CENTRE FOR HEALTH PROGRAM EVALUATION

WORKING PAPER 8

Methods for Establishing Quality Weights for Life Years

Erik Nord
National Institute of Public Helath, Oslo

Re-print of National Centre Working Paper No 8

First edition printed 1991
ISSN 1038-9547
ISBN 1 875677 15 1

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of Public Health and Community Medicine at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE receives core funding from the National Health and Medical Research Council and Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR(S) ACKNOWLEDGMENTS

In the preparation of this paper I have had a number of most stimulating discussions with Jeff Richardson. His comments have been exceedingly constructive and encouraging. I am also indebted to Elizabeth Nygaard, Johanna Cook, Richard Allison, Carole Butler, Martin Buxton and Caroline Selai for valuable comments to earlier versions of the paper.

E. Nord

ABSTRACT

Several valuation techniques are in use for quality adjusting life years in cost utility analysis. The paper gives an overview of the variability in results. A close inspection of a number of instruments with respect to their theme, instructions, decision framing and the phrasing of questions make many of the observed differences in results understandable. When judging the validity of the different techniques, three points should be kept in mind. One is that statements about validity should be made with respect to concrete versions rather than broad categories like "the rating scale", "time trade-off" etc. Another point is that a valuation technique that is valid in clinical decision analysis may not be valid in health program evaluation, and vice versa. The third point is that quality weights for life years are empirically more meaningful, in the sense that they are more amenable to empirical testing, if they are interpreted simply as preference weights rather than measures of amounts of well life in the utilitarian tradition. Time trade-off with a moderate time horizon is recommended in clinical decision analysis, while a combination of time trade-off and a variant of person trade-off is recommended in health program evaluation.

Methods For Establishing Quality Weights For Life Years

1 Introduction

In cost-utility analysis, the benefit of a health service is stated in terms of the number of quality adjusted life years gained by the service (see for instance Torrance, 1986, for a review). Quality adjustment consists of having people assign weights to life years in different health states on a scale from zero (dead) to unity (healthy). The general rule is to let people value health states that they are not in themselves. Several valuation techniques are in use, including the rating scale, magnitude estimation, standard gamble, time trade-off and person trade-off (see appendix).

The various techniques have been compared in several studies, and considerable differences in results have been observed. An overview of such differences is given in the next section.

In most comparative studies relatively little effort has been devoted to finding out why results differ. The reason for this may be that much of the work in the field has been carried out by doctors and economists rather than by psychometricians. Whatever the reason, several writers have pointed out that we need a better understanding of differences in results before applying the weights in concrete decision making (Rosser, 1983; Sutherland et al, 1983; Llewellyn-Thomas et al, 1984; Read et al, 1984; Loomes and McKenzie, 1989).

It is the belief of this author that considerable understanding can be obtained merely by looking more closely at differences between valuation instruments with respect to what is being asked and how it is asked. This means a detailed inspection of theme, instructions, decision framing and phrasing of questions in each instrument. Section three of this paper offers a number of explanations of observed differences in results based on such "instrument analysis".

However, to know why results differ is not enough. When different instruments produce different results, the normative question still remains: Which instruments are more valid, and which are less? As we shall see below, there is much hesitancy and disagreement about this question in the literature. It is the belief of this author that much of the disagreement can be avoided if three points are recognised and kept in mind. One is that valuation techniques occur in different versions that produce quite different results. Statements about validity should therefore be made with respect to concrete versions rather than broad categories like "the rating scale", "time trade-off" etc. Another point is that the validity of a valuation technique depends on the use to which the valuations are put. In particular, a valuation technique that is highly valid in clinical decision

analysis may be less valid in health program evaluation, and vice versa.

The third point is that quality weights for life years are empirically more meaningful, in the sense that they are more amenable to empirical testing, if they are interpreted simply as preference weights rather than measures of amounts of well life in a life year in the utilitarian tradition. These three points are elaborated in section four. In section five some policy recommendations are made on the basis of the previous analysis.

Cost-utility analysis in terms of QALYs gained raises a number of other difficult questions. Among these are: Whose values should count? Can we really depend on the judgements of healthy people for assessment of the relative disutility of different states of illness and disability? Can life years be valued individually and put together in a simple additive model, or should the object of the valuation task rather be life scenarios? Should patient characteristics like age or family situation count? Is it sufficient to determine weights within a general health state classification system, or are disease-specific weights necessary in order to achieve satisfactory precision?

These are all important issues that need careful consideration before any overall judgement on the feasibility and validity of QALYs can be passed (see for instance Donaldson et al, 1988; Loomes and McKenzie, 1989; Mehrez and Gafni, 1989; Nord, 1989). They are, however, beyond the scope of this particular paper.

2 Differences in results

Selected representative results from a number of studies are presented in tables 1 and 2. We can make note of the following:

Standard gamble and time trade off generally give higher values than the rating scale. The study by Richardson et al (1989), however, is an exception.

Time trade-off lies substantially below standard gamble in two investigations, and at the same level as standard gamble in two others.

Magnitude estimation gives very divergent results - from very high values (Rosser & Kind, 1978; Buxton et al, 1987), via values as for the rating scale (Patrick et al, 1973) to values far below the rating scale (Kaplan et al, 1976).

The person trade-off technique also gives divergent results: about the same values as the rating scale in one study (Patrick et al, 1973), but much higher in another (Nord, 1991).

3 Reasons for the variation

It is not surprising that the results are heterogeneous. In each technique the subjects are faced with a cognitive task that differs from that used with other techniques. In addition, several of the techniques exist in different versions that frame the decisions in different ways. In the following we will point out a number of possible explanations of the observed differences in results. Some of these - relating to general differences between the five main techniques - have been offered earlier (see for instance Tversky and Kahneman, 1981; Schoemaker, 1982; Kaplan and Ernst, 1983; Llewellyn-Thomas et al, 1984; Froberg and Kane, 1989; Loomes and McKenzie, 1989; Richardson, 1989), while others - relating to particular versions of the various techniques - seem to have been somewhat overlooked in the literature.

Differences in the level of abstraction

There is a crucial distinction between standard gamble, time trade-off and person trade-off on the one hand, and the rating scale and magnitude estimation on the other. The former may be called equivalence techniques or trade-off techniques. They face the subjects with a choice between pairs of conditions. The question is: How much are you willing to sacrifice of certainty (SG), life span (TTO) and the health of others (PTO), respectively, in order to improve your own quality of life (SG and TTO) or that of an imaginary patient (PTO). With the rating scale and magnitude estimation, on the other hand, subjects are asked to apply numerical scales directly to clinical conditions. But few people - if any - use numerical scales in everyday situations when thinking of or expressing quality of life (Mulkay et al, 1987; Nord, 1990). A priori, there seems to be no reason why answers to such abstract questions should correspond closely with answers to concrete questions about trade-offs (Bombardier 1982; Carr-Hill 1988; Loomes 1988; Richardson 1989; Nord 1990).

Differences with respect to what is being valued

In a computation of QALYs, a weight for a life year in a particular health state is supposed to reflect the goodness or badness of the state as perceived by the individual concerned. But most valuation techniques capture more than or something different from such a pure quality of life consideration, and they do so in different ways.

When standard gamble is used, risk aversion in general and reluctance to "gambling with own health" in particular may make subjects demand very low probability of death in order that they should prefer the gamble alternative rather than to continue (with certainty) in a given state of illness (see for instance Bombardier, 1982). This points in the direction of relatively high weights on states of illness when the standard gamble method is used.

On the other hand, people are generally less concerned with losses in the distant future than with losses in the near future (so-called "positive time preference"). It seems reasonable to assume that this be true also of losses of life years (see for instance Lipscomb, 1989). In a time trade-off exercise, subjects will therefore probably express higher willingness to trade in life years in order to win life quality the longer the time horizon. The weights for health states that this method produces may therefore have a tendency to lie lower than what a pure valuation of life quality might indicate.

Taken together, risk aversion and positive time preference may provide a reasonable explanation of the differences observed between standard gamble and time trade-off in the studies of Bombardier et al.(1982) and of Read et al. (1984). Positive time preference may also be the explanation of why these two studies provided lower values for time trade-off in relation to standard gamble than a study by Torrance (1976) did. Bombardier et al.(1982) used remaining

life expectancy as a time horizon. Read et al. used a 10 year horizon. Torrance, on the other hand, assumed a remaining life expectancy of only 3 months for the health states in question. In this way the subjects had minimal room for possible time preference.

More generally, it is well known that information about expected duration of a state has an effect on its valuation (Sackett and Torrance, 1978; Sutherland, 1982). There is great variability with respect to such information in the rating scale exercises referred to in Tables 1 and 2. Specified duration ranges from indefinite (Llewellyn-Thomas et al, 1984; Sintonen, 1981; Richardson et al, 1989) to five and ten years (Read et al, 1984), one year (Nord, 1991), three months (Torrance, 1976) to one day (Patrick et al, 1973; Kaplan et al, 1976). For this reason alone, there is no reason to expect that these studies should produce the same results.

An example is the choice of a one-day context made by Kaplan et al for the Quality of Well-Being Scale. The effect is for instance that a person whose sole health problem is a stuffy, runny nose receives a score of 0.83, while a person who in addition is in bed performing no major role activity receives a score of 0.61 (Kaplan and Anderson, 1990). In terms of undesirability of a single day, it is quite understandable that the difference between the two scores is not greater. After all, a stuffy, runny nose is always unpleasant, while a day in bed need not be so bad. But in the long run one would expect the difference in undesirability between the two states to be much more substantial. Thus, we may note that "bedridden" scored as low as 0.09 in Sintonen's study (table 2), where duration was unspecified.

When person trade-off is used, distributive considerations become a serious confounding factor. Subjects may consider that all or most of a budget should not be spent on any one patient, regardless of how much worse off the individual is than others. Such considerations will tend to limit differences in weights between serious and less serious conditions. It will work in the opposite direction if the subjects wish to distribute health as uniformly as possible and consequently think that those who are worst off must have priority regardless of the relative quality of life. The net effect of such distributive considerations is difficult to evaluate and will vary with social values

As noted above, results obtained with different versions of the person trade-off method differ considerably. Again, this probably has to do with differences in what is being valued. Patrick et al. (1973) asked the subjects to value saving the lives of a number of people in different states of illness, in relation to saving the lives of a number of healthy people. Nord (1991b), on the other hand, asked the subjects to value saving the life of one person in relation to curing a group of seriously ill patients. Clearly, these two problems are not identical. Knowing how sensitive people's responses in general are to how questions are framed (see for instance Tversky and Kahneman, 1981; Schoemaker, 1982), it is not surprising that the answers to the two questions are different.

The person trade-off version used by Nord (1991b) produced substantially higher values than both the standard gamble and time trade-off have done for comparable states. The reason for this may be that the standard gamble and time trade-off involve sacrificing own life, while person trade-off has to do with sacrificing other peoples lives. Subjects may have different values with respect to these issues. They may for instance feel that they have a right to do as they wish with their own life, including sacrificing life years or certainty of survival in order to gain quality of life. At the same time they may feel an obligation to give very high priority to life saving - as opposed to health improvement - when other people's lives are concerned. People with such a dual attitude will (implicitly) assign relatively high values to life per se when responding to a person trade-off question. At the same time they may assign lower values to the same states when they only have to take their own lives into consideration, as is the case with standard gamble and time trade-off.

Buxton et al (1987) and Richardson et al (1989) have pointed out that the magnitude estimation technique employed by Rosser and Kind (1978) is insensitive to psycho-social morbidity. Again, this may have to do with the focus of the valuation instrument. The main principle of direct

scaling techniques is that the subjects are asked to score clinical conditions according to how good or desirable they are, in other words, according to quality of life. However, Rosser and Kind (1978) are an exception here. They asked the subjects how many times more ill one is in condition A than in condition B. This is a more limited theme. There may be clinical conditions that many people regard as undesirable without necessarily thinking of them as states of illness. Anxiety and depression are probably examples of this. These conditions may therefore score higher on a healthy-ill-scale than they would on a quality of life-scale.

Differences in the use of anchoring points

Everything is relative. If a respondent is asked to assess a health state in relation to being well, one must expect the health state to score badly. But if the respondent is asked to compare the same state with being dead, there is reason to expect a rather more positive assessment (see for instance Schoemaker, 1982; Sutherland et al, 1983).

This so called anchoring effect probably explains why use of magnitude estimation in the studies of Rosser and Kind (1978) and Kaplan et al. (1979) gave such different results. As noted above, Rosser and Kind asked the subjects to indicate how much more ill patients were in different states of disability and distress than in a mild reference state. They then asked the subjects to score the condition "dead" in the same way. The mean response for dead was "200 times more ill". This ratio was much higher than those obtained for most illness. The effect of this was to compress the score for most of the other conditions into the top of the standardised scale from 1 (well) to 0 (dead).

Kaplan et al.(1979), on the other hand, did not include death in their measurement. This was because subjects were asked to score one day in different states. A reference day was chosen on which the patient had symptoms such as a "stomach upset". One healthy day was compared with this reference day. The subjects thought on average that the healthy day was 9-10 times as desirable as the reference day. This led to "stomach upset" and conditions of this order of severity being compressed into the lower end of the 1-0 scale.

Another example of the effect of anchoring points may be found in connection with Kaplan et al.'s use of a rating scale. The scale extended from 10 (well) to 0 (dead). In itself the use of "dead" as the lower end point explains why the values in general were much higher than those obtained in their magnitude estimation exercise. However, the instructions also included the following: "If you think the person's situation was about half-way between being dead and being completely well, then choose step 5". This is scarcely an unimportant detail. "Half-way between being dead and completely well" may have sounded like a very serious condition to many subjects, (note the resemblance to "half dead"). To the extent that step 5 was associated with such a serious condition, we may suspect that the values of many other severe states have been compressed in the upper half of the scale.

Neither Bombardier et al. (1982), Patrick et al. (1973) nor Sintonen (1981) anchored the mid-point of the visual analogue scale in this way. This may be another reason (in addition to differences in specified duration as noted above) why they obtained lower values for states such as "sitting in a wheel chair" or "being confined to bed" than Kaplan et al. did (see tables 1 and 2).

When Rosser/Kind and Kaplan et al. (1979) used magnitude estimation, there was no constraint upon the movement along the axis from "dead" to "well". That is to say, the subjects could use as high a ratio as they wanted, to compare health states. This made it possible for the subjects to express very large differences in value between a given reference condition ("dead" in the study by Rosser/Kind, "upset stomach" in that of Kaplan et al, 1979.) and the remaining conditions. Rating scales do not provide the same possibility. The scale has fixed upper and lower end points, and the subjects have to assess each state in relation to these two end points simultaneously. With the majority of conditions the subjects will probably want to express a certain distance from both end points. This may explain the more extensive use of the middle part of the scale when the rating scale is used, compared with results found by Rosser/Kind

(1978) and by Kaplan et al (1979).

The fact that studies by Patrick et al. (1973) and Sintonen (1981) obtain values between Rosser and Kind and Kaplan et al. fits with this explanation. In all four studies ratio questions were asked, but unlike the latter two, the two former studies used a scale with fixed upper and lower end points (1000-0 and 100-0 respectively). As noted by others, their scales thus acquired a strong resemblance to rating scales (Kaplan et al. 1979; Sintonen 1981).

Nord (1991, 1991b) found far higher scores when using the person trade-off technique than when using the EuroQol(c) rating scale. Different use of anchoring points is a fairly obvious explanation. The EuroQol(c) scale is numbered from zero to one hundred, the end points being labelled "worst imaginable health state", and "best imaginable health state" respectively. A follow up questionnaire revealed that subjects had a tendency to interpret the numbers on this particular scale in terms of "percentages of fitness", which indicates that they were primarily using the upper end as a single anchor point. This would tend to compress health states of illness down on the scale. In the person trade-off questions, "dead" was implicitly used as the reference condition, and there were no limits to how high values the remaining conditions could be given. This drew states of illness up when transferred to a 1-0 scale, as was the case in the study of Rosser and Kind (1978).

4 Which techniques are the better ones?

When different instruments produce different life year weights the inevitable question is: Which techniques are the better ones?

In international review articles there is some hesitancy about answering this question (Rosser 1983; Torrance 1986). Williams (1988) writes expressly that "the valuation part can be handled by a variety of methods ... No one of these -... is clearly superior or inferior to the others". Loomes and McKenzie (1989) call for "comprehensive and detailed comparisons between the main existing methods..to see..if a better understanding of the relationships between methods can be obtained".

Others have been willing to go further, although in different directions. Kaplan et al. (1979) favour the rating scale. Torrance and Feeny (1989) recommend weights based on standard gamble. Bombardier et al. (1982), Richardson (1989) and Mooney and Olsen (1990) advocate the use of time trade-off.

I see three reasons for this lack of clarity and disagreement in the scientific community.

One is that techniques are often discussed in a general manner without taking into consideration that most of them occur in different versions that produce quite different results. This is really the main message from the previous section. Overlooking this intra-technique variability easily leads to statements that are too bold.

An example is the general conclusion drawn by Kaplan et al (1979) that "magnitude estimation does not appear appropriate as a measurement method for a health status index .." (p.520). A premise for this statement is the dissimilarity of the magnitude estimation results not only to rating scale results, but also to results obtained in the person trade-off study of Patrick et al (1973). However, the magnitude estimation results stem from one particular exercise where the state "dead" was omitted. There is really no basis for transferring the results of this exercise to a 1-0 scale where "dead" is the bottom end point. In the magnitude estimation study of Rosser and Kind (1978), valuation of "dead" was a central part of the task. The results were not only dramatically different. They were very similar to those obtained in the person trade-off study conducted by Nord (1991, 1991b).

Another factor that contributes to confusion and disagreement seems to be insufficient recognition of the fact that the validity of any valuation technique will necessarily depend on the use to which the values are put. Two different areas of application are particularly worth mentioning. In clinical decision analysis, the health state values are used for assessing benefits related to alternative treatments. This application requires a valuation technique that focuses on patients' values regarding their own life. In program evaluation, on the other hand, health state values are used for comparing outcomes of programs for different groups of patients. This application requires a technique that captures people's values regarding the lives of other people. As noted in the previous section, the two kinds of values need not be the same. Obviously, then, valuation techniques that are highly valid in the first application may be less valid in the second, and vice versa.

A third reason for disagreement regarding validity is an ambiguity in the interpretation of quality weights for life years.

In one interpretation, the weights are supposed to express - at a cardinal level - how much well life years in different states contain (see for instance Torrance, 1976b). The amount of well life in a year of perfect health is used as numeraire. We may call this the utilitarian interpretation.

In another interpretation, life year weights are supposed to express, at a cardinal level, the trade-offs that subjects are prepared to make between different health care outcomes. We may call this the preference interpretation (see for instance Patrick et al, 1973).

The two interpretations are related, in the sense that subjects' opinions in terms of trade-offs certainly reflect their feelings about the goodness of life in different states (see for instance Kaplan, 1979). But there is a clear difference between the two interpretations in terms of testability and meaning.

With the utilitarian interpretation, testing the validity of a valuation technique amounts to examining if it counts well life correctly. One way of doing this is to compare weights obtained with the technique in question with weights obtained by means of a technique for counting well life that is generally accepted as valid. This would be a test of criterion validity. However, a criterion technique, often referred to as a "gold standard", is not easy to find for "well life".

The reason for this is that very few people, if any, use numerical scales in everyday situations when thinking of or expressing quality of life. "The amount of well life" is therefore for most people a concept without meaning at a cardinal level (Mulkay et al, 1987; Klein, 1989; Nord, 1989).

To make this point clear, let me take an example. Let us assume that two health states A and B are assigned values 0.4 and 0.8 respectively. Given the utilitarian interpretation, the following statement may then be made: "Curing a person in state A for ten years leaves society with an increase in well life that is three times as high as curing a person in state B for ten years." It is the belief of this author that to most people this statement is mystifying rather than meaningful and that there will never be agreement in society as to how the statement could be verified or falsified.

Torrance (1987) points out that if the basic assumptions of expected utility theory are accepted, the standard gamble is a valid technique for determining the amounts of well life associated with different health states. However, basing a validity test on theoretical assumptions is really to avoid the problem. The assumptions must have empirical support. In fact, substantial evidence indicates that people in practice often behave in contradiction to the assumptions of the expected utility model (see for instance Schoemaker, 1982; Sutherland, 1982; Richardson, 1989). It is quite difficult to see how standard gamble can be regarded as the criterion method for counting well life.

Within the framework of the utilitarian interpretation, an alternative approach to the question of validity suggested by Torrance (1987) would be to see how strongly health state valuations correlate with other trusted measures of health related quality of life. This would be a test of construct validity (see for instance Kaplan, 1976). Churchill et al. (1987) conducted such a test. They found that patients' time trade-off values had a correlation of 0.40 with nephrologists' scores of the same patients by means of Spitzer's Quality of Life Index. The result is interesting, but far from convincing in our context. Firstly, a correlation of 0.40 is not very high. Secondly, as noted by Read et al. (1984), correlation analysis is a poor way of revealing systematic differences between methods.

As long as the methods rank states similarly, the correlation may be high even if one of the methods systematically produces lower values than the other. Thirdly, Spitzer's Quality of Life Index is not at a cardinal level. Hence, even if there were complete concordance between the two sets of scores, one could not draw the conclusion that the time trade-off technique yields valid cardinal estimates of well life.

Altogether, the interpretation of quality weights as "amounts of well life" in life years in different states does not seem to be a viable one from a scientific point of view, as there seems to be no convincing way of testing - or falsifying - statements based on this interpretation.

As noted above, an alternative interpretation of quality weights is in terms of strength of preference. Within the framework of this interpretation, the question of validity of a valuation instrument turns into a question of whether the instrument captures the subjects' preferences in a correct way.

A criterion test of validity may in this case consist of testing for so called reflective equilibrium (see Rawls, 1971), i.e. examining to what extent preference statements that are inferred from health state valuations are in accordance with preferences elicited directly. Without using the technical term, Loomes and McKenzie (1989) have called for precisely this kind of test. An example will demonstrate how it would work. Let us assume that a subject by means of some instrument has assigned the value 0.4 to a state A and the value 0.8 to a state B. If these values are interpreted as preference weights for life years, statements of the following two kinds may be inferred:

> (S_1) The subject is indifferent between living 2 years in state A and living 1 year in state B (2x0.4=1x0.8).
>
> (S_2) The subject is indifferent between making 1 patient in state A well for 2 years and making 6 patients in state B well for 1 year (2x1x0.6 = 1x6x0.2).

The valuation instrument is validated with respect to clinical decision analysis to the extent that subjects, for a variety of states, agree with inferred statements of type S_1. Similarly, the valuation technique is validated with respect to program evaluation to the extent that subjects, for a variety of states, agree with inferred statements of type S_2.

Essentially, this is the same as saying that the proof of the pudding lies in the eating. Unfortunately, there are few instances of such testing in the science of valuing health states. An exception is Rosser and Kind (1978), who pointed out to subjects how their initial responses in terms of magnitude estimation would be interpreted in terms of preferences for resource allocation on individual patients as well as preferences for programs involving different numbers of people. The subjects were encouraged to modify their initial responses if they were uncomfortable with these interpretations. A reflective equilibrium in terms of program evaluation was in other words striven for. Accordingly, Rosser/Kind health state values have considerable criterion validity for such evaluation. But it must be noted that this is due to elements in the evaluation process that effectively turned their valuation technique into something far more than pure magnitude estimation.

Another study which comes close to testing for reflective equilibrium is Nord (1991). Here, preferences inferred from health state valuations on the EuroQol(c) rating scale were compared with preferences elicited directly by means of the person trade-off technique. There were large discrepancies between the two sets of preferences, which indicates that valuations obtained by means of the particular rating scale in question have low validity as quality weights for life years in program evaluation.

In the absence of further empirical studies focusing on reflective equilibrium, one may have recourse to a second-best way of validating valuation instruments. One may

examine to what extent statements put forward on the basis of the valuations are clearly embedded in or follow logically from what the subjects have actually expressed in their responses to the valuation task. In doing so, we may distinguish between two levels. Firstly, we may study the immediate similarity between the subjects' responses and the inferred preference statements in terms of concepts involved. This is much like a test of face validity. Secondly, for valuation instruments with low face validity, we may make explicit the assumptions that need to be made in order to allow inference from the subjects' responses to the derived statements. The bolder the assumptions, the lower is the content validity of the instruments with respect to establishing the derived statement.

Consider again the statements S_1 and S_2 above and the two underlying valuations (A=0.4 and B=0.8). For each of the five main valuation techniques Table 3 lists the responses that would yield these two valuations. Each response may be compared with each of statements S_1 and S_2 with respect to immediate similarity in terms of concepts involved.

Statements S_1 and S_2 are essentially expressions about either self or others. They express trade offs in numerical terms between level of well-being, duration of state and (for S_2) the number of people concerned. A tentative conclusion based on examination of the various responses listed in Table 3 would then be that time trade-off responses resemble very much preference statements concerning clinical decisions (type S_1). Time trade-off and person trade-off responses both have much in common with preference statements concerning program evaluation (type S_2). Standard gamble and person trade-off responses have some similarity with preference statements concerning clinical decisions, in as much as a trade off is expressed, but the object of trade is certainty (SG) and number of people (PTO) rather than duration of state. The rating scale and magnitude estimation responses have conceptually relatively little in common with either of the inferred preference statements. Table 4 summarises these conclusions in terms of face validity of the various valuation techniques.

Valuation techniques with low face validity may still have high content validity if it can be safely assumed that subjects actually think in terms of trade offs between level of well-being, duration of state and numbers of people concerned when responding.

For pure versions of the rating scale, magnitude estimation and standard gamble there is little a priori reason to expect thinking in such terms. Occurrence of such thinking would therefore have to be empirically demonstrated before any of these techniques could be accepted as valid in terms of content validity. Unfortunately, there has been little research on how people think when they value health states. The evidence that exists is not encouraging. Nord (1991) let a group of subjects value a set of health states on a rating scale numbered in fives from 0 to 100 and then asked them what they meant by the numbers they had selected.

None of 67 subjects made any reference to life year weights, trade offs or equivalence in numbers of treatments. Eleven subjects explicitly stated that they did not mean anything in particular by the numbers or that the numbers were randomly chosen.

Nord's study has later been replicated by Morris and Durand (1989), with corresponding results, especially as regards the lack of depth of intention of the subjects when choosing numerical values.

5 Conclusion

Different instruments for valuing health states produce different results. A close inspection of a number of instruments with respect to theme, instructions, decision framing and phrasing of questions makes many of the observed differences in results understandable. In this paper we have particularly focused on such explanatory factors as differences in levels of abstraction, differences in what is being valued and differences in the use of anchoring points.

One main lesson to be learnt from the data is that there may be significant differences both in design and results between instruments that are alike in terms of basic approach. Generalisations with respect to such broad categories as "the rating scale", "magnitude estimation", "standard gamble", "time trade-off" and "person trade-off" should therefore be made with great care.

Another important lesson is that one single factor, namely differences in the treatment of the state "dead", seems to explain a large part of the observed variance in health state valuations across instruments. The stronger the presence of this state as a reference state, the higher the values of all states of illness.

When different instruments produce different values, the question arises: which values are the correct ones?

In answering this question, we have distinguished between two applications of health state valuations. One concerns clinical decision analysis, the other program evaluation.

In both cases, the purpose of valuing health states is to establish quality weights for life years in the various states. We have discarded the interpretation of such weights as "amounts of well life", as there seems to be no convincing way of testing - or falsifying - statements based on this interpretation. Instead we have focused on quality weights in terms of strength of preference for different outcomes. Given this interpretation, the best test of validity is to examine to what extent subjects agree with preference statements that are inferred from the subjects' responses to the valuation task. In one study, the rating scale scored poorly on such a validity test. For other techniques, such empirical studies are lacking and should be strongly encouraged.

A second best solution to the validation problem is to examine to what extent QALY-type statements put forward on the basis of health state valuations seem to be directly embedded in or follow logically from what the subjects have actually expressed in their responses to the valuation task. Using this face validity criterion, rating scales and pure magnitude estimation score very poorly. Time trade-off, preferably with a short time horizon, would then seem to be the most valid technique for establishing preference weights for life years in clinical decision analysis. In program evaluation, a procedure that seems worth consideration would be to combine values elicited by means of time trade-off and person trade-off. In the absence of more precise knowledge, a pragmatic solution may be to average the two.

The rating scale technique is generally accepted as the easiest for subjects to understand. It may well be that responses to such scales have stable mathematical relationships to preference weights established through direct trade-off methods. If this be the case, transformed rating scale values may be used as preference weights for life years. Studies by Torrance (1976) and Loomes (1988) suggest that such mathematical

relationships do exist. Further research is needed, however, to ascertain the strength and form of these relationships. As noted by Richardson et al (1989), such research will have to be specific to concrete versions of the rating scale.

It must be kept in mind that tests of face or content validity are second best solutions. I would like once more to emphasise that the real proof of the pudding lies in the eating. It may therefore well be that an instrument with low face and content validity through interaction of different design factors randomly produces values that perform well on a test of reflective equilibrium. If such goodness of fit is empirically demonstrated, the instrument should be regarded as highly valid for eliciting health state values however meaningless or biased the instrument in itself may seem. The magnitude estimation procedure adopted by Rosser and Kind (1978) is possibly an example of this in program evaluation, confer the likeness of the results to the results of the person trade-off study of Nord (1991, 1991b).

As noted in the introduction, cost-utility analysis in terms of QALYs gained raises a number of difficult questions that lie beyond the scope of this paper. One of these questions does, however, deserve a brief comment. Health state valuation studies are generally based on proxies. That is, subjects are asked to consider states that they are not in themselves. This is true of all the studies considered in this paper. Applying the results of any of these studies in decision making clearly presupposes that judgements by proxies correctly reflect the relative disutility of different states of illness or disability as perceived by patients in those states.

There is some evidence that proxies score states lower than patients do (e.g., Epstein et al, 1989). But in general, far too little is known about this issue. It is felt by this author that asking patients rather than proxies should be given high priority in future research on health state valuations.

REFERENCES

Bombardier C. et al. Comparison of three preference measurement methodologies in the evaluation of a functional status index. In: Deber RB, Thompson GG, (ed). Choices in health care. Department of Health Administration, University of Toronto, 1982.

Buxton M, Ashby J, O'Hanlon M. Alternative methods of valuing health states. Mimeo. Health Economics Research Group, Brunel University, 1987.

Carr-Hill RA. Assumptions of the QALY procedure. Social Science & Medicine 1989; 29:469-477.

Churchill D. et al. Measurement of quality of life in end-stage renal disease: The time trade-off approach. Clinical and Investigative Medicine 1987; 10:14-20.

Donaldson C, Atkinson A, Bond J. Should QALYs be programme-specific? Journal of Health Economics 1988; 7: 239-257.

Epstein AM, Hall JA, Tognetti J, Son LH, Conant L. Using proxies to evaluate quality of life. Medical Care 1989; 27:S91-98.

Froberg, DG, Kane L. Methodology for measuring health-state preferences - II: Scaling methods. Journal of Clinical Epidemiology & 1989; 42:459-471

Kaplan RM, Anderson JP. The general health policy model: An integrated approach. In: Spilker B (ed.): Quality of life assessments in clinical trials. Raven Press, Ltd. New York 1990.

Kaplan RM, Bush JW, Berry CC. Health status: Types of validity and the index of well-being. Health Services Research 1976; 11:478-506.

Kaplan RM, Bush JW, Berry CC. Health status index. Category rating versus magnitude estimation for measuring levels of well-being. Medical Care 1979; 17:501-25.

Kaplan RM, Ernst JA. Do category rating scales produce biased preference weights for a health index? Medical Care 1983; 21:193-207.

Klein R. The role of health economics. British Medical Journal 1989; 299:275-276.

Lipscomb J. Time preference for health in cost-effectiveness analysis. Medical Care 1989; 27 S:233-253.

Llewellyn-Thomas H. et al. Describing health states. Methodological issues in obtaining values for health states. Medical Care 1984; 22:543-52.

Loomes G. Disparities between health state measures: An explanation and some implications. Mimeo, Department of Economics and Related studies, University of York, 1988.

Loomes G, McKenzie L. The use of QALYs in health care decision making. Social Science & Medicine 1989; 28:299-308.

Mehrez A, Gafni A. Quality-adjusted life years, utility theory and healthy year equivalents. Medical Decision Making 1989; 9:142-149.

Mooney G, Olsen JA. QALYs: Where next. Forthcoming.

Morris J, Durand MA. Category rating methods: Numerical and verbal scales.
Mimeo. Centre for Health Economics, University of York, 1989.

Mulkay M, Ashmore M, Pinch T. Measuring the quality of life. Sociology 1987; 21:541-564.

Nord E. The significance of contextual factors in valuing health states. Health Policy 1989; 13:189-198.

Nord 1990. A comment on the meaning of the numerical valuations of health states. Social Science & Medicine 1990; 30:943-944.

Nord E. The validity of a visual analogue scale in determining social utility weights for health states. The International Journal of Health Planning and Management 1991 (in press).

Nord E. The Rosser/Kind index revisited. 1991b (forthcoming).

Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. Health Services Research 1973;228-45.

Rawls J. A theory of justice. Harvard University Press. Cambridge 1971.

Read JL. et al. Preferences for health outcomes. Comparison of assessment methods. Medical Decision Making 1984; 4:315-29.

Richardson J. Cost utility analyses in health care: Present status and future issues. Working paper no. 8. Monash University, Public Sector Management Institute, 1989.

Richardson J, Hall J, Salkeld G. Cost utility analysis: The compatibility of measurement techniques and the measurement of utility through time. Paper to the 11th Annual Conference of the Australian Health Economist Group, Australian National University, September 1989.

Rosser R. Issues of measurement in the design of health indicators: A review. In: Culyer AJ (ed). Health Indicators. Oxford, Martin Robertson, 1983.

Rosser R, Kind P. A scale of valuations of states of illness: Is there a social consensus? International Journal of Epidemiology 1978; 7:347-58.

Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. Journal of Chronic Diseases 1978; 31:697-704.

Schoemaker PJ. The expected utility model: Its variants, purposes, evidence and limitations. Journal of Economic Literature 1982; 20:529-563.

Sintonen H. An approach to measuring and valuing health states. Social Science & Medicine 1981; 15c:55-65.

Sutherland HJ, Dunn V, Boyd NF. Measurement of values for states of health with linear

analogue scales. Medical Decision Making 1983; 3:477-87.

Sutherland HJ, Llewellyn-Thomas H, Boyd NF, Till JE. Attitudes toward quality of survival. The concept of "maximum endurable time". Medical Decision Making 1982; 2:299-309.

Torrance GW. Social preferences for health states: An empirical evaluation of three measurement techniques. Socio-Economic Planning Science 1976; 10:129-36.

Torrance GW (b). Health status index models: A unified mathematical view. Management Science 1976; 22:990-999.

Torrance GW. Measurement of health state utilities for economic appraisal. Journal of Health Economics 1986; 5:1-30.

Torrance GW. Utility approach to measuring health related quality of life. Journal of Chronical Diseases 1987;40:593-600.

Torrance GW, Feeny D. Utilities and quality-adjusted life years. International Journal of Technology Assessment in Health Care 1989;5:559-75.

Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science 1981; 211:453-458.

.

Williams A. The measurement and valuations of improvements in health. University of York, Centre for Health Economics. Newsletter 3/1988.

Table 1: Selected results from comparative valuation studies

Author	Year	N	Kind of Subjects	Selected Results SG	RS	TTO	State
Torrance	1976	43	Students	.75	.61	.76	Not indicated.
				.73	.58	.70	
				.60	.44	.63	
				.44	.26	.38	
Bombardier et al	1982	52	Health Care personnel Patients Family	.85	.65	.78	Walking Stick
				.81	.47	.58	Walking frame
				.64	.29	.41	Needs supervision when walking.
				.55	.18	.28	Needs one assistant for walking.
				.38	.08	.11	Needs two assistants.
Llewellyn-Thomas et al	1984	64	Patients	.92	.74		Tired. Sleepless.
				.84	.68		Unable to work. Some pain.
				.75	.53		Limited walking. Unable to work. Tired.
				.66	.47		In house. Unable to work. Vomiting
				.30	.30		In bed in hospital. Needs help self care Trouble remembering.
Read et al	1984	60	Doctors	.90	.72	.83	Moderate angina.
				.71	.35	.53	Severe angina.
Richardson et al	1989	46	Health care personnel				Breast Cancer:
				.86	.75	.80	Removed breast. Unconcerned.
				.44	.48	.41	Removed breast. Stiff arm. Tired. Anxious. Difficulties with sex.
				.19	.24	.16	Cancer spread. Constant pain. Tired. Expecting not to live long.

Table 2 Selected results from comparative valuation studies

Author	Year	N	Kind of Subjects	Selected results RS	ME	PTO	TTO	State
Patrick et al	1973	30	Students	.78	.85	.71		Skin defect
				.60	.66	.58		Pain in abdomen. Limited in social activities.
				.50	.54	.42		Visual impairment. Limited travelling and social activity
				.37	.46	.36		Wheel chair Unable to work.
				.28	.36	.32		In hospital. Limited walking.Back pain. Needs help for self care.Loss of consc
Kaplan et al	1979	54	Psychology	.93	.44			Polluted air.
			Students	.67	.13			Limited walking. Pain in arms/legs.
				.49	.06			Wheel chair. Help for self care. Large burn.
				.25	.02			Small child. In bed Loss of conscious.
Sintonen	1981	60	Colleagues	.61	.72			Difficulties in moving outdoors.
				.45	.51			Needs help outdoors
				.25	.34			Needs help indoors also.
				.09	.15			Bed-ridden
				.04	.04			Unconscious.
Buxton et al [1]	1987	121	Health care personnel	.997	.72			Breast cancer: Removed part of breast. Occasionally concerned.
			University staff	.994	.70			Removed breast. Occasionally concerned.

[1] Magnitude estimation values were obtained by applying the Rosser/Kind index (Rosser and Kind, 1978)

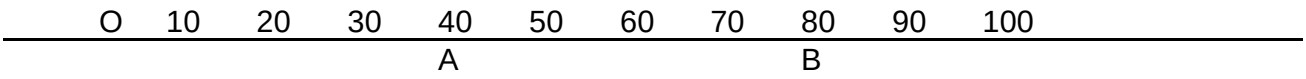
Table 2 (Cont'd)

Author	Year	N	Kind of Subjects	Selected results RS	ME	PTO	TTO	State
					.987		.68	Removed breast. Occasionally concerned, also about appearance.
					.917		.27	Removed part of breast. Stiffness of arm. Engulfed by fear. Unable to meet people.
					.910		.24	Removed whole breast. Otherwise as previous case.
Nord [2]	1991 1991b	22	General public	.71		.985		Moderate pain. depressed.
				.65		.98		Unable to work, moderate pain.
				.30		.97		Unable to work, limited leisure activity, moderate pain, depressed.
				.20		.90		Problems with walking, unable to work, limited leisure activity, strong pain, depressed.

[2] The person trade-off values are transformed from raw scores published in Nord (1991). This study did not include the state "dead". The transformations to a 1-0 scale are based on a subsequent separate valuation of "dead", still using person trade-off (Nord, 1991b).

Table 3: Different subject statements that two life year weights (A = 0.4, B = 0.8) may be derived from.

Category rating:



Magnitude estimation:

I think state A is half as good as state B
(Kaplan et al., 1979)

Standard gamble:

I think it is just as good to gamble with a chance of 0.4 of getting well immediately and a chance of 0.6 of dying immediately as to live with certainty in state A.

Similar for state B.

Time trade-off:

I think it is just as good to live 0.4 year as healthy as living 1 year in state A.

Similar for B.

Person trade-off

I think it is just as good to cure 1 person in state A as curing three people in state B.

Table 4: Face validity of different valuation techniques with respect to determining utility weight for life years in clinical decision analysis and program evaluation.

	Preference concepts covered in response						Face validity		
	Self	Others	Trade Off	Numeric	QoL	Duration	Number /people	Clinical decision	Program eval.
Visual analogue scale	+			+	+			Low	Low
Magnitude est.	+			+	+			Low	Low
Standard gamble	+		+	+	+			Medium	Low
Time trade-off	+		+	+	+	+		High	Medium
Person trade-off	+	+	+	+			+	Medium	Medium