prevention programmes.

PTO-2 (treatment) (sample 4, n = 170). The respondents were asked to imagine two *treatment programmes*. However, only one of the treatment programmes could be implemented. Treatment *programme B* could treat X persons who were in a given health state. If these persons did not receive any treatment they would probably spend their remaining lifetime in this health state. However, if treated they would be fully recovered and able to live a normal life. Treatment *programme A* could treat 1000 persons who were all in a life-threatening health state. A life-threatening health state meant that the person would die immediately without treatment. If they did receive treatment, they would be fully recovered and able to live a normal life. The objective was to locate *X*, the number of persons that the respondent judged equivalent in choosing between the two treatment programmes.

In all four samples, the respondents were given visual aids (props) to help them visualise the trade-offs they had to make. We modified props used in an earlier TTO study that have also been used to elicit valuations for both the UK TTO EQ-5D tariffs and for the Health Utility Index (HUI) [Dolan *et al.* 1995; Furlong *et al.* 1990].

In order to help respondents make trade-offs, a *ping-pong strategy* for PTO-1 was applied. The respondents were asked to choose between an explicit number of persons instead of facing them with *open-ended* trade-offs.[3] For PTO-2, a simple design where the number of persons to choose was gradually lowered from 1000 to 0 was applied. The trade-off procedures for PTO-1 and PTO-2 are illustrated in Figure 1.

Figure 1. Stepwise procedure used in the elicitation of PTO valuations.



The most common way to ask PTO questions is to apply the PTO-1 approach [Nord *et al.* 1993; Prades 1999; Green 2001]. Initially the focus is only on PTO-1 (treatment) and PTO-2 (treatment). PTO-1 (prevention) and PTO-2 (prevention) are discussed later in this section. An example is required in order to show how this method can be used to calculate the value of a health state. Imagine that the PTO-1 (treatment) technique is used to ask a respondent to locate X. Assume that the respondent is indifferent at X being 50 persons. As the benefit of returning somebody who is in health state A to

[3] The ping-pong strategy was implemented as a consequence of personal communications with Erik Nord, Senior Researcher at the National Health Institute in Oslo, Norway.

perfect health is $1 - A$, and the benefit of returning somebody to perfect health in stead of dying is 1, the respondent makes a trade-off which, put in a formal way, is:

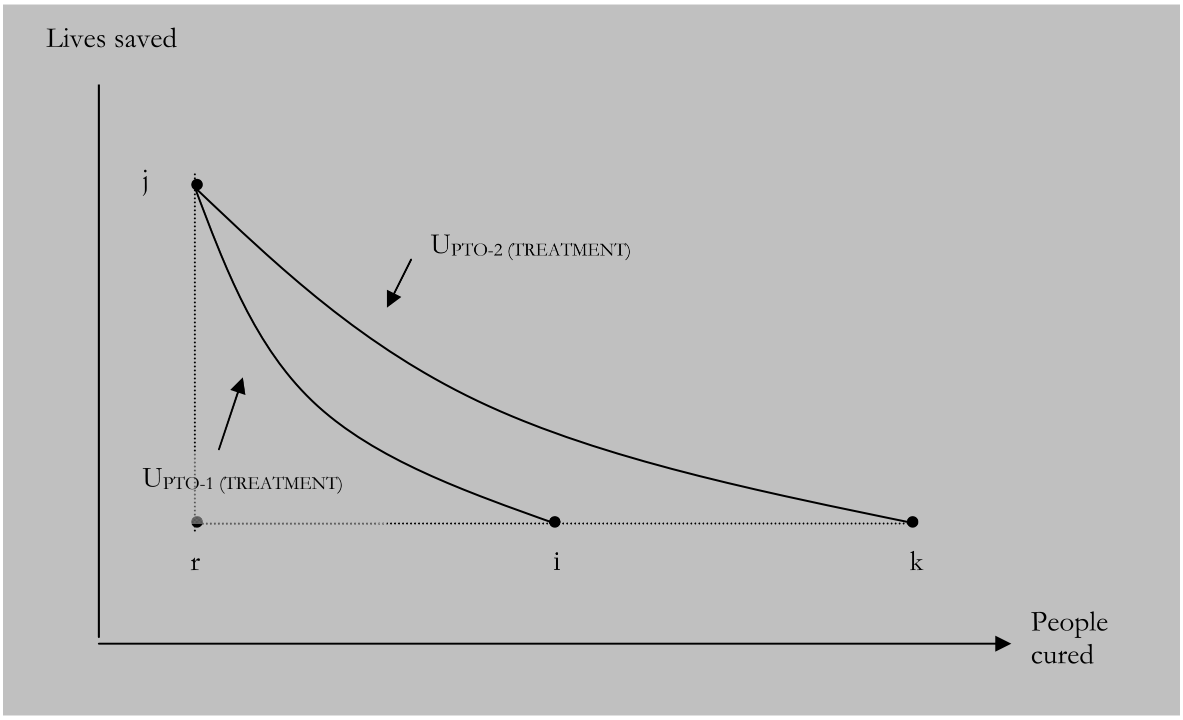
$$50 \times [1 - U(A)] = 10 \times 1. \tag{1}$$

Solving the above equation reveals that $U(A) = 0.8$, which means that the given health state A has a health-utility of 0.8 on a 0 to 1 cardinal scale. The line of reasoning is similar in the case of the PTO-2 (treatment) technique. Here, the trade-off concerned the number of fatalities accepted for returning X persons in health state A to full health. As described in Prades (1999), the PTO-1 (treatment) and PTO-2 (treatment) methods are equivalent, since we are comparing (j, lives) with (k, cured). In the PTO-1 we matched (10, lives) and (?, cured) and in the PTO-2 technique we matched (?, lives) and (1000, cured). Assume that in applying the PTO-2 technique, the respondent states X to be equivalent to 750. In other words the respondent is saying that returning 750 persons from a given health state A to full health is equivalent to the acceptance of letting 1000 persons die. Put more formally:

$$750 \times 1 = 1000 \times [1 - \mathrm{U(A)}]. \tag{2}$$

Where $j = 750$, $U(A) = 0.25$. In PTO-1 we arbitrarily establish $j = 10$ and ask about k. In PTO-2 we establish $k = 1000$ and ask about j. Hence, U(A) ought to be constant across the two contexts. However, where preferences are references-dependent as suggested in prospect theory, starting with j and asking about k may not be the same as starting with k and asking about j [Kahnemann & Tversky 1979; Tversky & Kahnemann 1991]. As suggested by Prades (1999), the PTO-1 (treatment) technique is a matching between two *gains* (saving lives against curing people) whereas in PTO-2 (treatment) the matching is between a *loss* (fatalities accepted) and a *gain* (curing people). According to Tversky & Kahnemann (1991) the shape of the value function is steeper for losses than for gains. Transferring the assumptions of prospect theory to our context, the indifference curve between j and k will have a different slope, which implies that the marginal rate of substitution (MRS) between lives saved and people cured is different. The principle is illustrated in Figure 2.

Figure 2. Reference effect and indifference curves.



In PTO-1 (treatment) we asked a matching question from a references-point referred to as *r*, but in PTO-2 (treatment) the question was asked from the point referred to as *j*. Due to the movement from *j* to *r* being a loss, by loss aversion it had a larger impact than a movement from *r* to *j*. In order to compensate for a loss, a larger increase in the other attribute was needed. In this really was the case, the indifference curve between the two attributes - 'lives saved' and 'people cured' - would change, and more people cured would be needed in order to compensate for one life lost (PTO-2) than for one life not saved (PTO-1).

The PTO-1 (prevention) and PTO-2 (prevention) were included in order to investigate another, but different, framing effect. The objective was to investigate whether it matters if the respondents were confronted with a PTO scenario where people were receiving treatment or a PTO scenario where people were prevented from disease/dying. The hypothesis was that it *did* matter since it seems plausible to believe that respondents have different perceptions of the content/implication of the two words *treatment* and *prevention*. Respondents may put more emphasis on the treatment scenario than on the prevention scenario, since the former deals with people already being in a more or less severe health state compared to the latter where people (currently) were in full health. Where there was a difference regarding the scenario with which the respondents were confronted, the difference would have to show in the valuations of the health states. Due to our hypothesis, the valuations of EQ-5D health states in the treatment scenario would consequently be lower than the valuations in the prevention scenario.

There is evidence that respondents, when asked PTO questions about very mild health states, may state an infinite number of people in order to obtain a level of equivalence.[4] The reason is that some respondents have preferences for preventing persons from dying, no matter how many other persons this prevents from being cured from a given health state. An example could be a mild health state such as *headache*. It seems reasonable that some respondents may not be willing to make a trade-off unless an enormously high number of persons are being cured from headaches in order to let 10 persons die. Consequently, we divided the PTO exercise in two parts. In the first part we asked the respondents to make trade-offs where we presented them with eleven health states excluding the five mildest EQ-5D health states (11211, 11112, 12111, 11121, 21111). In the second part we replaced the wording of one of the alternatives. Instead of saying that 10 persons would die if not treated, they now suffered from *multiple sclerosis,* and if they did not receive any treatment they would probably remain with this condition for the rest of their lives. The respondents were given a full description of the disease and how it influences mobility and other functions for the patient suffering from multiple sclerosis. Since we wanted to present the respondents with the corresponding EQ-5D health states that resemble the health status of an average multiple sclerosis patient, we contacted the *Australian Association of Multiple Sclerosis* for an expert opinion. They replied that the EQ-5D health state 22322 would be the best in describing the health state of the average multiple sclerosis patient.[5] Consequently, this description was applied to all four samples when asking the respondents about valuations for the five mild EQ-5D health states mentioned above.

The weighting for severity-of-illness exercise

In a paper by Nord *et al.* (1999), they propose a theoretical framework using the PTO technique to elicit equity-weights for severity-of-illness and potential-for-health. This line of reasoning and their theoretical framework were adapted and used to investigate whether such weights could be elicited in practice, and if so, what the numerical values would be.

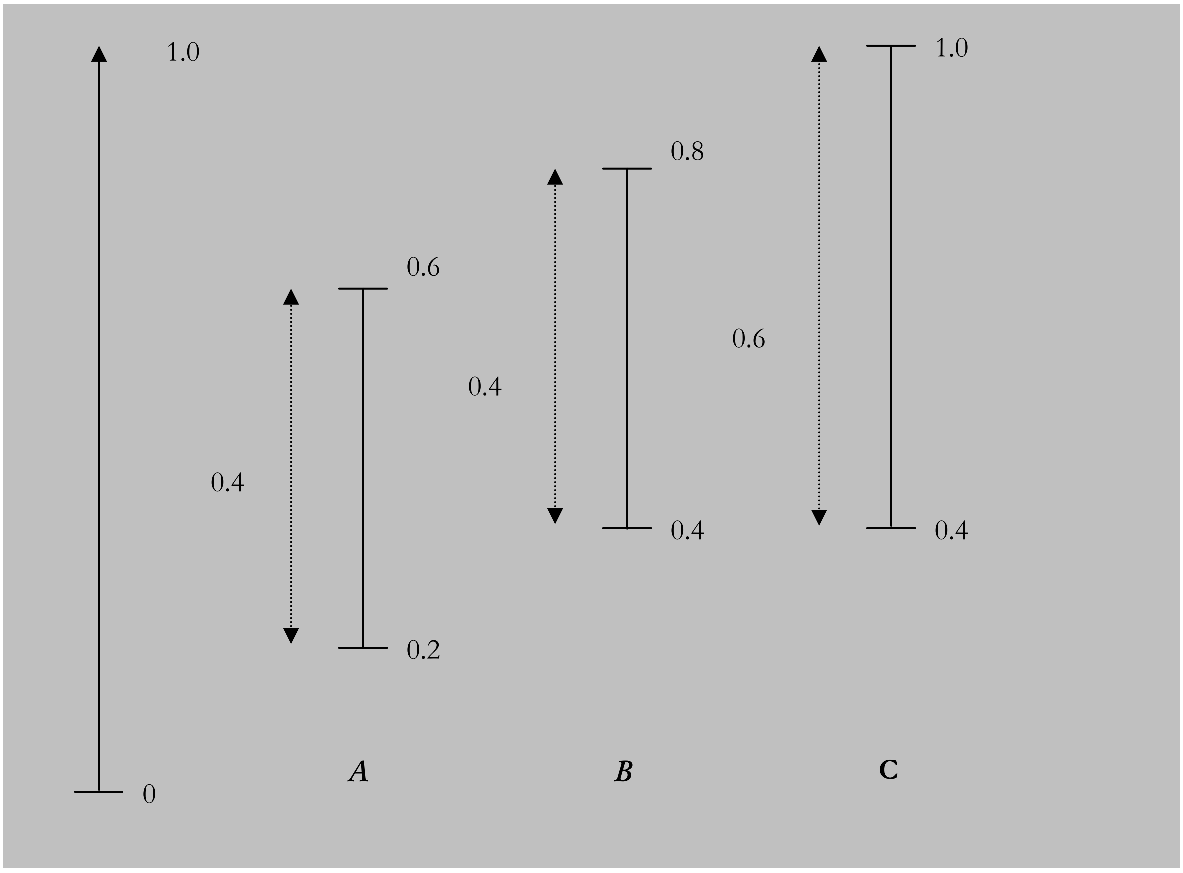
In order to illustrate the problem, Nord *et al.* (1999) developed yet a simple but very illustrative figure that describes the issues. Consider Figure 3, where A, B, and C are three groups of patients (or alternatively - three individuals) who on average are alike in all respects, expect that they suffer from different diseases. Their Health-Related Quality of Life (HRQoL) is measured on an interval scale using a scaling technique. A scale from zero (0) to unity (1) is used to represent the three groups' HRQoL utility.

[4] Personal communication with Erik Nord, Institute of Public Health, Oslo, Norway.

[5] We contacted the Australian Association of Multiple Sclerosis since one of the authors was preparing the interview manuals for the PTO study when he worked as a visiting research fellow in Melbourne, Australia.

The bottom endpoints (starting points) of the three vertical lines in the figure indicate the three groups' utility before the intervention and the top endpoints (finishing points) indicate their utility after the intervention. Assuming that all three groups have the same life expectancy without an intervention and that none of the interventions has any effect on life expectancy, the implication is that improvements in health are purely related to HRQoL improvements. Furthermore, it is assumed that the intervention 'costs per patient' are the same in all three groups.

Figure 3. Three improvements in health on a scale from 0 - 1.0.



Given these assumptions, Figure 3 illustrates that group *A*, without intervention, is more severely ill than both *B* and *C*. This is measured by the lower initial utility at the starting point. Secondly, using the same amount of health care resources for *A* and *B* will produce the same amount of health gain, measured at the individual utility level. Thirdly, while *B* and *C* experience the same low level of HRQoL without an intervention, the same amount of resources will result in higher health gain for *C* than for *B*. This difference is caused by the fact that the three groups differ according to disease, which corresponds differently to the intervention. In other words, one could argue that *C* has a higher potential-for-health than does *B*.

If society wants to maximise health benefits (measured as number of QALYs), *C* will have first priority in getting the intervention, while *A* and *B* will share second place. However, there are two considerations concerning equity, which are important in a priority context of scarce health care resources, which are not in line with the above-mentioned ordering of the three patient groups.

Due to concerns for severity-of-illness and potential-for-health, the former equity consideration implies that patients with relatively more severe diseases ought to be given a higher weight because they are being discriminated against in QALY calculations. This is because an intervention would eventually result in a lower number of QALYs compared with the number of QALYs for *B* or *C*. Consider Figure 3, where *A* is worse off without an intervention than is *B* or *C*, which is why *A*, *ceteris paribus*, has a higher quest for treatment than *B* or *C*, even though the health gain for *A* is just as large as for *B* and larger than for *C*. Another equity issue arises when comparing *B* and *C*. Considering the number of QALYs gained to be most important, *C* ought to receive the intervention before *B*. However, it may be argued that this order discriminates against *B,* since this group has a lower potential to take advantage of the intervention compared to *C*.

In theory, it is possible to estimate equity-weights for both severity-of-illness and potential-for-health that can be incorporated in the calculation of QALYs, as shown in Nord et al. (1999). Thus QALYs would comprise both a utilitarian as well as an egalitarian aspect (all groups treated equally regardless of the utility derived from the intervention), which may be more in line with how society believes that scarce health care resources ought to be distributed across patient groups. To date most of the work that tries to establish equity weights in the allocation of health care resources has been of a theoretical nature [Wagstaff 1991; Dolan 1996; Bleichrodt 1997; Williams 1997; Dolan 1998], and only two studies have looked into the elicitation of weights for severity-of-illness and/or potential-for-health [Nord 1995; Cabasés *et al.* 1999]. In another study, a different yet similar method has been reported by Dolan (1998), where the objective was, from an empirical point of view, to investigate whether is was possible to estimate social values (i.e. a social welfare function). This study was, however, merely a pilot study, since it was conducted among 35 third-year economics undergraduates at an English university [Dolan 1998].

In order to capture these two considerations for fairness in the standard calculation of QALYs, Nord et al. (1999) have proposed a multiplicative social welfare model, which could take into account the societal value of an improvement from utility level U_1 to utility level U_2. The functional form of the multiplicative model is:

$$SV = \mathrm{d}U \times SW \times PW \tag{3}$$

where *SV* is the societal value, d*U* represents the utility gain ($U_2 - U_1$), *SW* is the weight representing the severity-of-illness factor, and *PW* is the weight representing the potential-for-health factor. *SV* is measured on a common 0 to 1 interval scale. In order to achieve this, the *SW* factor must be set at 1 for $U_1 = 0$ (being in a life-threatening health state) and increase with the increasing values of U_1 where *SW* finally = 0 for $U_1 = 1$. However, the functional form of the utility is not known and has not been investigated in other studies.

The PW factor is more complicated. In order to estimate weights for PW one either has to devalue utility gains for patients with relatively high potential-for-health, or revalue utility gains for patients with relatively low potential. Nord *et al.* (1999) decided to base the theoretical fra-mework on the latter since this makes it easier to keep the measurement of SV within the conventional 0-1 range. The ratio $(U_2 - U_1)/(1 - U_1)$ represents the ratio between the actual potential for a given group of patients and the theoretically highest possible health gain, where the highest possible health gain is the movement from the initial level to full health (U_1 to 1). The ratio is called the *relative potential ratio* (RPR). In order to keep *SV* within the scale endpoints of 0 and 1, *PW* needs to be set at 1 when $U_1 = 0$ and $U_2 = 1$, given that d*U* and *SW* are then both equal to 1. In addition *RPR* in this case equals 1, which results in *PW* = 1 when *RPR* = 1. Since patients with less potential have to be given a higher weight, *PW*, all other things being equal, must increase with falling values of the *RPR*.

An illustration of the line of reasoning when estimating weights for *SW* and *PW* is given in Table 7. The table illustrates the utility gain for four different health care interventions *A* to *D*. The *SW* and *PW* weights presented here are only illustrative. With the arbitrary weights chosen, the most attractive intervention is *C*, since it results in the highest societal value (*SV*). In we did not adjust for fairness considerations and were only interested in maximizing the utility gain, intervention *A* would be the most attractive of the four interventions. Applying fairness weights to the QALY calculations results in both interventions *B* and *C* becoming more attractive than *A*.

Table 7. Illustrative example of how to calculate the societal value for four different interventions.

Intervention	**Utility**			**SW**	**RPR**	**PW**	**Societal value (SV)**
	Initial (U_1)	End (U_2)	Utility gain				
A	0.60	0.85	0.25	0.20	5/8	1.90	0.095
B	0.65	0.75	0.10	0.15	2/7	6.50	0.098
C	0.60	0.80	0.20	0.20	1/2	2.50	0.100
D	0.70	0.90	0.20	0.10	2/3	1.85	0.037

When deciding how to elicit both severity-of-illness and potential-for-health weights the theoretical guidelines provided by Nord *et al.* (1999) were followed. All four samples were asked questions concerning SW and PW, which were used to estimate fairness weights. By applying a *paired comparison method* it was possible to cover the whole range of the 0 to 1 utility scale. To estimate SW fairness weights the respondents were asked to make PTO trade-offs, where they were to compare utility movements on the 0 to 1 scale. As in the PTO exercise, visual aids were used to help the respondents make the trade-offs. The respondents were presented with PTO questions of the following kind:

"Imagine that there exist two treatment programmes. Only one of these programmes can be implemented. The first programme (programme I) can treat 10 persons, who are all in the worst health state, similar to level F. The treatment will bring these persons to level E. The second alternative (programme II) can treat X number of persons, who are all in a health state corresponding to level E. If treated, these persons are brought to level D. The costs of the two treatment programmes are the same. How many persons should programme II comprise in order for you to feel that the two programmes are equivalent?"

The respondents were presented with three additional, similar, trade-off questions, where only the starting and ending points for programme II were changed, but the total interval between the start and end utility levels in programme II did not change.[6] The alternative, as described for programme I, remained the same throughout the remaining three trade-off questions. An example: assume that the respondent states that X should be 30 persons, which means that bringing 10 persons from level F to level E is equivalent to bringing 30 persons from level E to level D. Consequently, the SW can be calculated to be 0.33 (10/30). In total, direct valuations of severity weights for four initial utility levels on the 0 to 1 scale were obtained. The SWs for the remaining levels were estimated using linear extrapolation.

Potential-for-health weights were estimated in a similar way to SW. The respondents were faced with similar, yet slightly different, trade-off questions. Again the respondents were presented with visual aids. The trade-off question for estimating PW was as follows:

"Imagine again that there exist two treatment programmes. Again, only one of these programmes can be implemented. The first programme (programme I) can treat 10 persons, who are at the health state E. The treatment will bring these persons to level A. The alternative treatment programme (programme II) can treat X number of persons, who likewise are at the health state level E. If treated, they

[6] It was explicitly assumed here that the scale had ratio properties and interval scores were excluded so that they were not visible to the respondents. In practice the 0 to 1 scale was implicitly divided into 0.2 intervals, resulting in the following set up: Level A = 1.0, level B = 0.8, level C = 0.6, level D = 0.4, level E = 0.2 and level F = 0.0. Later in this chapter it will be tested empirically whether such an assumption is valid.

can be brought to level B. The costs of the two treatment programmes are the same. How many persons should programme II comprise in order for you to feel that the two programmes are equivalent?"

In the PW exercise, respondents were asked to make trade-offs between movements that were equal in terms of severity (starting point), but different in terms of utility gains. Again, paired comparisons were chosen to cover the whole range of possible utility gains on a 0 to 1 scale. In total, direct PW valuations for three RPRs were obtained.[7] The remaining weights were calculated using linear extrapolation.

To calculate direct weights of potential-for-health is a little more complicated than in the case of severity-of-illness weights. To illustrate the mathematics behind the potential-for-health weights consider the following three movements, illustrated in Table 8 on the 0 to 1 utility scale. The corresponding RPR ratios were arbitrarily chosen.

Table 8. Calculation of potential-for-health weights (PW): An illustration.

Movement	Utility	RPR
I	Level E ➔ A (0.2 ➔ 1.0)	1/1
II	Level E ➔ B (0.2 ➔ 0.8)	3/4
III	Level E ➔ C (0.2 ➔ 0.6)	1/2
IV	Level E ➔ D (0.2 ➔ 0.4)	¼

Using the arbitrary numbers illustrated in the table above, the respondents were asked how many persons X brought from level E to B were equivalent to bringing 10 persons from level E to A. Let us say that the respondent gave a figure for *X* of 20 persons. Then the potential weight (PW) for movement II, for which RPR equals ¾, would be calculated as follows:

$$20 \times (0.8 - 0.2) \times PW_{II} = 10 \times (1.0 - 0.2) \times PW_{I}.$$

Since it is known that PW_I is 1, PW_{II} is estimated to be 0.67, which is then the potential weight for movements for which RPR equals ¾. Similarly, the potential weights for movements III and IV are calculated in the same manner.

[7] As in the SW exercise it was again explicitly assumed that the scale used to elicit weights for potential-for-health had ratio properties and the interval scores were not visible to the respondents.

The distribution exercise

In this exercise the respondents were presented with two direct questions instead of trade-off questions. The aim was to investigate whether respondents' preferences for the allocation of health care resources support utilitarian or egalitarian theory. The focus was also on whether there were any significant differences regarding person characteristics in how respondents chose to answer. All respondents were faced with the first question:

"Imagine two diseases I and II. They influence the same number of people and result in the same serious illness. If treatment is given to patients with disease I, they may feel ***some improvement*** *in their health, while if treatment is given to patients with disease II, they may feel a* ***substantial improvement*** *in their health. The costs of the two treatments are the same. The treatment capacity is too low for both diseases and it has been proposed to increase the budget."*

<u>There were two solutions:</u>

1) The majority of the increase in the budget should be allocated to the treatment of disease II since the treatment effect is higher, that is, the treatment is more cost-effective.
2) The increase in the budget should be divided equally between the two diseases since patients in both cases are suffering from severe illness and have the same right to be treated.

Which of the two solutions do you agree with the most?
A. Prioritise disease II, since it has the largest treatment effect.
B. Divide equally.
C. Not sure/unable to answer the question.

After finishing the above question, all respondents were faced with a similar, yet different, question:

"Imagine that I is a disease, which results in severe illness, while II is just as widespread a disease, which results in moderate illness. Treatment will result in patients suffering from disease I experiencing ***some improvement,*** *while patients suffering from disease II will experience a* ***substantial improvement.*** *The costs of the two treatments are the same. The treatment capacity is too low for both diseases and it has been proposed to increase the budget."*

<u>There were three solutions:</u>

1) Most of the increase in the budget should be allocated to treatment of disease II, as the treatment effect is highest.

2) Most of the increase in the budget should be allocated to treatment of disease I, as the patients suffering from this disease are worst off initially.
3) The budget should be divided equally between the two diseases.

Which of the three solutions do you agree with the most?
A. Prioritise II, since this has the largest treatment effect.
B. Prioritise I, as patients here are suffering the most.
C. Divide equally.
D. Not sure/unable to answer the question.

Weighting exercise - supplement

As mentioned earlier, when the respondents were faced with the weighting exercises, the visual aids did not show the interval numbers, but in fact were divided by 0.2 intervals on the 0 to 1 scale. The reason for not showing the respondents the intervals and only the levels was that we wanted to investigate whether the respondent's perception of the 0 to 1 scale was indeed that of an interval scale with equal intervals between the levels. A way to do this was simply to ask the respondent about the values they believed each level (A to F) should have on the 0 to 1 scale. The exercise resembled a category scale where respondents had to indicate a value for each level relative to the degree of the other levels.

Results

Personal characteristics

The mean age of the total sample was around 45 years, where the youngest/oldest respondent was 18/89 years. The results are illustrated in Table 9.

Table 9. Personal characteristics and assessment of exercises.

	Total sample (n = 580)
Age:	
Mean (SD)	44.9 (17.2)
Median	42.0
Min./Max.	18/89
Years in school:*	
≤ 7 years	13.6 %
8 – 10 years	47.1 %
≥ 11 years	39.3 %
University or college degree:	
Yes**	11.4 %
No	88.6 %
*Income per month before tax***:*	
< 8,000 DKK	19.7 %
8,000 – 14,999 DKK	24.5 %
15,000 – 19,999 DKK	17.0 %
20,000 – 29,999 DKK	26.0 %
30,000 – 39,999 DKK	9.0 %
40,000 – 49,999 DKK	2.7 %
	1.2 %

≥ 50000 DKK	
Problems with the ranking exercise:	
No	58.3
Yes, some	32.9
Yes, a lot	8.8
Don't know	0.0
Problems with the valuation exercise:	
No	63.6
Yes, some	29.8
Yes, a lot	6.6
Don't know	0.0
Problems with the PTO exercise:	
No	54.0
Yes, some	34.7
Yes, a lot	11.2
Don't know	0.2

*Primary and high school.

**Education lasting five years or more.

***1 DKK = 0.13745 Euro.

The majority had 8-10 years of primary school and around 40 per cent had a high school qualification. 11 per cent had a university degree. Around 20 per cent of the sample had a monthly income before tax of less than 8,000 DKK and approximately 13 per cent had an income of 30,000 DK or more. The majority of respondents indicated that they had no problems with the ranking, valuation, or PTO exercises. Relatively more respondents had problems with the ranking exercise than with the valuation exercise. Not surprisingly, more respondents had problems with the PTO exercise compared to both the ranking and valuation exercises. Around 11 per cent indicated that they had a lot of problems with the PTO exercise.

Health status using Visual Analogue Scale (VAS)

In total the mean VAS score was 90.3 with a standard deviation of 12.3. The lowest VAS score was 25 and the highest 100. There were no large differences across samples. The results are illustrated in Table 10.

Table 10. Results from the VAS exercise in each sample and total. (n = 580).

VAS scores	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**	**Total**
	(n = 120)	**(n = 170)**	**(n = 120)**	**(n = 170)**	**(n = 580)**
Mean (SD)	92,7 (8.5)	87.8 (14.8)	90.6 (11.5)	90.9 (12.1)	90.3 (12.3)
Median	95.0	90.0	95.0	95.0	95.0
Min./Max.	60/100	25/100	40/100	30/100	25/100

The PTO exercise

All 580 respondents valued 16 EQ-5D health states. However, due to different wording in the PTO questions, we had to interpret the sample as consisting of four independent samples. The direct valuations for PTO-1 (prevention) and PTO-2 (prevention) are illustrated in Table 11. The direct valuations for PTO-1 (treatment) and PTO-2 (treatment) are illustrated in Table 12.

Table 11. PTO valuations of EQ-5D health states in samples 1 & 3. (n = 120).

	PTO-1 (prevention)		**PTO-2 (prevention)**		
Health States	**Mean (SD)**	**Median**	**Mean (SD)**	**Median**	***P-value****
11112	0.8705 (0.3193)	0.99700	0.8655 (0.3191)	0.99800	0.87
11121	0.8463 (0.3470)	0.99800	0.8726 (0.3089)	0.99800	0.35
11211	0.8024 (0.4569)	0.99700	0.8607 (0.3287)	0.99800	**0.06**
12111	0.8743 (0.3094)	0.99800	0.8756 (0.3095)	0.99800	0.97
21111	0.8371 (0.3551)	0.99800	0.8276 (0.3625)	0.99800	0.78
11122	0.7143 (0.4010)	0.96000	0.6291 (0.4504)	0.90000	**0.04**
11113	0.6968 (0.4178)	0.90000	0.6449 (0.4449)	0.90000	0.20
21232	0.6670 (0.4360)	0.96000	0.6371 (0.4473)	0.90000	0.47
22222	0.6607 (0.4305)	0.90000	0.6460 (0.4379)	0.90000	0.71
22233	0.6993 (0.4189)	0.93000	0.6443 (0.4447)	0.90000	0.18
22331	0.6677 (0.4267)	0.90000	0.6751 (0.4308)	0.90000	0.85
22323	0.6390 (0.4410)	0.90000	0.8260 (0.3175)	0.96000	**0.00**
32111	0.7508 (0.3825)	0.90000	0.6705 (0.4357)	0.90000	**0.05**
33321	0.6575 (0.4292)	0.90000	0.6406 (0.4572)	0.96000	0.69
33333	0.6028 (0.4568)	0.85000	0.6318 (0.4600)	0.96000	0.49
UN	0.6873 (0.4211)	0.90000	0.5999 (0.4690)	0.90000	0.04

P-values with bold script *(p < 0.10).

Note: UN = unconscious.

*One-sample *t*-test.

Table 12. PTO valuations of EQ-5D health states for samples 2 & 4. (n = 170).

	PTO-1 (treatment)		**PTO-2 (treatment)**		
Health States	**Mean (SD)**	**Median**	**Mean (SD)**	**Median**	***P-value****

11112	0.8822 (0.2838)	0.99900	0.8940 (0.2459)	0.99000	0.54
11121	0.8750 (0.2954)	0.99900	0.9051 (0.2325)	0.99900	0.10
11211	0.8623 (0.3147)	0.99900	0.8770 (0.2910)	0.99900	0.52
12111	0.8866 (0.2781)	0.99900	0.9091 (0.2325)	0.99900	0.21
21111	0.8837 (0.2903)	0.99900	0.9112 (0.2258)	0.99900	0.12
11122	0.7437 (0.3610)	0.99000	0.7469 (0.3240)	0.90000	0.90
11113	0.7577 (0.3651)	0.99000	0.7751 (0.3251)	0.99000	0.49
21232	0.7649 (0.3604)	0.99000	0.8082 (0.2942)	0.99000	**0.00**
22222	0.7217 (0.3662)	0.90000	0.7630 (0.3205)	0.90000	0.10
22233	0.7432 (0.3645)	0.99000	0.7499 (0.3221)	0.90000	0.79
22331	0.7420 (0.3692)	0.99000	0.7804 (0.3192)	0.99000	0.13
22323	0.7724 (0.3500)	0.99000	0.7959 (0.3078)	0.99000	0.33
32111	0.7963 (0.3394)	0.99000	0.8678 (0.2750)	0.99900	**0.00**
33321	0.7575 (0.3584)	0.99000	0.7878 (0.3132)	0.99000	0.22
33333	0.7335 (0.3615)	0.99000	0.7574 (0.3272)	0.90000	0.35
UN	0.7357 (0.3767)	0.99000	0.7322 (0.3520)	0.90000	0.90

P-values with bold script ($p < 0.10$).

Note: UN = unconscious.

*One-sample *t*-test.

As illustrated in Table 11, using the PTO (prevention) scenario resulted in 12 valuations out of 16 EQ-5D health states being statistically different (at the 10 per cent level) when using either the PTO-1 (prevention) or the PTO-2 (prevention) technique. There was a clear trend that the more severe the health state, the lower the (mean) valuation on the 0 to 1 scale. The lowest state was ‘unconscious’, valued at a mean value of approximately 0.60 in the PTO-2 (prevention) scenario. Both in the PTO-1 (prevention) and PTO-2 (prevention) scenarios the standard deviation was quite high, indicating a large deviation in valuations across respondents. There was no clear trend regarding one scenario yielding higher valuations than the other.

Table 12 is similar to table 11, illustrating results from the two other scenarios PTO-1 (treatment) and PTO-2 (treatment). Here, in 14 out of the 16 valuations, the difference between the two techniques was statistically insignificant at the 10 per cent level. There was again a clear trend that the worse the health state, the lower the corresponding value on the 0 to 1 scale. In 15 out of the 16 valuations the PTO-2 (treatment) technique resulted in higher valuations than the PTO-1 (treatment) technique.

Table 13 illustrates the results from the comparison of the EQ-5D mean valuations between PTO-1 (prevention) versus PTO-1 (treatment) and PTO-2 (prevention) versus PTO-2 (treatment). In 6 out of the 16 valuations in the PTO-1 scenario there was a significant difference (at the 10 per cent level), and four of the remaining health states had relatively low *p*-values, indicating that it *could* matter whether the question was framed as a ‘prevention’ scenario or as a ‘treat-ment’ scenario. In the case of the PTO-2, there *was* a significant difference (at the 10 per cent level) between 11 of the 16 valuations, supporting the idea that it does matter how the question was framed.

The respondents valued all 16 EQ-5D health states in the PTO-1 (treatment) scenario higher (judged by mean value) than in the PTO-1 (prevention) scenario, which could indicate that respondents put more emphasis on preventing people who are in full health from being ill, than on treating people who are already ill. This tendency is also seen in the PTO-2 (prevention) versus the PTO-2 (treatment) scenario, where 15 EQ-5D health states had a higher mean value.

Table 13. Comparisons of mean EQ-5D valuations across 'prevention' and 'treatment' scenarios.

.Health states	PTO-1: prevention versus treatment	PTO-2: prevention versus treatment
11112	0.69	0.33
11121	0.37	0.25
11211	**0.09**	0.59
12111	0.67	0.24
21111	0.15	**0.01**
11122	0.43	**0.01**
11113	0.11	**0.00**
21232	**0.02**	**0.00**
22222	0.12	**0.00**
22233	0.25	**0.01**
22331	**0.06**	**0.01**
22323	**0.00**	0.35
32111	0.20	**0.00**
33321	**0.01**	**0.00**
33333	**0.00**	**0.00**
UN	0.21	**0.00**

P-values with bold script ($p < 0.10$).

*One-sample *t*-test.

The weighting exercise

The weighting exercise proved to be feasible and resulted in direct valuations for four initial utility levels. The remaining severity weights were elicited applying linear extrapolation and based on the direct valuations. The valuations based on extrapolation are illustrated in brackets in Table 14. There was a clear tendency that the lower the initial utility level, the higher the corresponding (mean) severity weight. In Table 14 the column 'theoretical' severity weights covers the corresponding severity weights proposed by Nord *et al.* (1999). These severity weights were purely arbitrary and not based on empirical evidence. Comparing these 'theoretical' severity weights with our 'empirically' based severity weights shows that when the initial utility was relatively low, the empirical weights were lower than the theoretical weights, and vice versa when the initial utility is higher.

Table 14. Severity weights for initial utility on a 0 to 1 scale.

Initial utility	Severity weight	'Theoretical' severity weights[a]
0.0	1.0000	1.00
0.1	(0.7930)	0.80
0.2	0.5850	0.65
0.3	(0.5209)	0.50
0.4	0.4568	0.40
0.5	(0.4189)	0.30
0.6	0.3810	0.20

0.7	(0.3670)	0.10
0.8	0.3430	0.05
0.9	(0.1715)	0.01
1.0	0.0000	0.00

[a] The term 'theoretical' severity weights covers the illustrative weights proposed by Nord *et al.* (1999) p 32.

Note: The weights in brackets were elicited by linear extrapolation.

The potential weights for each RPR level are illustrated in Table 15. In total, direct potential weights for three RPR levels were elicited. Again, the remaining potential weights were estimated by using linear extrapolation. These valuations are illustrated in brackets in the table. The lower the RPR, the higher the corresponding potential weights. Due to no reference level (lower bound) for the RPR level, it was impossible to elicit potential weights for the two RPR levels 0.2 and 0.1. As in Table 14, the column 'theoretical' potential weights covers arbitrary potential weights proposed by Nord *et al.* (1999). Comparing our 'empirical' potential weights with the 'theoretical' potential weights shows that for all RPR levels the empirical potential weights were much higher.

Table 15. Potential weights of different relative potential ratios.

RPR	Potential weight	'Theoretical' potential weight[a]
1.0	1.0000	1.00
0.9	(1.2524)	1.05
0.8	(1.5049)	1.15
0.75	1.6311	-
0.7	(1.8200)	1.30
0.6	(2.0089)	1.45
0.5	2.5757	1.60
0.4	(4.2993)	1.80
0.3	(6.0229)	2.00
0.25	6.8847	-
0.2	(?)	2.50
0.1	(?)	4.00

[a] The term 'theoretical' potential weights covers the illustrative weights proposed by Nord *et al.* (1999) p 32.

Note: The weights with brackets were elicited by linear extrapolation.

The priority exercises

The majority of the respondents, around 70 per cent, wanted to divide the increase in the budget equally across the two alternatives, implicitly indicating that they did not wish to prioritise the alternative giving the largest health gain, and hence went for equity. There were no large deviations across the four samples. The results are illustrated in Table 16.

Table 16. Priority exercise A. (Per cent).

Answers	Sample I	Sample II	Sample III	Sample IV	Total
A. Prioritise alternative II due to largest effect	35 (29.2)	45 (26.5)	33 (27.5)	51 (30.0)	164 (28.3)
B. Divide equally	81 (67.5)	120 (70.6)	82 (68.3)	117 (68.8)	400 (69.0)
C. Unable to answer	4 (3.3)	2 (1.2)	5 (4.2)	2 (1.2)	16 (2.8)
Total	120 (100.0)	170 (100.0)	120 (100.0)	170 (100.0)	580 (100.0)

In the second priority exercise, respondents were again confronted with an alternative resulting in the largest health gain and the alternative dividing the increases in the budget equally between the two. However, a third alternative was added where respondents could choose to give priority to those patients that were worst off. As in the former priority exercise the majority of the respondents, around 50 per cent, chose to divide the budget equally and around 32 per cent thought that the worst off patients should be given priority. The remaining 18 per cent chose to give priority to the alternative resulting in the largest effect. The results are illustrated in Table 17.

Table 17. Priority exercise B. (Per cent).

Answers	Sample I	Sample II	Sample III	Sample IV	Total
A. Prioritise alternative II due to largest effect	24 (20.0)	33 (19.4)	22 (18.3)	26 (15.3)	105 (18.1)
B. Prioritise alternative I due to severest patients	41 (34.2)	53 (31.2)	36 (30.0)	55 (32.4)	185 (31.9)
C. Divide equally	52 (43.3)	81 (47.6)	59 (49.2)	86 (50.6)	278 (47.9)
D. Unable to answer	3 (2.5)	3 (1.8)	3 (2.5)	3 (1.8)	12 (2.1)
Total	120 (100.0)	170 (100.0)	120 (100.0)	170 (100.0)	580 (100.0)

It was investigated whether respondents' preferences correlated with socio-demographic characteristics. The results are shown in Table 18. In both exercises, there was a significant difference at the 5 per cent level for age, implying that, in general, respondents who chose to divide equally were older then the other respondents. Gender also had a significant effect at the five per cent level. Females had a significantly higher tendency to divide equally than males. In distribution exercise A, respondents who spent fewer years in school had a statistical higher tendency at the 10 per cent level to divide equally, than respondents who had a high school qualification. Although the *p*-value for 'income before tax' was significant at the 10 per cent level, the results were inconclusive, since no matter the income level, the majority wished to divide equally. In distribution exercise B, the *p*-value for 'years in school' was significant at the 1 per cent level, indicating that here also more respondents who had spent fewer years in school chose to divide equally. The *p*-value for 'income per month' was significant at the 1 per cent level. In order to double-check the results shown in Table 18, we performed a logistic regression analysis. This exhibited the same results.

Table 18. Distribution of health care resources and socio-demographic characteristics.

	Distribution exercise A			Distribution exercise B			
	Priori-tise II	Divide equally	*P*-value	Priori-tise II	Priori-tise I	Divide equally	*P*-value
Age (mean)	43.4	45.6	0.01**	46.1	42.4	45.5	0.00****
Gender (%):							
Male	35 %	65 %	0.04***	23 %	42 %	35 %	0.02***
Female	25 %	75 %		15 %	31 %	54 %	
Years in school:*							
≤ 7 years	19 %	81 %	0.09***	16 %	20 %	64 %	0.00***
8 – 10 years	29 %	71 %		21 %	28 %	51 %	
≥ 11 years	33 %	67 %		17 %	42 %	41 %	
Income per month (before tax):							
≤ 14,999 DKK	28 %	72 %	0.09***	17 %	28 %	55 %	0.00***
15,000-29,999 DKK	26 %	84 %		16 %	40 %	44 %	
30,000-49,999 DKK	45 %	55 %		16 %	40 %	44 %	
≥ 50,000 DKK	20 %	80 %		17 %	50 %	67 %	

*Both primary and high school.

**One-sample *t*-test.

***χ^2-test.

****One-way ANOVA.

The weighting exercise – supplement

Respondents, when faced with the weighting exercise, were not told where each level was placed on the 0 to 1 scale. However, implicitly the scale was divided into intervals of 0.2 and hence a secondary objective was to test whether it was acceptable to anticipate that the respondents also interpreted the scale as an interval scale, where each level was divided with a 0.2 interval. The results are presented in Table 19.

Judged by median value, the responses corresponded very well to the pre-assumptions made about the scale. Mean and median values were quite close even though there were a considerable range of valuations as illustrated by the minimum and maximum. However, given the low standard deviations, these must be regarded as outliers.

Table 19. Scoring level A to F on the 0 to 1 scale using category scaling (n = 580).

Scale	Mean (SD)	Median	Min./Max.
Level A	97.81 (8.40)	100.0	10/100
Level B	73.93 (14.24)	80.0	9/99
Level C	55.56 (14.28)	60.0	8/97
Level D	35.46 (13.29)	35.0	2/85
Level E	19.78 (11.57)	20.0	1/70
Level F	6.55 (9.84)	5.0	0/65

Discussion

The person trade-off exercise

Previous PTO studies, regardless of their framing, have shown that the compression of the valuations towards the upper end of the 0 to 1 scale is very high compared with other scaling techniques [Ubel *et al.* 1998; Cabasés *et al.* 1999]. The study reported here is no exception. One explanation may be that individuals value poor health relatively higher when the valuations of these states are elicited *via* techniques that measure social preferences, rather than *via* techniques that measure individual preferences. According to Ubel *et al.* (1998) the explanation is that the individual gives a high social value to saving lives compared with curing individuals from poor (even extremely poor) states of health. Nord (1996) has undertaken an experiment using a variety of multi-attribute utility (MAU) instruments compared to the PTO technique, which he interprets as societal values. His findings are illustrated in Table 20. It can be seen that the societal values are much higher than the individual utilities elicited from MAU-instruments across all three levels. The line of reasoning which Nord (1996) uses is that the MAU-instruments elicit utilities that are too low and do not correspond with societal values, and thus are inappropriate as input in the calculation of QALYs. In the study reported here, societal values correspond quite well with the findings reported by Nord (1996). However, the difference between societal values and MAU utilities may be caused by reasons other than simply the PTO technique being capable of eliciting 'true' societal values whilst MAU instruments do not.

Table 20. Societal values for health states versus individual utilities from MAU-instruments.

Instrument		**Problem level**[a]	
	Severe	**Considerable**	**Moderate**
Societal values	0.65-0.85	0.90-0.94	0.98
QWB	0.45-0.55	0.65-0.70	<0.80
HUH	0.10-0.20	0.30-0.40	<0.85
HUHI	0.40	0.70	0.90-0.94
EQ-5D	0.20	0.60	0.70
York EuroQol (TTO)	0.20-0.25	0.40-0.50	0.80
IHQL (3D)	0.50-0.70	0.75-0.85	0.89-0.93
IHQL (complex)	0.70-0.75	0.80-0.90	0.90-0.94
15D	0.77	0.86	0.91-0.93
Rosser/Kind	0.68	0.94	0.97-0.98

Adapted from Nord (1996).

[a] The three states were described as follows:

Severe:	Sits in a wheel-chair, has pain most of the time, is unable to work.
Considerable:	Uses crutches for walking, has light pain intermittently, is unable to work.
Moderate:	Has difficulties in moving about outdoors and has slight discomfort, but is able to do some work and has only minor difficulties at home.

In designing the four independent PTO studies a ping-pong strategy was applied which presented the respondents with close-ended questions, i.e. a steady number of pre-defined responses. In the PTO-1 scenarios, the respondents could state that the equivalence level was at 10 persons. If they did so, they were implicitly saying that they felt that the health state was equal to a life-threatening state and consequently gave the value of 0 (zero). If they felt that 100 persons was too high a number and consequently settled at 50, the health state would be given the value of 0.80. In other words, the respondents implicitly can not, apart from a value of 0.0, give a health state a value lower than 0.80. This is a serious limitation in our design, which evidently must compress the (mean) valuations toward the upper end of the scale. In the PTO-2 scenario, this limitation was not present. However, the respondents in the PTO-2 scenario could not express preferences for health states between 0.25 and 0.0, simply due to the design, which also caused a serious limitation in eliciting true preferences for health.

The obvious question now is whether using an open-ended questionnaire would remove these limitations? The answer is yes, since using an open-ended design, *ceteris paribus*, would implicitly give the respondents the chance to be anywhere on the 0 to 1 scale. However, as shown by Cabasés *et al.* (1999), using an open-ended design did not change the mean valuations significantly. This is an interesting result since it gives some support to the proposition that our design is not flawed. However, there may be other limitations to using an open-ended design, since in reality respondents could judge equivalence to be from 0 to ∞. This may be difficult for the respondents to comprehend, that is, to find equivalent numbers of persons with respect to preferences for a health state.

In the study reported here negative valuations, that is, health states worse than death were not allowed. However, this could easily be accomplished empirically, as shown in the elicitation of Danish time trade-off (TTO) tariffs [Pedersen *et al.* 2003]. It is impossible to say anything conclusive about how negative valuations would affect the mean PTO valuations. It probably would lower them, however.

In addition to the aim of eliciting societal values using the PTO technique, four different frames were applied in order to test for possible framing effects, since there is empirical evidence that different frames result in different valuations [Prades 1999]. In the study by Prades (1999), the predictive power of three PTO frames and other scaling techniques (visual analogue scale and standard gamble) was tested and it was found that a direct comparison of the intervals showed a higher predictive power at the individual level.[8]

The study reported here gives two important findings: (1) based on the majority of the EQ-5D health states, people do not value gains and losses differently, and (2) based on the majority of the EQ-5D health states, people do value health states differently depending on whether a 'prevention' or a 'treatment' scenario is considered. In other words, people are more willing to allocate extra health care resources to treat people already than prevent persons who are in full health from becoming ill. The first finding implies that our results do not correspond with prospect theory, which suggests that preferences are reference-dependent. The same conclusion is reached by Prades (1999). The second finding shows that the results are in line with our hypothesis, i.e. people have preferences for helping persons who are already ill rather than spending extra resources on preventing the same number of persons in full health from being ill. The result is not surprising. Empirical studies have shown that people in general have strong preferences for curing persons [Ubel 1998].

Although many researchers are in favour of the PTO method due to its attempt to elicit social value instead of individual utility, others do not support the method. An example is Dolan (1998), who points out that the PTO method is unable to separate individuals' relative weights for at least four different aspects: (1) pre-treatment severity of illness, (2) post-treatment severity, (3) gains in health from treatment, and (4) number of people treated. Dolan argues that using the PTO method makes it impossible to establish 'the definition of need' that is the most appropriate in terms of resource allocation. Further, he believes that the PTO method imposes a cognitive overload that affects the validity of the respondents' trade-offs. Instead he supports using individual estimates to construct a health welfare function, which may incorporate equity considerations, e.g. log-linear inequality-aversion functions.

Dolan's criticism of PTO is based on purely intuitive theoretical grounds and not on any empirical evidence. The findings presented here indicate that more or less the same amount of respondents find it 'a lot of trouble' to understand the PTO method as do respondents confronted with the TTO method. Our findings show that over half the respondents in this study said that they had no problems making PTO trade-offs, and only around 11 per cent said that they had 'a lot of trouble' understanding the exercise. Second, the design of the PTO study was based on the design of the Danish TTO study, using the same props, health states etc. [Pedersen *et al.* 2003]. In the TTO study, around 9 per cent of the respondents stated that they had 'a lot of trouble' understanding the exercise. We agree that trade-off questions may be too difficult for some respondents to comprehend, but not that the PTO is more cumbersome than the TTO. That the PTO method makes it impossible to separate individuals' relative weights can only be tested qualitatively by asking respondents what they think about when making the trade-offs. Like Dolan, we urge such a study, but before the results have been put on the table, it is difficult to say anything conclusive.

The weighting exercise

In considering the benefits of health care programmes where many patients or potential patients are targeted, the QALY calculation - according to Bryan et al. (2002) - would typically take account of at least four distinct features or characteristics: (1) the number of patients receiving the intervention, (2) the probability of the intervention being successful, (3) the survival gain if successful, and (4) the gain in quality of life if successful. The QALY method has been criticised for being a poor measure of health-related quality of life (HRQoL) since only one aspect of its constituent features, the one relating to quality of life, is based upon preferences [Dolan 1998; Nord 1994]. Thus far there has been little emphasis, both theoretical and empirical, on distributional preferences and preferences relating to risk in the calculation of QALY scores. In general, two sets of distributional concerns relating to the construction of QALYs can be distinguished: (A) those relating solely to the number of people who receive treatment (i.e. preferences for more beneficiaries rather than fewer in terms of a wider distribution of health benefits, regardless of who receives them), and (B) those relating to the personal characteristics of the individuals who receive treatment (e.g. the level of pre-treatment severity or potential-for-health).

In the study reported here, the focus is solely on (B), which has been the case for the majority of previous studies [Williams 1997; Nord *et al.* 1999]. However, recently Bryan et al. (2002) undertook an empirical study where the focus was on (A). Using the QALY method (as conventionally constructed) implicitly assumes that the *societal value* for a health intervention is directly proportional to the number of patients receiving treatment, and the level of risk associated with treatment. An example: an intervention providing health care for 10 patients will have a QALY score (i.e. societal value) that is twice that for an intervention providing care for 5 patients. The study by Bryan et al. investigated the robustness of these assumptions relating to the constant marginal returns of a QALY maximisation approach. By applying conjoint analysis they found that public preferences were not much at odds with the core proportionality assumptions concerning societal value in the QALY maximisation model assumptions. However, their results were at odds with reports from various previous studies.

[8] See Prades (1999) for an in-depth discussion.

As noted by many researchers, society's overall valuation of health output is a function not only of total output (QALYs gained), but also of the distribution of health output across individuals. As stated by Nord *et al.* (1999): *'... society may be prepared to make some sacrifices in the total production of health in order to secure a fair or equitable distribution of health ...To encapsulate such distributive concerns, economists have proposed to assign equity weights to QALYs according to characteristics of their recipients'.* In other words, the aim of resource allocation in health care would be to maximize the sum of equity-weighted QALYs rather than an unweighted sum.

This part of our study is largely based on the theoretical framework of Nord *et al.* (1999), in which they suggest theoretically how to estimate weights for severity-of-illness and potential-for-health. If such equity weights are to be incorporated into the calculation of QALYs, they propose the term *health-related societal value.*

We believe, as do Nord *et al.*, that the QALY method (in its conventional form) is flawed since it is unable to incorporate concerns for other than the maximization of health.[9] In other words, in its present form, the QALY method does not represent societal preferences and consequently the QALY score is not a social value. Using the theoretical framework suggested by Nord et al. (1999), we estimated equity weights for both severity-of-illness and potential-for-health. In both cases we came up with consistent results. That is, the lower the initial utility level the higher the severity weight, and the lower the RPR the higher the potential weight. Comparing our empirical weights with the theoretical weights suggested by Nord *et al.* (1999) revealed a mixed pattern.

For severity weights we found that the corresponding weight for an initial utility level of 0.1 was 0.7930. The theoretical weight was suggested to be 0.80. However, as the initial utility level became higher, the differences between our empirically-based and the theoretical-based weights became larger. We found that the corresponding weight for the initial utility level 0.9 was 0.1715, where the theoretical weight was suggested to be as low as 0.01. Our results show that even though the initial utility level was close to 1 (full health), the respondents had preferences for giving it a relatively high weight.

In the case of estimating weights for potential-for-health, we found the corresponding weight for the RPR 0.9 level to be 1.25, close to the theoretical suggested weight of 1.05. However, as the RPR level became lower the differences became larger. The corresponding empirical weight for the 0.3 RPR level was 6.02, much higher than the theoretical weight at 2.00. The results showed that respondents held strong preferences towards patients having a relatively low potential benefit from a given intervention.

[9] When discussing maximisation of health care we explicitly think about the maximisation of health care in society as a whole, and not the maximisation of health in a given group of patients, since there is a difference between the two scenarios. While the objective within a group of patients is to maximise health, due to the budget this may not be the case for the overall health for the population as a whole. Here one may have to take equity considerations into account when allocating health care resources.

The distribution exercise

The framing of the distribution exercise used in this study has previously been tested among Norwegian decision-makers [Nord 1993]. Our findings are very similar. When asking respondents to allocate an increase in the health care budget between two alternatives - one maximising health gain (utilitarian) and the other dividing equally across patient groups (egalitarian) - over two-thirds of the respondents chose the egalitarian option. In the second distribution exercise, we added a third alternative, which would help the worst-off patients. Again, the majority (around half of the respondents) would divide equally across patient groups, whilst around 30 per cent of respondents would give priority to the worst-of patients, and fewer than 20 per cent would maximise health gain.

Our findings add further evidence to previous studies, that respondents have strong egalitarian preferences when allocating scarce health care resources, thus not providing much support for utilitarianism, that is, the distribution of health care based solely on the maximisation of health gain. Our background material shows that the reluctance to maximise health care is not the same across socio-demographic characteristics. For example, females are more inclined to divide equally than are males.

Also, as suggested by Nord (1993), the distribution exercise asked about *attitudes*. That is, the respondents had to choose from a discrete number of alternatives. A more specific approximation would have to been to ask respondents to choose between a dozen alternatives, varying the number of persons affected, etc. The problem with this solution is that each alternative has to be very precisely specified, which may impose a cognitive overload on the respondents.

Since the majority of respondents 'failed' to put most emphasis on helping the worst-off patients, it may be argued that there is no need to incorporate equity weights, i.e. fairness considerations, in the calculation of QALYs [Walker & Siegel 2002]. These authors discuss these issues based on the empirical findings made by Nord (1993) and others. Walker and Siegel, correctly, see the move to incorporate preferences for social values into the calculation of QALYs to be inspired in part by a commitment to justice and fairness in the allocation of scarce health care resources. They state that: *"In so far as the authors we have discussed are motivated by the desire to find a practical way to incorporate such social values into health care allocation schemes, we have a great deal of sympathy for their work. While we do not want to deny that social values may play a legitimate role in formulating allocation policies, we stick to the claim that the use of SVPs [social value preferences] needs to be justified. Without this justification, we worry that the move from CUAs [cost-utility analyses] to SVPs is just a move from an implicit reliance on a questionable utilitarian standard to an explicit reliance on the popularity of moral values."* In other words what Walker and Siegel are missing is a non-circular way of showing that social values are indeed represented by SVPs.

Martin *et al.* (2002) undertook a study concerning 'accountability for reasonableness' which is a framework by which the fairness of priority setting in health care can be evaluated.[10] They conducted an empirical study identifying elements of fairness by decision-makers engaged in priority setting for new technologies in Canada (a primarily publicly funded system).[11] Their aim was not to identify social value preferences (or functions), but to address elements of fairness that are important to health care decision-makers in setting priorities, that is the required elements for priority-setting be fair. They identified eleven specific elements of fairness, which are illustrated in Table 21.

Table 21. Accountability for reasonabless and fairness according to decision-makers.

Conditions of accountability for reasonabless	Elements of fairness according to decision-makers
(1) Publicity	External transparency
(2) Relevance	Multiple perspectives
	External consultation
	Consensus
	Honesty
	Identify potential conflict of interest
(3) Appeals	Appeals mechanism
(4) Enforcement	Leadership
	Internal transparency
	Understanding
	Opportunity to express views
	Agenda setting

Source: Adapted from Martin et al. (2002) p.4.

Fairness is relative, which means that the elements of fairness illustrated in Table 20 are based on average perspective. Martin et al. (2002) found that none of the decision-makers they interviewed identified any elements of fairness that conflicted with or were missing from accountability for reasonabless. They concluded that this provides evidence that the conditions of accountability for reasonabless are familiar and acceptable within that context.

It is not surprising that decision-makers wants transparency in order to make fair decisions in setting priorities in health care. The implications for researchers who want to identify social value functions are that we have to be very explicit on how to incorporate equity weights into the calculation of QALYs. As shown earlier, the incorporation of the two equity weights severity-of-illness and potential-for-health does not complicate the calculation process very much. However, when taking other equity

[10] Accountability for reasonabless is based on the work of Daniels & Sabin (1999) and states that health care institutions engaged in priority setting have a claim to fairness if they satisfy four conditions: (1) rationales for priority setting decisions must be publicly accessible (*publicity condition*), (2) these rationales must be considered by fair-minded people to be relevant to priority setting in that context (*relevance condition*), (3) there must be an avenue for appealing these decisions and their rationales (*appeals condition*), (4) these must be some means, either voluntary or regulatory, of ensuring that the first three conditions are met (*enforcement condition*).

[11] Since the framework of incorporating fairness considerations into priority setting was developed in the context of a primarily privately funded setting, the aim of their study was to assess its applicability in a primarily publicly funded system.

weights, e.g. age, into account, one could easily make the process cumbersome and incomprehensible to outsiders, e.g. decision-makers.

Menzel et al. (1999) report on a study involving 150 Norwegian politicians "accountable for health policy at the county level" and the result was that there was no consensus on how to divide scarce health care resources. The politicians were asked to choose an allocation of health care resources between treatment alternatives that would offer a little help to a group with severe illness, or treatments that would offer considerable help to a group with moderate illness. The results showed that 45 per cent chose to divide equally regarding resources, 37 per cent gave priority to the worst-off, and 11 per cent gave priority to those suffering from moderate illness. According to Walker and Siegel these findings give *"... most assuredly evidence that this group generally favours helping the least well off first (even when this will not maximize utility) it could not possibly count as evidence for the presumed consent of any particular politician to such an allocation scheme. Similarly, for any given SVP, one can easily assume that the individual value preferences that go into the calculation of the 'societal value' may diverge from one another by wide margins."* In other words, while rationing is necessary, presumed consent based on an appeal to SVPs does not seem to offer adequate moral justification for such rationing. The same issues have been discussed by Menzel (1990), who states that *"in reality we question only a sample of people to establish our basic map of proportional quality ratings, and we question only a relatively small sample of patients to place them on the spectrum of health states."*

Although using a sample may lead to the same results as a general vote in terms of 'winning' preferences, the democratic procedural justification, according to Walker and Siegel (2002) is lost. Of course basing results on a sample instead of the whole population will never be as robust (or valid) and on this point we agree with the criticism put forward by Walker & Siegel (2002). However, we do not believe that samples are useless simply because they never 100 per cent represent the true preferences of the population seen as a whole.

Nord *et al.* (1999) have proposed a two-step procedure. In the first step actual patients are asked to determine utility measures for their particular conditions, and the second step is to use these utility measures when asking a representative and randomised sample of the general public to determine preferences for hypothetical allocation schemes. However, their proposal is based on purely theoretical arguments and has not yet been applied in an empirical setting. Nevertheless, we urge that this line of reasoning be tested empirically.

In conclusion, the results from this study more or less follow previous findings, even though this study is the first of its kind, i.e. randomly drawn respondents from the general publication and also included the highest number of respondents reported thus far. We urge that more PTO studies are performed in eliciting valuations for health states. The next step would be to perform a representative PTO study to elicit valuations for EQ-5D health states, and then model a set of Danish EQ-5D tariffs similar to the Danish TTO estimates to see how they correspond to each other [Pedersen et al. 2003]. However, this would demand a new design, since we feel that the design used in this study put too many limitations on which values the respondent could assign to any given EQ-5D health state.

Since the aggregated QALY model fails to take distributive effects into account, the approach remains contentious. Here we have presented how one can find equity weights empirically for severity-of-illness and potential-for-health and subsequently incorporate these fairness weights into the calculation of QALYs. The result may indeed be a 'super QALY'. As much as we believe in this approach to the calculation of QALYs, there may be more equity weights that ought to be incorporated into the calculation process, e.g. age-related weights where younger people are given a relatively higher weight than elderly people, since younger people may benefit from the improvement in HRQoL in relatively more years. Nevertheless, we feel that this study adds important evidence that the PTO method is a valid alternative to other valuation techniques such as the TTO or the SG. To what degree the PTO method is indeed capable of estimating social values - where the TTO and SG estimates are individual values - is still too early to postulate. In order to do so, much more evidence is required - both theoretical and empirical.

References

1. Bleichrodt H. Health utility indices and equity considerations. *Journal of Health Economics* 1997; 16(1): 65-91.
2. Boyle MH, Torrance GW, Sinclair JC, Horwood SP. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *New England Journal of Medicine* 1983; 308: 1330-1337.
3. Brooks R, with the EuroQol Group. EuroQol: The current state of play. *Health Policy* 1996; 37: 53-72.
4. Broome J. *Weighing goods*. Blackwell: Oxford. 1991.
5. Bryan S, Roberts T, Heginbotham C, McCallum A. QALY-maximisation and public preferences: results from a general population survey. *Health Economics* 2002; 11(8): 679-693.
6. Cabasés JM, Ugalde JM, Gaminde I. *Societal perspective on the eliciting of health states preferences*. Paper presented at the 16th Plenary Meeting of the EuroQol Group. Sitges, Spain. 6th – 9th November 1999.
7. Cabasés JM, Gaminde I, Ugalde JM, Pozo F. *Social elicitation of EQ-5D health state preferences through Person Trade Off (PTO).* Paper presented at the 17th Plenary Meeting of the EuroQol Group. Pamplona, Spain. 28th – 29th September 2000.
8. Cohen B. Utility measurement and the allocation of health care resources. *Medical Decision Making* 1995; 15: 287-288.
9. Culyer AJ. The normative economics of health care finance and provision. *Oxford Review of Economic Policy* 1989; 5: 34-58.
10. Danish Institute for Health Technology Assessment. *National strategy for Health Technology Assessment*. The National Board of Health. 1996.
11. Dolan P, Gudex C, Kind P, Williams A. *A social tariff for EuroQol: Results from a UK general population study*. Discussion Paper 138. Centre for Health Economics. The University of York. 1995.
12. Dolan P. *The allocation of benefits in health care: Does equity matter?* Paper for the first World Conference of the International Health Economics Association, Vancouver, May 1996.
13. Dolan P. The measurement of individual utility and social welfare. *Journal of Health Economics* 1998; 17: 39-52.
14. Dolan P & Green C. Using the person trade-off approach to examine differences between individual and social values. *Health Economics* 1998; 7: 307-312.
15. Dolan P & Olsen JA. *Desperately seeking numbers: The not-so-holy grail of the 'super-QALY'*. Mimeo University of Sheffield. 1999.

16. Dolan P & Cookson R. A qualitative study of the extent to which health gain matters when choosing between groups of patients. *Health Policy* 2000; 51: 19-30.

17. Drummond MF, O'Brien B, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes*. Oxford University Press: Oxford. 1997.

18. Eddy DM. Oregon's methods: Did cost-effectiveness analysis fail? *Journal of the American Medi-cal Association* 1991; 266: 2135-2141.

19. Feeny D, Furlong W, Boyle M, Torrance G. Multi-attribute health status classification systems: Health Utilities Index. *PharmacoEconomics* 1995; 7: 490-502

20. Fischhoff B. Value elicitation: Is there anything in there? *American Psychologist* 1991; 46: 835-847.

21. Furlong W, Feeny D, Torrance GW, Barr R, Horsman J. *Guide to design and development of health-state utility instrumentation*. Working Paper No. 90-9. Centre for Health Economics and Policy Analysis, McMaster University. 1990.

22. Green C. On the social value of health care: What do we know about the person trade-off technique? *Health Economics* 2001; 10: 233-243.

23. Gold MR, Siegel JE, Russell BR et al. *Cost-effectiveness in health and medicine*. Oxford University Press: New York. 1996.

24. Gudex C. *QALYs and their use by the Health Service*. Discussion Paper No. 20. Centre for Health Economic, University of York. 1986.

25. Hadorn DC. Setting health care priorities in Oregon: Cost-effectiveness meets the rule of rescue. *Journal of American Medical Association* 1991; 265: 2218-2225.

26. Hershey J, Kunreuther H, Schoemaker P. Sources of bias in assessment procedures for utility functions. In Bell D, Raiffa H, Tversky A (eds) *Decision making*. Cambridge University Press: Cambridge. 1991.

27. Jonsen AR. Bentham in a box: Technology assessment and health care allocation. *Law, Medicine and Health Care* 1986; 14: 172-174.

28. Kahnemann D & Tversky A. Prospect theory: An analysis of decision under risk. *Econometrica* 1979; 4: 263-291.

29. Kahnemann D, Slovic P, Tversky A. *Judgment under uncertainty. Heuristics and biases.* Cambridge University Press: Cambridge. 1982.

30. Kaplan RM & Anderson JP. The general health policy model: An integrated approach. In Spilker B (eds) *Quality of life and pharmacoeconomics in clinical trials*. Lippincott-Raven: Philadelphia. 1996.

31. Lauer J. *Person-trade-off methods and the globen burden of disease: Implications for health-preference structures*. Poster presented at the 3rd World Meeting of the International Health Economics Association (iHEA). World Health Organization. 2000.

32. Libscomb J. Time preference for health in cost-effectiveness analysis. *Medical Care* 1988; 27: 233-253.
33. Loomes G & Sugden R. Regret theory: An alternative theory of rational choice under uncertainty. *Economic Journal* 1982; 92: 51-60.
34. Loomes G & McKenzie L. The use of QALYs in health care decision making. *Social Science and Medicine* 1989; 28: 299-308.
35. Martin DK, Giacomini M, Singer PA. Fairness, accountability for reasonabless, and the views of priority setting decision-makers. *Health Policy* 2002. (in press)
36. McCord M & Neufville R. Lottery equivalents: Reduction of the certainty effect problem in utility assessment. *Management Science* 1986; 32(1): 56-60.
37. Mehrez A & Gafni A. Quality-adjusted life years, utility theory and healthy-years equivalents. *Medical Decision Making* 1989; 9: 142-149.
38. Menzel P. *Strong medicine.* Oxford University Press: Oxford. 1990.
39. Mooney G & Olsen JA. QALYs: Where next? In McGuire A et al. (eds) *Providing health care: The economics of alternative systems of finance and delivery.* Oxford University Press: London. 1991.
40. Morris J & Durand A. *Category rating methods: Numerical and verbal scales.* Mimeo. Centre for Health Economics, University of York. 1989.
41. Mulley AG. Assessing patients' utilities. Can the end justify the means? *Medical Care* 1989; 27: 269-281.
42. Murray CJL & Lopez AD (eds). *The global burden of disease.* Harvard University Press: Harvard School of Public Health and WHO/World Bank. 1996.
43. Nord E. The significance of contextual factors in valuing health states. *Health Policy* 1989; 13: 189-198.
44. Nord E. The validity of a visual analogue scale in determining social utility weights for health states. *International Journal of Health Care Planning and Management* 1991; 6: 234-242.
45. Nord E. Methods for quality adjustment of life years. *Social Science and Medicine* 1992A; 34: 559-69.
46. Nord E. An alternative to QALYs: The saved young equivalent (SAVE). *British Medical Journal* 1992B; 305: 875-877.
47. Nord E. *Helsepolitikere ønsker ikke mest mulig helse per krone* (Health care decision makers are not interested in maximizing health per krone). Tidsskrift for Norsk Lægeforening 1993A; 11: 1371-1273. (In Norwegian)
48. Nord E. The trade-off between severity of illness and treatment in cost-value analysis of health care. *Health Policy* 1993B; 24: 227-238.

49. Nord E, Richardson J, Macarounas K. Social evaluation of health care versus personal evaluation of health states. *International Journal of Technology Assessment in Health Care* 1993; 9: 463-478.

50. Nord E. The QALY – a measure of social value rather than individual utility? *Health Economics* 1994; 3: 89-93.

51. Nord E. The person-trade-off approach to valuing health care programmes. Medical Decision Making 1995; 15: 210-208.

52. Nord E, Richardson J, Street A, Kuhse H, Singer P. Who cares about cost? Does economic analysis impose or reflect social values? *Health Policy* 1995; 34: 79-94.

53. Nord E, Pinto JL, Richardson J, Menzel P, Ubel P. Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics* 1999; 8: 25-39.

54. Nord E. *Cost-value analysis in health care*. Cambridge University Press: Cambridge. 1999.

55. Olsen JA. Persons vs. Years: Two ways of eliciting implicit weights. *Health Economics* 1994; 3: 39-46.

56. Olsen JA. A note on eliciting distributive preferences for health. *Journal of Health Economics* 2000; 19: 541-550.

57. Olsen JA, Dolan P, Richardson J, Menzel P. *The moral relevance of personal characteristics in setting health care priorities*. Paper presented at the 21st Nordic Health Economists' Study Group Meeting. Lund, Sweden, August 25th – 26th. 2000.

58. Pedersen KM, Wittrup-Jensen KU, Brooks R, Gudex C. Valuation of health. The theory of quality adjusted life-years and a Danish application. Syddansk Universitetsforlag, 2003. (In Danish)

59. Richardson J. Cost utility analysis: What should be measured? *Social Science and Medicine* 1994; 39(1): 7-21.

60. Treadwell JR & Lenert LA. Health values and prospect theory. *Medical Decision Making* 1999; 19: 344-352.

61. Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Value measurement in cost-utility analysis: Explaining the discrepancy between rating scale and person trade-off elicitations. *Health Policy* 1998; 43: 33-44.

62. Ubel PA. How stable are people's preferences for giving priority to severely ill patients? *Health Economics* 1999; 49: 895-903.

63. von Neumann J & Morgenstern O. *Theory of games and economic behavior*. Princeton University Press: Princeton. 1953.

64. Wagstaff A. QALYs and the equity-efficiency trade-off. Journal of Health Economics 1991; 10: 21-41.

65. Wakker P & Stiggelbout A. Explaning distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making* 1995; 15: 180-186.

66. Walker RL & Siegel AW. Morality and the limits of societal values in health care allocation. *Health Economics* 2002; 11: 265-273.

67. Weinstein MC & Stason WB. Foundations of cost-effectiveness analysis for health analysis and medical practice. *New England Journal of Medicine* 1977; 296: 716-721.

68. Weinstein M, Fineberg H, Elstein A. *Clinical decision analysis*. W.B. Saunders: Philadelphia. 1980.

69. Williams A. Intergenerational equity: An exploration of the 'fair innings' argument. *Health Economics* 1997; 6(2): 117-132.

70. Williams A. Economics of coronary artery bypass grafting. *British Medical Journal* 1985; 291. 326-329.

71. Williams A. Ethics and efficiency in the provision of health care. In Bell JM & Mendus S (eds) *Philosophy and medical welfare*. Cambridge University Press: Cambridge. 1988.