IS THE PERSON TRADE-OFF A VALID METHOD FOR ALLOCATING HEALTH CARE RESOURCES?

JOSE-LUIS PINTO PRADES*
Department of Economics and Centre for Health Economics, Pompeu Fabra University, Barcelona, Spain.

SUMMARY

The Person Trade-Off (PTO) is a methodology aimed at measuring the social value of health states. It is claimed that other methods measure individual utility and are less appropriate for taking resource allocation decisions. However, few studies have been conducted to test the apparent superiority of the method for this particular kind of decision. We present a pilot study to this end. The study is based on the results of interviewing 30 undergraduate students in economics. We compare two well known techniques, the Standard Gamble and the Visual Analogue Scale, with the PTO. The criterion against which the performance of the methods is assessed is the directly obtained preference about how to establish priorities among hypothetical patients waiting for treatment. Apparently the PTO performed better than the others. We also compare three different frames for the PTO. One of them seems to predict people's preferences. © 1997 by John Wiley & Sons, Ltd.

INTRODUCTION

One of the main characteristics of QALYs is the use of quality of life weights (health state utilities) to estimate the benefit of different health policies. Choosing among health policies implies comparing the benefit that these policies will have for different groups of people. Some authors1–7 have pointed out that the usual techniques (Rating Scales, Time Trade-Off, Standard Gamble) used to calculate health state utilities contain important problems. One of the main problems is that they ask questions about how people value their own health and from this valid conclusions cannot be drawn about how society wants to balance health benefits for different groups of the population. For example, Menzel7 argues 'suppose one has expressed a preference for a shorter, healthier life over a longer, less healthy one... Has one consented to anyone saving the shorter but healthier life of one person rather than the longer, lower quality of another?' Nord2, 5 argues that society not only values the size of the health status improvement but the severity of the initial state as well. This concern for the worst off would not be picked up by the techniques already mentioned. Cohen6 concludes that 'society may want to direct

*Correspondence to: Jose-Luis Pinto Prades, Department of Economics and Centre for Health Economics, Pompeu Fabra University, Edifici Jaume I, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain.
Contract/grant sponsor: Fondo de Investigaciones Sanitarias; Contract/grant number: 0767/95
Contract/grant sponsor: Merek, Sharp and Dohme.

© 1997 by John Wiley & Sons, Ltd.
resources preferentially to those who are farthest from good health, even if larger aggregate benefits could be obtained under a different distribution'.

One method proposed by some of these authors to overcome this problem is the Equivalence of Numbers or Person Trade-Off (PTO) technique. Richardson describes this technique as follows: ‘in this technique respondents are asked a question of the following kind: “if there are x people in adverse health situation A and y people in adverse health situation B, and if you can only help (cure) one group, which group would you choose?” One of the numbers x or y can be varied until the subject finds the two groups equivalent in terms of needing or deserving help. The undesirability (disutility) of situation B is x/y times as great as that of situation A’. PTO has the advantage of asking the right question. If the values of the health states were used to make trade-offs between people, the best thing to do would be to ask that question directly. ‘Better than standard gamble would be equivalence of numbers, whereby trade-offs between different people’s lives are clear. Best of all would be explicit QALY bargain questions. Without some such explicit link between the questions used to establish quality indices and the allocations generated by QALYs, QALYs will remain persistently suspicious’.7

The main objective of this paper is to test this claimed superiority of the PTO in measuring health state utilities for the purpose of resource allocation. Our method follows an approach based on decision theory. We take individual preferences as given; so our task is to select the method that best reflects these preferences. In our case, we both estimate quality indices using two well-known techniques such as the Visual Analogue Scale (VAS) and the Standard Gamble (SG) and also use the PTO. We compare resource allocation decisions that would have been taken using these three techniques and examine how well they predict the preferences about resource allocation decisions directly elicited from the public. We will call this the predictive power of the methods.

A second objective of the paper is to compare different frames for the PTO. As we have said, one of the hypothetical advantages of the method is that it asks the right question. However, asking the right question is not enough in all contexts. We have to be very cautious with PTO answers, given the extensive literature showing that preference elicitation methods can be subject to important biases.8–10 These problems are particularly important when we try to elicit preferences for questions that are very unfamiliar to ordinary people.11 Nord has documented some biases in the method. Can we find other important biases in the way we ask PTO questions? We use three different frames and check if the numbers obtained may be influenced by the framing of the questions.

In the next section we describe the three frames we have used for the PTO. We will describe them in some detail because some are fairly new in the literature. We also suggest how the theory predicts the direction of the potential biases among the three frames. Given that the VAS and the SG are well known techniques for health economists we will not explain them in any detail.

**DIFFERENT FRAMES FOR THE PERSON TRADE-OFF**

We used three different representations that we call PTO-1, PTO-2 and PTO-3 (see Appendix 1).

(i) PTO-1: we ask about the number of chronically ill people cured which is equivalent to saving 10 lives. This is the frame most commonly used at the moment. A sample will help to illustrate how this method is used to elicit the value of a health state.

Imagine that somebody is asked how many people who are in a chronic health state (A) should be returned to perfect health in order to produce the same social benefit as returning to perfect health 10 people who were about to die. Imagine that this person says that the number is 100 (an equivalence of 1 to 10). As the benefit of returning to perfect health somebody who is in A is 1–U(A), and the benefit of returning to perfect health somebody who is about to die is 1, this person is saying that

\[
100 \times \left[1 - U(A)\right] = 10 \times 1
\]

We can then say that \(U(A) = 0.9\).

(ii) PTO-2: we ask about the number of fatalities accepted for curing 1000 people with a certain chronic condition.

In theory both methods are equivalent since we are comparing (y, lives) with (z, cured). In PTO-1

we match (10, lives) and (?, cured) and in PTO-2 we match (?, lives) and (1000, cured). If we have estimated in PTO-1 the utility of health state A as $U(A)$, the benefit of preventing $y$ number of fatalities will be equivalent to returning to perfect health $z$ people in A when:

$$y \times 1 = z \times [1-U(A)]$$  \hspace{1cm} (2)

In PTO-1 we arbitrarily establish $y = 10$ and we ask about $z$. In PTO-2 we establish $z = 1000$ and ask about $y$. Of course $U(A)$ needs to be constant across contexts.

However, if preferences are reference-dependent, as Prospect Theory suggests, starting with $y$ and asking about $z$ may not be the same as starting with $z$ and asking about $y$. In PTO-1 the matching is between two gains (saving lives against curing people) whereas in PTO-2 the matching is between a loss (fatalities admitted) and a gain (curing people). Prospect theory suggests that the shape of the value function is steeper for losses than for gains. In our case this would suggest that the indifference curve between $y$ and $z$ would have a different slope in both cases giving rise to a different marginal rate of substitution between lives saved and people cured. We illustrate this in Fig. 1.

In PTO-1 we ask a matching question from a reference point like s, but in PTO-2 the question is asked from a point like y. A s a movement from y to s is judged as a loss, by loss aversion it has a larger impact than a movement from s to y. To compensate for a loss a larger increase in the other attribute is needed. If this is so, the indifference curve among the two attributes will change and more people cured will be needed in order to compensate for one life lost (PTO-2) than for one life not saved (PTO-1).

(iii) PTO-3: this is a several-step procedure. When we have several health states, we first apply PTO-1 to the worst health state. Then we continue asking PTO questions about health states that are adjacent to each other. Assuming that benefits are additive we estimate the value of health states. We illustrate the method with the help of Fig. 2.

In PTO-1 we compare arrows 1, 2, 3 and 4 with 5 (arrows a, b, c, d and e will be explained later). In PTO-3 we first compare arrows 4 and 5, then 3 and 4, then 3 and 2 and then 2 and 1. For example, suppose we have two health states A and B. The worst is B; so using the PTO we find that people are indifferent between saving 1 life and curing 10 B. Then, using the PTO, we compare A and B and discover that people are indifferent between curing 1 B and curing 10 A. We can then say that 1 life is equivalent to 100 A. That would imply that $U(B) = 0.9$ and $U(A) = 0.99$. Formally, we are obtaining $U(B)$ using equation (1) and $U(A)$ assuming that

$$[1-U(B)] = 10 \times [1-U(A)]$$  \hspace{1cm} (3)

that is, we assume that benefits are additive. In PTO-3 we make use of this assumption. If we accept this is true, it should be equivalent to estimate $U(A)$ asking for the number of people $n$ such that

$$n \times [1-U(A)] = 1$$  \hspace{1cm} (4)
or using $U(B)$ as an intermediate step. However, if additivity is violated PTO-1 and PTO-3 may differ.

Another explanation of a possible disparity between PTO-3 and PTO-1 lies in the existence of diminishing sensitivity (DS), which is a characteristic of the value function proposed by Prospect Theory. According to this theory there is an effect called diminishing sensitivity which states that marginal value decreases with the distance from the reference point. For example, the difference between a yearly salary of $60,000 and a yearly salary of $70,000 has a bigger impact when current salary is $50,000 than when it is $40,000.\footnote{In our case, suppose we have three health states A, B, and C, with increasing level of severity. By DS it may happen that an improvement from B to A may have a larger impact when evaluated from B itself than from C. The distance from B to A can be perceived as larger when it is measured directly than when it is measured as a difference between two distances from C. That is,

$$A - B > (A - C) - (B - C)$$

For example, suppose that we ask how many people in B should be returned to perfect health in order to produce the same social benefit as returning 10 people in C to perfect health. Suppose the answer is 100. That is,

$$10 \times [1 - U(C)] = 100 \times [1 - U(B)]$$

Now we ask the same question but with A and C. If B and A are perceived as similar (due to DS) when the reference point is C, the answer should now be close to 100. Let us suppose it is 120. That is,

$$10 \times [1 - U(C)] = 120 \times [1 - U(A)]$$

which implies that

$$100 \times [1 - U(B)] = 120 \times [1 - U(A)]$$

However, if we directly ask the question of how many improvements from A to full health are equivalent to 10 improvements from B to full health it might well be that A and B are not perceived as similarly as before. If this is so, people may ask for more improvements from A to full health in order to compensate for the 10 improvements from B to full health. If, for example, they now say that the number is 20, we have,

$$10 \times [1 - U(B)] = 20 \times [1 - U(A)]$$

Substituting the value of $1 - U(B)$ into equation (6), we have

$$10 \times [1 - U(C)] = 200 \times [1 - U(A)]$$

If we give to $U(C)$ the value of zero we would then have $U(A) = 0.916$ from equation (7) (when the reference point is C) and $U(A) = 0.95$ from equation (10) (when the reference point is B). This suggests that if DS is present, higher PTO values would be obtained with PTO-3 than with PTO-1.

This reference effect is illustrated (Fig. 3). In Figure 3(A) we show the value function proposed by Prospect Theory when reference point is C. In this case the values of B and A are similar. In Figure 3(B), owing to the change in the reference point, both values are much more further apart. In Figure 3(C) both situations are placed together.

### Design of the Study

With the objective of trying to test the predictive power of different preference scaling methods, we personally interviewed 30 undergraduate economics students. The mean duration of the interview was 1.5 h. Questions were asked using four EuroQol (15) health states (see Appendix 2). The students were paid 2000 pesetas (ca $16) for their collaboration. All of them had previously attended a lecture on health status valuation.
We obtained the values of the health states using the three PTO frames, the VAS and the SG. We also asked subjects to order these four health states (say A, B, C and D) from the best to the worst (imagine A is the best and D is the worst). We then asked them the next question: imagine there are five people with health problems. One is about to die, another is in health state D, another in C, another in B and another in A. If you give treatment to one of them he/she will pass to the next best health state: the one who is about to die will be in D for the rest of his/her life, the one who is in D will be in C, the one who is in C will be in B, the one who is in B will be in A and the one who is in A will be in perfect health. If you could only treat one of them, who would be the first, the second, etc.? We can illustrate this question with Table 1.

In Fig. 2 these improvements are represented by the arrows on the left (a, b, c, d and e).

A) Value of B and A fairly close

B) Value of B and A much further apart

C) Change in the value of A when the reference point changes

Figure 3. Diminishing sensitivity and value of health states.
Table 1. Examples of health improvements

<table>
<thead>
<tr>
<th>Individual</th>
<th>Health state before treatment</th>
<th>Health state after treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>About to die</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>Perfect health</td>
</tr>
</tbody>
</table>

What we did was to ask them to compare directly the value of the intervals. We then compared those answers with the priorities that would have been produced using the PTO, the SG and the VAS.

The predictive power will be tested at the individual and at the aggregate levels. We will measure the degree of agreement between the expected ordering and the ordering directly obtained at the individual level. This degree of agreement is measured using Kendall’s τ. This is a statistic that measures the degree of association between two orderings. It oscillates between 1 (maximum agreement) and -1 (maximum disorder) and it is zero when the two orderings are independent.

At the aggregate level we will compare the ranking of the intervals directly obtained and the ranking implicit in the valuations performed by our individuals. We will interpret the ranking performed by the individuals as a voting exercise. Ranking one interval over another will be interpreted as a vote in favour of the former. We will then compare the ‘votes’ obtained by the intervals with the final (aggregate level) sizes of the intervals (distances between the values of health states). We will use an example with the help of Table 1. Suppose that, in Table 1, somebody prioritizes individual 2 (improvement from D to C), then individual 3 (C to B) and then individual 4 (B to A). We will then assume that when having to choose (vote for) 2; between 3 and 4, he/she would choose (vote for) 2; between 3 and 4, he/she would vote for 3 and between 2 and 4, he/she would vote for 2. If individual 2 receives more votes than individual 3 in the rankings performed by our subjects, we would expect that

\[ U(C) - U(D) > U(B) - U(C) \]  \hspace{1cm} (11)

where \( U(B), U(C) \) and \( U(D) \) are the mean values of health states B, C and D. If this is not so, the values obtained for these health states would not be appropriate for taking decisions about who should be prioritized. For example, if \( U(D) = 0.25, U(C) = 0.75 \) and \( U(B) = 0.9 \) more benefit would be obtained with the improvement from D to C than with the improvement from C to B. This would suggest that individual 2 should be prioritized over (ranked higher than) individual 3. As this decision would be taken assuming that \( U(B), U(C) \) and \( U(D) \) reflect society’s preferences, we check the validity of this assumption by asking people to perform this ranking directly.

This method is far from perfect because majority voting does not take into account the intensity of preferences, whereas preference elicitation methods do. In spite of this hypothetical problem we think that this method can shed some light on the predictive power of the different methods used.

RESULTS

We show in Tables 2–5 the results of our experiment. In Table 2 we present the value of the Euroqol health states using the VAS, the SG and the three PTO frames. It can be seen that VAS is lower for all states.

Table 2. Values of the health states using different preference elicitation methods

<table>
<thead>
<tr>
<th>Mean (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAS</td>
</tr>
<tr>
<td>SG</td>
</tr>
<tr>
<td>PTO-1</td>
</tr>
<tr>
<td>PTO-2</td>
</tr>
<tr>
<td>PTO-3</td>
</tr>
</tbody>
</table>

Note: differences between VAS and SG are statistically significant (\( \alpha = 1\% \)). Differences between VAS and PTO are statistically significant (\( \alpha = 1\% \)). Differences between SG and PTO are statistically significant for any PTO. There are no statistically significant differences (\( \alpha = 5\% \)) between PTO-1 and PTO-3 answers. Between PTO-1 and PTO-2 all differences are statistically significant except for health state 21313. Between PTO-2 and PTO-3 all differences are statistically significant except for the case of health state 12121. N=30.
Table 3. Predictive validity of different preference elicitation methods

<table>
<thead>
<tr>
<th>Method</th>
<th>VAS</th>
<th>SG</th>
<th>PTO-1</th>
<th>PTO-2</th>
<th>PTO-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau )</td>
<td>-0.06</td>
<td>0.387</td>
<td>0.393</td>
<td>0.34</td>
<td>0.621</td>
</tr>
<tr>
<td></td>
<td>(0.072)</td>
<td>(0.067)</td>
<td>(0.053)</td>
<td>(0.07)</td>
<td>(0.077)</td>
</tr>
</tbody>
</table>

Note: the differences between VAS and other methods are statistically significant (\( \alpha = 0.001 \)). The differences between SG, PTO-1 and PTO-2 are not statistically significant. The differences between PTO-3 and the rest are statistically significant (\( \alpha = 0.01 \)). \( N =30. \)

between the expected ordering and the ordering directly obtained at the individual level using Kendall’s \( \tau \).

As we can see the value of \( \tau \) for the VAS is close to 0 (independence of the two orderings), which seems to suggest that people used different criteria to answer both questions. If this were confirmed for larger samples, it would mean that VAS values cannot be interpreted as reflecting people’s preferences for establishing priorities.

We can also see that SG, PTO-1 and PTO-2 are similar in predicting people’s preferences. Using this finding the three methods would be equivalent. Of course, this is so at the ordinal level. What we cannot say using this methodology is what size of intervals better reflects people’s preferences.

We can approach this question using Tables 4 and 5. In Table 4 we have the sizes of the intervals. In Table 5 we have the number of hypothetical votes that each interval received when compared with each of the rest. In rows we have the number of favourable votes that each interval received when compared with the rest of intervals (in columns). For example, 18 individuals ranked the improvement from ‘about to die’ to state 32331 higher than the improvement from state 32331 to 23232 and 24 individuals ranked the improvement from ‘about to die’ to state 32331 higher than the improvement from 12121 to ‘perfect health’. We then have then the number of unfavourable votes that each interval received when compared with the rest in columns. We can clearly see that the treatment that always received most votes was the one which allowed people to avoid death, even though their final health state was not good. The second most voted interval was the improvement from 32331 to 23232. That is, that interval received more votes than the rest when compared in pairs with the others, except when compared with the interval death–32331. The next most preferred interval was the improvement from 23232 to 21312. The next was the interval 21312–12121. The least preferred was the interval 12121–perfect health. It seems that the order of preference is inversely related to the severity of the initial health state. However, it cannot be concluded from this that the less well off were prioritized even though the health gains in that group were less than those that would have been obtained by those who started from a more healthy position. We just know the ranking and not the reasons behind it.

The result of this hypothetical voting exercise can be perfectly explained by the three PTO frames. The sizes of the intervals (Table 4) correspond precisely with the result of the voting. Using the PTO-based health states utilities we would have prioritized the improvement from ‘about to die’ to 32331, because this is the largest interval, then the improvement from 32331–23232 because it is the second largest interval; the improvement from 12121 to ‘perfect health’ comes in the end. This ranking is the same ranking directly elicited with the voting exercise. Using SG creates a problem with intervals 23232–21312 and 21312–12121. The results of the SG seem to suggest that people prefer the second over the former while the result of the votes is the opposite. One explanation is that, although they saw the first interval smaller than the second for themselves, they chose to improve the health of people who were in 23232 due to inequality aversion. That is, they preferred to have two people in 21312 than have one in 23232 and another in 12121, which would have increased inequality. The fact that the SG does not take into account inequality while, in principle, the PTO does, may explain why the SG has the above problem whereas the PTO does not.

In Table 2 we can see that there is no apparent discrepancy between PTO-1 and PTO-3. This result was not expected. It suggests that the diminishing sensitivity effect played no role on this occasion. Clearly more research is needed to
Table 5. Number of votes for each interval

<table>
<thead>
<tr>
<th></th>
<th>Death–32331</th>
<th>32331–23232</th>
<th>23232–21312</th>
<th>21312–12121</th>
<th>12121–perfect health</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death–32331</td>
<td>-</td>
<td>18</td>
<td>18</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>32331–23232</td>
<td>10</td>
<td>-</td>
<td>16</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>23232–21312</td>
<td>10</td>
<td>12</td>
<td>-</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td>21312–12121</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>-</td>
<td>25</td>
</tr>
<tr>
<td>12121–perfect health</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: the number of favourable votes that each interval received when compared with the interval in the column is shown in rows (N = 28; there are two missing).

test if this result could be reproduced in a larger sample. There is a clear discrepancy between PTO-2 and PTO-1, as expected. It is clear that the framing of the question influences the results. To ask in terms of admissible fatalities for every 1000 people cured leads to different PTO values from asking in terms of the number of people cured equivalent to saving 10 lives, although in both cases we are trading off lives against people cured. The bias found is in the direction predicted by the theory, that is, the values obtained with PTO-2 are higher than those obtained using PTO-1. What our results show is that the values we obtain using the PTO may be manipulated when using a different framework.

It can be argued that to avoid manipulation we could standardize the methodology and accept only values obtained using one method. The question is, which method best reflects people’s preferences?

If we answer the last question using predictive power as the criterion, the method that apparently best reflects preferences for establishing priorities is PTO-3. What could be the reason for the superiority of PTO-3 over the rest? It seems that we have to attribute its (apparent) superiority to the only feature that clearly differentiates the method from the rest: it makes people directly compare the intervals. It seems reasonable to expect that the more distant the objects we have to compare the more difficult the matching exercise. We think that this result may be quite important for the PTO method for two reasons:

(i) Although we have shown that different framing may lead to different numbers, this test shows that PTO-3 appears to be better than the others. It may work better because it uses an incremental approach which could be cognitively easier for respondents to understand.

(ii) People may find it easier to accept, from an ethical point of view, making these trade-offs between people when the health states they have to compare are not too different. In our subjects we perceived a better attitude towards PTO judgements with PTO-3 than with PTO-1 and PTO-2. This was particularly important for mild health states. For example, when they were asked how many people in health state 12121 had to be cured in order to produce the same benefit as saving 10 lives, 35% of the sample gave an answer like ‘I don’t know, it had to be a lot, say a million people’. This answer just reflects the difficulty of making the judgement. With PTO-3 we only had this problem in a couple of subjects. However people seem to accept very well the trade-off between lives and the complete recovery of people who were in health state 32331. A severe health state 32331 is severe, they were much less reluctant to accept that a trade-off had to be done. The reason is that they realised that the benefit gained by accepting the death of somebody would also be very important. So if we want to know how many headaches are equivalent to saving one life we had better not ask that question directly because people will probably not be able to give reasonable answers.

One conclusion of all this is that in order to choose a method that reflects people’s preferences, we would have to take into account how well the method adapts to subjects’ possibilities of processing information and giving answers to our questions. ‘If subjects cannot use the response mode most convenient to investigators, then investigators must find a response mode that works for subjects.’

The difficulty that people have in expressing their preferences for very mild health states may also explain why the discrepancy between the intervals obtained using the SG and the PTO is larger for mild health states than for more severe health states. For example, if we look at the intervals (Table 4) we see that the largest discrep-
The main objective of this paper is to test the predictive power of several methods proposed in the literature to obtain health states utilities. We think that the main conclusions we can draw from this study are the following:

1. At the aggregate level the PTO performed slightly better than the SG.

2. At the aggregate level the PTO (and the SG) performed much better than the VAS. In this case we have strong evidence against the VAS. The low predictive power of VAS is a matter of concern. We have to take into account that the method used in Oregon to value health states was very similar to the VAS. In fact, in a telephone survey subjects were described the health state and were asked for a value between 0 (death) an 100 (no health problems). This is in fact very similar to the VAS, with the added problem that the interview was not personal and subjects did not have much time to think. As the results were used to compare the benefits that different health policies would provide for different groups of people, it is not a surprise that the results were highly controversial.

3. PTO-3, based on a direct comparison of the intervals, showed a higher predictive power at the individual level. We have suggested some reasons why this frame of the PTO may perform better than the rest. What the highest predictive power of PTO-3 suggests is that an improvement from 

\[
B \rightarrow C 
\]

suggests that an improvement from B to C may be valued differently if we value directly this improvement or if it is calculated as the difference between a reference state A and the two health states B and C [equation (5)].

A part from the specific results obtained, what this paper tries to emphasize is the importance of checking if the decisions we recommend should be taken, using cost-utility analysis, agree with the decisions people would like to see implemented. When we estimate health state utilities and use them to establish priorities, we are using these numbers to compare the benefit of different groups of people. Are we entitled to interpret numbers in this way? We have suggested one way of approaching this question. This approach is not new at all. Some years ago Loomes and McKenzie wrote17 ‘there should be attempts to compare the choices between various alternatives implied by the stated valuations and QALY scores derived, with decisions made in direct choices between those same alternatives’. We have tried to follow this approach.

ACKNOWLEDGEMENTS

I thank A lan Williams, Erik Nord and Xavier Badía for their helpful comments. Special thanks are due to Han Bleichrodt, who patiently helped me to understand my own ideas better. The usual disclaimer applies. Financial support was received from Fondo de Investigaciones Sanitarias, project 0767/95, and from Merck, Sharp and Dohme.
APPENDIX 1: FRAMES USED IN PERSON TRADE-OFF QUESTIONS

PTO-1. Imagine that you are a health planner and you have to decide among two competing health care programmes. One of them consists in buying more ambulances. Health authorities have estimated that if they buy more ambulances 10 more lives would be saved each year. The other programme consists in funding a new medicine that will completely cure a group of people with a health state like this one (Appendix 2). There is no treatment for this health problem at the moment and if you do not fund the medicine they will probably remain in this health state for the rest of their lives. How many people would have to be cured so that you are indifferent between these two programmes?

PTO-2. Imagine that you are a health planner and you have two groups of 1000 people with different health problems. The members of one of the groups have an illness that you think they can overcome in a short amount of time without giving them medication. However, if you do not give them medication there is a probability that they will die. If you give them medication it is sure that they will recover. The members of the other group are in health state X (Appendix 2). If you give them medication it is certain that they will recover. If you do not treat them, they will be in that situation for the rest of their lives (health state X for the rest of their lives). What is the risk of death you would accept for the first group which would make you indifferent between treating any of these two groups? Take into account that if you accept an individual risk of death for each patient of 1/1000 this means that you accept that one patient will die in order to cure the 1000 patients in the other group. What is then the maximum number of fatalities that you would accept in order to cure the 1000 patients in the other group?

PTO-3: (The first question is the same as for PTO-1.) Now you have two groups of patients, one in health state X and the other in health state Y (milder). The origin of the problem is different in both cases. Two medicines have recently been discovered that would completely cure both groups. However, you can fund only one. With the budget that you have you can cure 10 people who are in health state X. How many many would have to be cured of illness Y so that you are indifferent between these two programmes? We repeated this question starting from Y and comparing this program with funding another medicine for a group of people who were in health state Z (better than Y).

Note Although these are the frames that we initially used, some people needed further explanation. We then talked to them trying to explain the example better, showing them examples about the use of probabilities, etc. In general, for all the interviewees we not only showed them the frames, we also asked them if they thought they had understood, and we asked them some questions to check they had understood; and if we saw they had not, then we tried to explain the same questions with different words.

APPENDIX 2: EUROQOL HEALTH STATES USED IN THE STUDY

12121
No problems with walking about.
Some problems washing or dressing self.
No problems with performing usual activities (e.g. work, study, housework, family or leisure activities).

21312
Some problems with walking about.
No problems with self-care.
Unable to perform usual activities.
No pain or discomfort.

23232
Some problems with walking about.
Unable to wash or dress self.
Some problems with performing usual activities.
Extreme pain or discomfort.

32331
Confined to bed.
Some problems washing or dressing self.
Unable to perform usual activities.
Extreme pain or discomfort.
Not anxious or depressed.

© 1997 by John Wiley & Sons, Ltd.
REFERENCES