CENTRE FOR HEALTH
PROGRAM EVALUATION

WORKING PAPER 23

SOCIAL EVALUATION OF HEALTH CARE VERSUS PERSONAL EVALUATION OF HEALTH STATES: EVIDENCE ON THE VALIDITY OF FOUR HEALTH STATE SCALING INSTRUMENTS USING NORWEGIAN AND AUSTRALIAN SURVEY DATA

Erik Nord
Nation Institute of Pubic Health, Norway

Jeff Richardson
Professor and Co-Director, Centre for Health Program Evaluation

Kelly Macarounas-Kirchmann
Research Assistant, Centre for Health Program Evaluation

September 1992
ISSN 1038-9547
ISBN 1 875677 00 3

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

The authors are indebted to Gouke Bonsel for giving valuable comments at an early stage on the interpretation of the Norwegian data. All conclusions are entirely the authors' responsibility.

ABSTRACT

In most of the cost-utility literature, QALY gains are interpreted as a measure of social value. Given this interpretation, the validity of different multi-attribute health state scaling instruments may be tested by comparing the values they provide on the 0-1 QALY-scale with directly elicited preferences for person trade-offs between different treatments (equivalence of numbers of different patients treated). Norwegian and Australian public preferences as measured by the person trade-off suggest that the EuroQol Instrument assigns excessively low values to health states. This seems to be even more true of the McMaster Health Classification System. The Quality of Well-Being Scale appears to compress states towards the middle of the 0-1 scale. By contrast the Rosser/Kind index fits reasonably well with directly measured person trade-off data.

Social evaluation of health care versus personal evaluation of health status: evidence on the validity of four health state scaling instruments using Norwegian and Australian survey data

1 Introduction

Quality adjusted life years (QALYs) are increasingly used as an outcome measure in health program evaluation. The measure is based on a procedure in which health states are assigned a value on a scale from unity (healthy) to zero (dead) and sometimes to a value below zero (worse than dead). Several multidimensional, non-disease-specific, scaling instruments are available that allow decision makers to look up values for any health state in which they might be interested. The values incorporated in these instruments have been elicited by means of various techniques, including the standard gamble, time trade-off, rating scales and magnitude estimation. The techniques give different results (for reviews, see Richardson 1991; Nord 1992). The question inevitably arises: Which valuation technique, and which scaling instrument, is the most appropriate and how should this be judged?

Researchers engaged in developing health state scaling techniques have been hesitant to answer these questions (Rosser, 1983; Torrance, 1986; Williams, 1988). This uncertainty at the theoretical level has permitted an arbitrary use of health state values in applied cost-utility analysis. In a review of 15 studies, Nord (1992) found that 24 out of 36 valuations were based on the author's own judgement. Four valuations were obtained by applying an established scaling instrument, but in none of these cases was there an explanation of why the instrument was preferred to the other available instruments. None of the fifteen articles included a discussion of the health state values that could support them theoretically or make them seem plausible in terms of their implications for social choice. Similarly in the studies reported in Torrance's (1986) review, each of the four scaling techniques mentioned above was described as a method for measuring 'utility' although, as noted by Torrance, the relationship between the techniques is not well understood.

We believe that some of the uncertainty and arbitrariness in this field may be overcome. As a first step to achieve this, it is helpful to distinguish between two interpretations of the number of "QALYs gained" (Q):

1. As a measure of <u>production</u> : Q = the increase in the amount of subjectively perceived well-life, i.e. the sum of increments to individual utilities.
2. As a measure of <u>social value</u> : Q = the social value assigned to the program by the people from whom the health state values were elicited.

In the former case, the QALYs gained may be entered as an argument in a social welfare function which, in addition, includes ethical and other distributional considerations (Richardson, 1991; Mooney & Olsen, 1991; Wagstaff, 1991). In the latter case, the number of QALYs gained is itself supposed to encapsulate considerations of distribution as well as efficiency.

In most of the cost-utility literature, QALY gains have been implicitly interpreted in the latter sense, i.e. as a measure of social value. This is clearly seen in the increasingly widespread practice of publishing cost-per-QALY league tables (Williams, 1985; Smith, 1990; O'Kelly & Westaby, 1990). Weinstein and Stason (1977) state the position explicitly: "Alternative programs or services are then ranked, from the lowest value to the highest, and selected from the top until available resources are exhausted". Williams (1987) puts it similarly: "The implications of such calculations seems to me to be that we should not expand treatment capacity where cost-per-QALY is high <u>if</u> there are untreated patients due to lack of capacity in technologies offering low-cost QALYs". It would be difficult to disagree with these statements if equity considerations were explicitly deemed to be irrelevant to decision making or if each of the projects under consideration had the same distributional implications. If either of these preconditions is met then there would be no distinction between the two interpretations of QALYs. However, with the special status attached to health in most countries it cannot simply be assumed that they are fulfilled. The authors arguing for the use of league tables do not appear to recognise the significance of these preconditions.

If QALYs gained are interpreted as measures of social value, the validity of the values obtained from different scaling techniques may be tested by asking whether the people from whom the values were elicited actually agree with the consequences in terms of the implied priorities for different health programs (Loomes & McKenzie, 1989; Nord, 1991; Nord, 1992). This is a test of so called reflective equilibrium (Rawls, 1971).

One method for performing such a test is to look at the implications of scale values for the number of people receiving treatment of one kind that would be equivalent in value to one person receiving treatment of another kind, and to compare this number with people's directly elicited preferences for such a trade-off. In the literature the latter measurement technique is referred to as the equivalence of numbers procedure (Torrance, 1986). Following Nord (Nord, 1992), we will in the present study refer to it as the "person trade-off technique".

Suppose for example a particular technique was used with a group of subjects to assign values v_a and v_b to states A and B on the 1-0 scale, where $v_a > v_b$. According to the social value interpretation of QALYs, the following judgement could then be inferred: Curing one person in State B for one year and gaining $(1-v_b)$ QALYs is of equal value for these subjects as curing N people in state A for one year, and thereby gaining $N(1-v_a)$ QALYs (where $N=(1-v_b)/(1-v_a)$). The statement may then be tested using the person trade-off technique, i.e. by asking the group directly the number of patients receiving the second treatment that they would consider as being of equivalent value to a specified number receiving the first treatment. Such a test does not, of course, allow for all considerations of equity or access. But if the directly elicited equivalence number differs significantly from the inferred one, there is a strong case for rejecting the health state scalings as a basis for calculating QALYs in the social value sense.

As noted by Mulley (1989), there are surprisingly few tests of this kind published in the QALY literature. However, some data do exist. Nord (1991) tested values elicited with the EuroQol rating scale in the way described above. In addition, the public reactions to the first priority list in Oregon (Hadorn, 1991) may be seen as an instance of health state scalings being subject to an informal, real world test of reflective equilibrium (Eddy, 1991; Nord, 1991). A further simple test of reflective equilibrium is to use the published values of existing multi-attribute instruments to determine whether their implications are plausible in terms of person trade-offs.

There are presently three well known non-disease-specific health state scaling instruments available that cover a wide range of health states and provide values for each state. These are the Quality of Well-Being Scale (see for instance [Kaplan & Anderson, 1988]), the McMaster Health Classification System (Torrance, et al. 1982) and the Rosser/Kind index (Rosser & Kind, 1978). In addition there is the EuroQol Instrument (The EuroQol Group, 1990), which also covers a wide range of states but where values to date have only been determined for a limited number of states.

The purpose of the present paper is to report the results of a joint Norwegian-Australian study in which the person trade-off technique is used to test the validity of the four generic instruments as measures of social value. Originally, our empirical data were collected with a view to validating the EuroQol Instrument. We therefore start by addressing this instrument and proceed to comment on the three other scales in subsequent sections.

We emphasise that the paper concerns the use of the four instruments in question for assessing the social value of different health care programs, and hence to their role in allocating resources in the health sector. No judgement is made with respect to their validity as a measure of production or as a measure of the individual utility gains of different therapies. For a discussion of these issues, see (Richardson, 1991; Nord, 1992).

The EuroQol Instrument

A group of European researchers known as The EuroQol Group has recently developed a standardised non-disease-specific instrument for describing and valuing health related quality of life (The EuroQol Group, 1990; Essink-Bot, et al. 1990; Kind, 1990; Brooks, et al. 1991; Nord, 1991). The instrument consists of a self administered questionnaire, in which subjects are asked to value health states on a visual analogue rating scale running from zero ("worst imaginable health state") to one hundred ("best imaginable health state"). In the first version of the instrument all states were described by six dimensions with two or three levels of functioning along each dimension (mobility (1/2/3), self care (1/2/3), major activity (1/2), leisure activity (1/2), pain (1/2/3), anxiety/depression (1/2)).

As noted above, Nord (1991) carried out a formal test of reflective equilibrium by asking subjects to locate various health states on the EuroQol rating scale and then to make pair-wise valuations of some of the health states. In each pair, the subjects were asked to nominate the number of cured patients in the less severe state that would be equivalent to curing one patient in the more severe state.

With both valuation techniques, analysis was carried out using median values, since, with the person trade-off technique, mean values may be unduly influenced by individuals responding with very high numbers which do not have real cardinal significance. The median may be interpreted simply as a measure of central tendency but also in terms of a majority view: If X is the median, then a majority are against assigning a lower value than X and also against assigning a higher value than X.

In six different subgroups, the median equivalence numbers were consistently much greater than those obtained by using the median rating scale scores as life year weights in the conventional QALY algorithm. Table 1 gives some examples.

Nord's study focused on programs which cured patients in various states of illness. There was no anchoring of the valuations to the value of life saving. The joint Norwegian-Australian study was constructed to rectify this omission. A self administered questionnaire was designed in which subjects were asked how they, as members of Parliament, would evaluate two proposed, equally costly, special units A and B. Unit A would save ten people per year from dying and give them full health. Unit B would cure a number of people in a state of chronic illness and return them to full health. The question put to the subjects was: How many patients must be treated per year in unit B in order that you would find it just as valuable to spend the money on unit B as on unit A? Each subject was asked to evaluate one state of chronic illness in this way.

As a first step in both the Norwegian and the Australian study, two EuroQol states were used to describe states of chronic ill health (states A and B in table 2). As will be shown below, the results obtained for these two states in Norway were illogical, as the less severe state was given a lower median equivalence number than the more severe state. This may have been due to the complexity of the health state descriptions. An additional round of data collection with less complex descriptions (Z and W in table 2) was therefore conducted in both countries.

In Norway the subjects were partly a random population sample, selected by the Central Bureau of Statistics and partly a sample of the respondents to the preceding study (Nord, 1991). Altogether, 386 people were sent questionnaires. 109 were returned, representing a response rate 28.2 %. There were 102 useable answers.

In Australia a convenience sample of students and nurses was selected randomly from the personnel records of a university and a nearby hospital. In total 1442 people were sent a questionnaire and 384 useable answers were returned, representing a response rate of 27%.

The Australian study included a sub-experiment in which the presentation of the EuroQol states was varied in order to test for a possible framing effect. In half of the questionnaires, the lay out was as for states A and B in table 2. In the other half, the order of the items was changed, so that the positive items ("no problems") came first. The results were very much alike for both layouts. This is in accordance with a previous finding by Nord (1991). Using a Mann-Whitney test it was not possible to reject the null hypothesis that the median values were the same from both questionnaires. Consequently, the data were pooled for the analysis reported here.

Table 3 summarises the personal characteristics of the respondents. There is a moderate selection bias in favour of men and the well-educated in the Norwegian sample. In the Australian study, nurses gave a higher median equivalence number than students for state B, but a lower median than the students for state A. Neither difference was significant as judged by the Mann-Whitney test, and the data from the two groups were pooled.

In 28 cases in the Australian sample, zero was given as the equivalence number. Since it is difficult to see any meaning in such a response, these cases are excluded from the analysis below.

The distribution of equivalence numbers for each health state are shown in table 4. The responses are highly dispersed, indicating the likelihood of a high sampling error for the median values. Confidence intervals are indicated in the final row of the table (Gardner & Altman, 1989).

Table 5 reports the results in terms of both the 0-1 scale and person trade-offs. The directly measured EuroQol values in column A are the average values obtained in the four countries in which the instrument has been tested (Nord, 1991). States W and Z are scored on the basis of the authors' own judgement: Z is located somewhat below the two EuroQol states in question, since it includes sitting in a wheel chair (see table 2), while W is located somewhat below EuroQol state 112121, which had an average valve of .65 points in three countries (The EuroQol Group, 1990).

Results in the two countries are quite similar. With three of the four states, the implied values differences on the 0-1 scale are five percentage points or less. (As noted above, the Norwegian value for state A is anomalous.) In all eight cases in table 5, the values on the 0-1 scale implied by the person trade-off responses by far exceed the EuroQol values.

If the EuroQol instrument was measuring the same values as the person trade-off technique then the median values in table 5 should be equal for each health state with differences only attributable to randomly distributed measurement errors. That is, the probability of the person trade-off score exceeding the EuroQol score for any given observation would be 0.5. The probability of this occurring in all 8 comparisons, as in table 5, is therefore $(0.5)^8$ - or less than 0.5 per cent. As noted above, the equivalence number obtained for state A in one of the Norwegian sub-samples seems unreasonable, as it exceeds the equivalence number obtained for the less severe state B in both Norway and Australia. However, even with the omission of this observation the probability of the remaining results occurring by chance is less than 1 per cent. This strongly supports the view that the EuroQol instrument seriously underestimates the social value placed upon the health states when the alternative is death. Casual observation of the results further suggests that the problem is particularly acute at the lower end of the scale. This is consistent with earlier findings (Nord, 1991).

The implication of these conclusions is that the scaling technique used for the EuroQol does not accurately reflect the social value of life per sé relative to the quality of life. It is possible that the EuroQol scores could be transformed to obtain a valid indicator of social value. Nord (1992) has examined this possibility using Norwegian data. Other inquiries into the problem of transforming rating scale values have been made by Torrance (1976), Loomes (1988) and Richardson et al (1990). However, it must be stressed that to date the theoretical and empirical relationships between rating scale values and social values are not sufficiently understood.

Other instruments

The Quality of Well-Being Scale (QWB) was developed by Kaplan and colleagues at the University of California in San Diego (for a brief history, see [Kaplan & Anderson, 1988]). It is based upon a rating scale extending from ten (well) to zero (dead) and a health state classification system consisting of three different dimensions of function and 25 symptom/-problem complexes. Community surveys have been conducted in which respondents were asked to use the rating scale to indicate the disutility of a single day with each kind of dysfunction and symptom. On the basis of these uni-dimensional disutility judgements, the value of any composite health state within the classification system can be determined on the standard 1-0 QALY-scale by means of a simple additive formula (see appendix).

The Rosser/Kind index (Rosser & Kind, 1978) covers 28 combinations of disability and pain/distress, as well as the state dead. The valuations were obtained by means of magnitude estimation: "No disability and mild distress" was chosen as a reference state. Each of the other states was scaled by asking a sample of 70 doctors, nurses, patients and others "how many times more ill" a patient in that state would be than a patient in the reference state. To clarify the meaning of the question, respondents were asked to imagine that their answer would define the proportion of resources that should be allocated to the relief of each health state.

The McMaster Health Classification System was developed by Torrance et al (1982). It has four dimensions - physical function, role function, social-emotional function and health problem - each subdivided into a number of levels. Each level has a weight between unity and zero based on a combination of rating scale and time-trade off interviews in a community sample of healthy adults. Any health state may be scaled by entering weights for levels fitting that state into a multiplicative formula (see appendix).

To our knowledge, no person trade-off test has been carried out on states described by the descriptive system of the Quality of Well-Being Scale, the Rosser/Kind index or the McMaster Health Classification System. However, we have mapped the states included in our study (A, B, W and Z) into each of these instruments and subsequently scaled them (in the same way as we scaled states W and Z by means of the EuroQol instrument in the previous section).

The mapping is inevitably inexact, as each of the systems differ with respect to the dimensions used and the functional descriptions within each dimension. Some of the weights assigned to the health states by the three non-EuroQol instruments are therefore given as ranges. For details of the mapping, see appendix. Table 6 presents the values obtained from the mapping together with the values implied by the person trade-off results reported in the previous section.

Results in the last section supported the view that the rating scale based EuroQol assigns values that are consistently too low if the objective is to measure the social values that are incorporated in the person trade-off procedure. The results presented in this section are more tenuous because the number of independent observations obtained from our surveys is small. However, inspection of table 6 suggests some tentative conclusions. The first is that the McMaster instrument may underestimate social value even more than the EuroQol. Secondly, while the QWB produces results which correspond better with the person trade-off technique, it appears to underestimate social value quite strongly at the upper end of the scale. Thirdly, the Rosser/Kind index assigns values that in all four cases are quite close to the values implied by our person trade-off data.

Discussion

The article commenced by noting the distinction between the "production of well life" and "social value". In welfare economics, the latter would be measured by the "Social Welfare Function" which would combine information about production, distribution and possibly considerations of "process" to determine the social desirability of different health programs. In this framework resource allocation should be determined by social value and not by the value of production per se.

With some exceptions (Mooney & Olsen, 1991; Richardson, 1991; Wagstaff, 1991) this distinction has been largely ignored in the cost utility literature and even when discussed it has been in theoretical terms. This is not surprising. Quantifying social value is conceptually hazardous as there is no agreement either about how it should be measured or about which ethical theory should form the basis for measurement. (Even this later statement presumes that societal judgements should be based upon a single theory.) There is an ethical system under which QALYs as a measure of production would correspond with social value as implicitly assumed in much of the literature. This system has been described by Mooney and Olsen (1991) as "Quasi - utilitarianism". Social welfare is equal to a weighted average of individually determined utilities where the weights ensure that each persons life year is equally important irrespective of the individuals personal characteristics or capacity to appreciate life years. The rule is almost certainly defective as it ignores distributional considerations and issues of entitlement which are known to be of importance in decision making, especially in the health sector (Harris, 1987; Nord, in press).

In the present article we have employed one plausible device for eliciting social choice namely the person trade-off technique. Its present application does not permit all issues of process or equity to influence the calculation of social values as might occur in a more general Social Welfare Function, but it does permit a clear distinction to be made between the production of health and its social value, all else equal.

Our empirical data have obvious limitations. They were elicited from small samples of people (particularly for states W and Z in Norway), and the response rates were low. The results are therefore not necessarily representative of the Norwegian or Australian population at large, let alone populations in other countries. It is also a disturbing fact that in one of our sub-samples, we obtained a logically inconsistent equivalence number for state A. The data nevertheless form an interesting pattern in relation to the health state scaling instruments in question, and we feel that they must cause considerable concern about the use of several of the instruments for health program evaluation.

The lower end compression of states on the McMaster scale is remarkable, not only in relation to our person trade-off data, but also relative to the other three scaling instruments. A closer look at the set of weights that the instrument uses makes this lower end compression quite understandable. A number of the weights are such that the implied person trade-offs are strongly counter intuitive. As shown in table 7, having "some limitations in physical ability to lift, walk run, jump or bend" scores 0.87. This implies that restoring 8 such patients to full health is equivalent to saving a life (given equal life expectancy after treatment). The same trade-off (8 cured versus 1 life saved) is supposed to apply to curing people "needing a hearing aid" or "having pain or discomfort for a few days in a row every month". "Needing mechanical aids to get around but not needing help from other people" scores 0.73. The implied trade-off between curing and life saving is approximately 4. We believe that these values are unreasonable and that they would be widely rejected.

Similar observations can be made concerning the QWB (table 7). States of dysfunction only consisting of either a cough, runny nose, headache, trouble with sleeping, talking with a lisp, pimples or spells of feeling upset are all assigned values in the range from 0.74 to 0.83. When these values are put into a QALY algorithm (in the social value sense), the implication is that relieving 4-6 people of such symptoms should be considered equivalent to saving a life. Again, this is not a reasonable description of social values.

The undervaluing of relatively mild states of illness by the QWB was recently indicated by the public reaction to the priorities suggested by its use. The state of Oregon in the US employed a version of the scale to rank a large number of health care procedures according to their cost-effectiveness ratio. A draft priority list for the Medicaid program based on these rankings, released in May 1990, contained a number of counter intuitive rank orderings. Specifically, some procedures deemed highly beneficial or life saving were placed below routine procedures like headache treatment or tooth capping (Hadorn, 1991).

As noted by Eddy (1991), an important reason for this seems to be that the QWB assigned too low values to trivial states of illness. This again lead to low equivalence numbers for trivial treatments compared to treatments for severe conditions (for a further discussion, see (Nord, 1991).

While the QWB is clearly undervaluing less severe health states, the more serious states included in the present study were assigned values closer to the person trade-off scores. We see two possible explanations of this. Firstly, there is a possible "anchoring" effect associated with the standard version of the QWB (Nord, 1992). As noted above, the scale used to establish the standard set of disutility weights extended from ten to zero. The instructions 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". "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"). This could have forced quite severe states into the upper half of the scale. The present study lends some support to this hypothesis.

Another important feature of the standard QWB is that its disutility weights express the undesirability of a single day in a particular state. As noted elsewhere (Nord, 1992), even quite severe functional limitations may be well tolerable for such a short time period (for instance being bedridden), while trivial symptoms such as a runny nose or a headache may be perceived as quite unpleasant. In other words, the time frame presented to respondents is likely to have seriously compressed values into the middle of the scale.

The Rosser/Kind index values are quite close to those implied by person trade-off measurements. This may be due to the similarity in framing questions. As noted earlier Rosser and Kind informed subjects that their initial responses to the magnitude estimation questions would be interpreted in terms of preferences for resource allocation for 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. As a consequence, their technique was very similar to the person trade-off technique in the present study.

The Rosser/Kind index has been criticised on the grounds that most health states are assigned values close to unity. Consequently, quality adjusted life years will not be very different from ordinary life years. QALYs will then add very little to decision makers' customary indicator of life expectancy (Mulkay, et al. 1987).

The present results lend some support to the view that this "upper end compression" is appropriate. It does not follow from this that Mulkay et al's (1987) conclusion is correct. Upper end compression reflects the high value that people seem to place on life saving programs relative to health improving programs. At the same time, the weights clearly permit discrimination between different health improving program. To see this, assume that two states A and B are assigned weights 0.95 and 0.99 respectively. Curing one person in state A then renders five times as many QALYs as curing one person in state B. This information may certainly be of interest to decision makers.

The lack of upper end compression in the EuroQol, McMaster and the QWB instruments can be defended on the grounds that they may have been designed for individual utility measurement rather than as instruments for assigning social value to health care programs. Each of these instruments is based on valuation exercises in which people were asked how they thought they, personally, would feel in different states of illness. Ethical and social considerations were not entered into the calculations. However, these very important limitations are insufficiently emphasised in the promotion of the instruments and are likely to escape the attention of practitioners. This seems to be precisely the mistake made by the Oregon Health Services Commission, which attributed a person trade-off meaning to QWB-values in a way that was not intended by the subjects from whom the values were elicited.

Concluding remarks

Health state scaling instruments are commonly used in cost-utility analysis to estimate the (relative) value assigned to different health care programs by society. To date there has been little attempt to validate these scales as instruments for quantifying social as distinct from individual values. To our knowledge this is the first study that explicitly addresses this issue. The person trade-off as applied here may be viewed as an important test of the validity of these instruments.

Norwegian and Australian public preferences as measured by the person trade-off suggest that the EuroQol Instrument assigns excessively low values to health states. This seems to be even more true of the McMaster Health Classification System. The QWB appears to compress states towards the middle of the 1-0 scale. By contrast the Rosser/Kind index fits reasonably well with our directly measured person trade-off data. We recognise that, given the limited observations available from our surveys, these must be tentative conclusions. Clearly there is a need for more comprehensive person trade-off and other validation studies as well as studies in countries other than Norway and Australia. The potential cost of not adequately validating existing scales has, in our view, been demonstrated quite dramatically by the failure of the QWB-based draft priority list in Oregon. The EuroQol Group has recognised these problems and put on its agenda studies of the relationship between individual health state valuations and social utility weights for different health improvements (Essink-Bot, et al. 1990; Nord, 1992). Similar research is encouraged elsewhere.

Despite its limitations, the present study suggests that health state valuations provided by the other three well known instruments may need considerable adjustment before being appropriate for the calculation of QALYs, at least with the second interpretation of these discussed in this article, namely, as the measure of social value which is appropriate for decisions about the allocation of resources.

BIBLIOGRAPHY

1. Brooks, R.G., Jendteg, S., Lindgren, B., Persson, U., Bjørk, M.S. 1991 "EuroQol: Health related quality of life measurement. Results of the Swedish questionnaire exercise", *Health Policy*, vol. 18, pp. 37-48.

2. Drummond, M.F., Stoddart, G.L., Torrance, G.W. 1987 *Methods for the economic evaluation of health care programs*. Oxford University Press.

3. Eddy, D.M. 1991 "Oregon's Methods: Did cost-effectiveness analysis fail?", *JAMA*, vol. 266, pp. 2135-2141.

4. Essink-Bot, M.L., Bonsel, G., van der Maas, P.J. 1990 "Valuation of health states by the general public: Feasibility of a standardised measurement procedure", *Social Science and Medicine*, vol. 31, pp. 1201-1206.

5. Gardner, M.J., Altman, D.G. 1989 "Statistics with confidence. Confidence intervals and statistical guidelines", *British Medical Journal*.

6. Hadorn, D.C. 1991 "Setting health care priorities in Oregon", *JAMA*, 1991, 265, 2218-2225.

7. Harris, J. 1987 "QALYfying the value of life", *Journal of Medical Ethics*, vol. 13, pp. 117-123.

8. Kaplan, R.M., Anderson, J.P. 1988 "A general health model: Update and applications", *Health Services Research*, vol. 23, pp. 203-235.

9. Kind, P. 1990 *Measuring valuations for health states: Piloting the EuroQol questionnaire*, Discussion paper No. 76. York: University of York, Centre for Health Economics.

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

11. Loomes, G., McKenzie, L. 1989 "The use of QALYs in health care decision making", *Social Science & Medicine*, vol. 28, pp. 299-308.

12. Mooney, G., Olsen, J.A., 1991 "QALYs: Where next?", In: McGuire A et al (eds). *Providing health care: The economics of alternative systems of finance and delivery*. Oxford University Press.

13. Mulkay, M., Ashmore, M., Pinch, T. 1987 "Measuring the quality of life: A sociological invention concerning the application of economics to health care", *Sociology,* vol. 21, pp. 541-564.

14. Mulley, A.G. 1989 "Assessing patients' utilities. Can the ends justify the means?", *Medical Care*, vol. 27, pp. S269-281.

15. Nord, E. 1991 "EuroQol: Health related quality of life measurement:. Valuations of health states by the general public in Norway", *Health Policy*, vol. 18, pp. 25-36.

16. Nord, E. 1991 "The validity of a visual analogue scale in determining social utility weights for health states", *International Journal of Health Planning and Management*, vol. 6, pp. 234-242.

17. Nord, E. 1991 *Unjustified use of the Quality of Well-Being Scale in priority setting in Oregon*. Mimeo. Oslo: National Institute of Public Health.

18. Nord, E. 1991 "The relevance of health state after treatment in prioritising between different patients", *Journal of Medical Ethics* (in press).

19. Nord, E. 1992 "Methods for quality adjustment of life years", *Social Science & Medicine*, vol. 34, pp. 559-569.

20. Nord, E. 1992 "The use of EuroQol values in QALY calculations", In: Bjørk S (ed). *EuroQol Conference Proceedings*. IHE working paper 1992:2. Lund: Swedish Institute of Health Economics.

21. Nord, E. 1992 "Towards quality assurance in QALY calculations", *The International Journal of Technology Assessment in Health Care*, (in press).

22. O'Kelly, T.J, Westaby, S. 1990 "Trauma centres and the efficient use of financial resources", *British Journal of Surgery*, vol. 77, pp. 1142-1144.

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

24. Richardson, J. 1991 "Economic assessment of health care: Theory and practice", *The Australian Economic Review*, 1st quarter, 4-21.

25. Richardson, J. 1991 "What should we measure in health program evaluation?", In Smith S (ed). *Economics and health*. Proceedings from the 12th Australian Conference of Health Economists. Melbourne: Public Sector Management Institute, Monash University.

26. Richardson, J., Hal,l J., Salkeld, G. 1990 "Cost-utility analysis: The compatibility of measurement techniques and the measurement of utility through time", In Smith S (ed). *Economics and health*. Proceedings from the 11th Australian Conference of Health Economists. Melbourne: Public Sector Management Institute, Monash University.

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

28. Rosser, R., Kind, P. 1983 "A scale of valuations of states of illness: Is there a social consensus?", *International Journal of Epidemiology*, vol. 7, pp. 347-358.

29. Smith, G.T. 1990 "The economics of hypertension and stroke", *American Heart Journal*, vol. 119, pp. 725-728.

30. The EuroQol Group. 1990 "EuroQol - a new facility for the measurement of health related quality of life", *Health Policy*, vol. 16, pp. 199-208.

31. Torrance, G.W. 1976 "Social preferences for health states: An empirical evaluation of three measurement techniques", *Socio-Economic Planning Science*, vol. 10, pp. 129-136.

32. Torrance, G.W. 1986 "Measurement of health state utilities for economic appraisal", *Journal of Health Economics*, vol. 5, pp. 1-30.

33. Torrance, G.W., Boyle, M.H., Horwood, S.P. 1982 "Application of multi attribute utility theory to measure social preferences for health states", *Operations Research*, vol. 30, pp. 1043-1069.

34. Wagstaff, A. 1991 "QALYs and the equity-efficiency trade-off", *Journal of Health Economics*, vol. 10, pp. 21-41.

35. Weinstein, M.C., Stason, W.B. 1977 "Foundations of cost-effectiveness analysis for health analysis and medical practices", *New England Journal of Medicine,* vol. 296, pp. 716-721.

36. Williams, A. 1985 "Economics of coronary artery bypass grafting", *British Medical Journal*, vol. 291, pp. 326-329.

37. Williams, A. 1987 "Who is to live? A question for the economist or the doctor?", *World Hospitals*, vol. 13, pp. 34-36.

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

APPENDIX A:
Mapping of states into other descriptive systems.

State A (EuroQol 212232).

1. The Quality of Well-Being Scale

Mobility: Step 4, weight -.062: Would have used more help than usual for age to use public transportation, health related. (Next step: In hospital. Previous step: No limitations.)

Physical activity: Step 3, weight -.060: Problems with walking, but less than being in a wheelchair and being dependent on others for moving the chair. (Previous step: No limitations.)

Social activity: Step 2, weight -.061: Performed no major role activity, health related, but did perform self care activities. (Next step: Needed help with self care as well. Previous step: Limited in major role activity.)

Symptom/problem: (Groups 4, 7, 8, 13 16 and 18 all include pain. The weights vary from -0.170 to -0.349. A range of -0.250 to -0.350 is chosen since state A includes "strong pain".

Total QWB score: High: 1-0.062-0.060-0.061-0.250 = 0.567.
Low: 1-0.062-0.060-0.061-0.350 = 0.467

2. The McMaster Health Classification System

Physical function: Physical function: Step 3, factor 0.81: Being able to get around without help from another person and needing mechanical aid to walk or get around.

Role function: Step 3, factor 0.77: Being able to eat, dress, bathe and go to the toilet without help; and not being able to play, go to school or work.

Social/emotional function: Step 3 or 4, factor 0.86 or 0.77: Being anxious or depressed some or a good bit of the time, and having an average (step 3) or very few (step 4) friends and contacts with others.

Health problem: Step 4, factor 0.91: Having a medical problem which causes pain or discomfort for a few days in a row every two months. (This is the only step that includes pain.)

Total McMaster score: High: 1.42(0.81x0.77x0.86x0.91)-0.42 = 0.27
Low: 1.42(0.81x0.77x0.77x0.91)-0.42 = 0.20

3. The Rosser/Kind index

The scale assigns 0.700 to "unable to work and severe pain" and 0.000 to "unable to work, chair bound and severe pain". For state A, which includes dependency on crutches, we arbitrarily assign the value 0.4-0.6.

State B (EuroQol 112232).

1. The Quality of Well-Being Scale

Mobility: Step 5, weight -0.00 (no limitations) or step 4, weight -.062 (see state A).

Physical activity: Step 4, weight -.000: No limitations for health reasons. (Next step: Problems with walking).

Social activity: Step 2, weight -.061 (see state A).

Symptom/problem: Weight -0.250 to -.350 (see state A).

Total QWB score: High: 1-0.062-0.000-0.061-0.250 = 0.637
Low: 1-0.062-0.000-0.061-0.350 = 0.537

2. The McMaster Health Classification System

Physical function: Step 1 or 2, factor 1.00 or 0.91: Being able to get around without help ... and having no (step 1) or some (step 2) limitation in ability to lift, walk, run, jump or bend.

Role function: Step 3, factor 0.77 (see state A).

Social/emotional Step 3 or 4, factor 0.86 or 0.77 (see state A).

Health problem: Step 4, factor 0.91 (see state A).

Total McMaster score: High: 1.42(1.00x0.77x0.86x0.91)-0.42 = 0.44
Low: 1.42(0.91x0.77x0.77x0.91)-0.42 = 0.28

3. The Rosser/Kind index

Disability: Step V: Unable to work. (Next: chair bound.)

Pain/Distress: Step D: Severe pain/distress.

Rosser/Kind score: 0.70

State W.

1. The Quality of Well-Being Scale

Mobility: Step 1, weight -.000: No limitations, or step 4, weight -0.062 (see state A).

Physical activity: As for state A: Step 3, weight -.060.

Social activity: As for previous states: Step 2, weight -.061.

Symptom/problem: The least severe pain group is chosen, i.e. Group 18, weight -.170: Pain in ear, tooth, jaw, throat, lips or tongue.

Total QWB score: High: 1-0.000-0.060-0.061-0.170 = 0.709
Low: 1-0.062-0.060-0.061-0.170 = 0.647

2. The McMaster Health Classification System

Physical function: Step 3, factor 0.81 (see state A).

Role function: Step 3, factor 0.77 (see state A).

Social/emotional function: Step 2, factor 0.96: Being happy and relaxed most of all of the time and having very few friends and little contact with others.

Health problem: Step 4, factor 0.91 (see state A).

Total McMaster score: 1.42(0.81x0.77x0.96x0.91)-0.42 = 0.35

3. The Rosser/Kind index

Disability: Step V: Unable to work.

Pain/distress: Step B: Mild pain/distress.

Rosser/Kind score: 0.935.

State Z.

1. The Quality of Well-Being Scale

Mobility: Step 4, weight -.062 (see state A).

Physical activity: Step 3, weight -.060 (see state A).

Social activity: Step 2, weight -.061 or step 1, weight -.106 (see state A).

Symptom/problem: Group 7 (pain in side, neck, back, hips, joints, legs), weight -.299.

Total QWB score: High: 1-0.062-0.060-0.061-0.299 = 0.518
Low: 1-0.062-0.060-0.106.0.299 = 0.473

2. The McMaster Health Classification System

Physical function: Step 3, factor 0.81 (see state A), or step 5, factor 0.61 (needing help <u>and</u> mechanical aid).

Role function: Step 3, factor 0.77 (see state A), or step 5, factor 0.50 (needing help for self care and not being able to attend work).

Social/emotional function: Step 3, factor 0.86, or step 4, factor 0.77 (see state A).

Health problem: Step 4, factor 0.91 (see state A).

Total McMaster score: High: 1.42(0.81x0.77x0.86x0.91)-0.42 = 0.27
Low: 1.42(0.61x0.50x0.77x0.91)-0.42 =
-0.12

3. The Rosser/Kind index

Disability: Step VI, chairbound.

Pain/distress: Step C, moderate.

Rosser/Kind score: 0.680.

Table 1
Validation using reflective equilibrium: Comparison of the Rating Scales (RS) and Person Trade-off (PTO)

RS Median Values		Number cured from second state equivalent to 1 cured from reference state		N
Reference State	Second State	Implied by RS (Column 1 & 2)	Direct PTO	
15	25	1,1	2	36
15	40	1,4	3	38
15	50	1,7	20	38
15	75	3,4	100	36
15	85	5,7	>10000	39

(Source: (16), Table 2).

Table 2
States of Chronic Illness

A (EuroQol 212232)	Unable to walk without a crutch. Unable to perform main activity. Unable to pursue family and leisure activities. Strong pain. Anxious. No problems with self care.
B (EuroQol 112232)	Unable to perform main activity. Unable to pursue family and leisure activities. Strong pain. Anxious. No problems with walking. No problems with self care.
W	Uses crutches for walking. Light pain intermittently. Unable to work.
Z	Sits in wheel chair. Pain most of the time. Unable to work.

Table 3
Characteristics of the Respondents in Norway and Australia

		Norway		Australia	
		Number	%	Number	%
Sex	Men	46	45.1	161	41.9
	Women	31	30.4	218	56.7
	Missing	25	24.5	5	1.3
	TOTAL	*102*	*100.0*	*384*	*100.0*
Age	18-29	22	21.6	296	77.1
	30-39	12	11.8	39	10.2
	40-49	17	16.7	27	7.0
	50-59	7	6.9	18	4.7
	60-69	11	10.8	1	0.2
	70-84	8	7.8	0	0.0
	Missing	25	24.5	3	0.8
	TOTAL	*102*	*100.0*	*384*	*100.0*
Education	Primary	11	10.8		
	Secondary	19	18.6		
	College/Univ.	10	9.8		
	Secondary or	41	40.2		
	College/Univ.	21	20.6		
	Missing				
		102	*100.0*		
	TOTAL				
				250	65.1
	Student			91	23.7
	Nurse			43	11.2
	Other				
				384	*100.0*
	TOTAL				

Table 4
Comparison of Norwegian and Australian Results.
Distribution in per cent, median values and confidence intervals

Number cured in State i equivalent to 10 saved lives	**State i**							
	A (212232)		**B (112232)**		**W**		**Z**	
	Norway	Aust.	Norway	Aust.	Norway	Aust.	Norway	Aust.
0 - 9	6.2	9.8	5.5	6.1	4.1	0.0	5.9	6.2
10	9.3	36.1	25.0	27.0	8.3	9.3	23.5	21.5
11 - 19	0.0	3.0	11.1	5.2	4.1	2.3	5.9	7.7
20 - 39	12.4	15.8	11.1	7.4	8.3	6.9	11.8	10.8
40 - 50	12.3	7.5	5.6	14.8	4.1	20.9	11.8	20.0
51 - 100	15.5	7.5	16.7	7.8	20.8	16.3	23.5	16.9
101 - 999	15.4	9.8	16.6	10.4	8.3	25.7	5.9	6.1
1000 -	27.8	10.5	8.4	11.3	41.7	18.6	11.8	10.8
TOTAL	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*	*100.0*
N	*32*	*133*	*36*	*115*	*24*	*43*	*17*	*65*
MEDIAN	*100*	*20*	*30*	*26*	*110*	*85*	*50*	*40*
90 % CI	*50 200*	*10 25*	*15 55*	*20 50*	*100 200*	*50 200*	*100 200*	*50 200*

Table 5
Comparison of Results from EuroQol and from Person Trade-off (PTO), Norway and Australia

State	Value on 0-1 Scale		Number returned to full health equivalent to 1 saved life		N[e]
	Direct EuroQol[a] by	Value implied PTO [b]	Direct PTO [c] by	Number implied EuroQol [d]	
A (212232)					
Norway	0.23	0.90	10.0	1.3	32
Aust.	0.23	0.50	2.0	1.3	133
B (112232)					
Norway	0.33	0.67	3.0	1.5	36
Aust.	0.33	0.62	2.6	1.5	115
W Noway	0.60	0.91	11.0	2.5	24
Aust.	0.60	0.88	8.5	2.5	45
Z Norway	0.20	0.80	5.0	1.25	17
Aust.	0.20	0.75	4.0	1.25	69

Notes:

(a) Average value reported from EuroQol studies (Nord, 1991).

(b) If X = number equivalent to saving 1 life; 1.0 = full health; V = value of initial health state (A, B, Z, W); then X(1-V)=1.0 and hence V=(X-1)/X.

(c) Medians in table 4 divided by 10.

(d) X = 1/(1-V), see note (b).

(e) Sample size in PTO survey.

Table 6
Comparison of Results from Four Valuation Methods on a 0-1 Scale

State	Value implied by PTO[a]	QWB [b]	McMaster [b]	Rosser/Kind [b]
212232		.47 - .57	.20 - .27	.40 - .60
Norway	.90			
Aust.	.50			
112232		.54 - .64	.28 - .44	.70
Norway	.67			
Aust.	.62			
W		.65 - .71	.35	.94
Norway	.91			
Aust.	.88			
Z		.47 - .52	(-.12) - .27	.68
Norway	.80			
Aust.	.75			

Notes:

(a) See table 5, column 2.

(b) See appendix for explanation.

Table 7
Equivalence numbers implied by selected values in the McMaster Health Classification System (MMHCS) and the Quality of Well-Being Scale (QWB)

State	Published value of state	Number cured equivalent to saving a life (approximated)
***MMHCS*[a]:** Some limitations in physical ability to lift, walk, run, jump or bend	0.87	8
Needing a hearing aid	0.87	8
Having pain or discomfort for a few days in a row every month	0.87	8
Needing mechanical aids to get around, but not needing help from others	0.73	4
***QWB*[b]** Stuffy, runny nose	0.83	6
Pimples	0.814	5
Lisp	0.763	4
Headache	0.756	4
Spells of feeling upset	0.743	4
Trouble with sleeping	0.743	4
Cough	0.743	4

Notes: (a) See (2).
(b) See for instance (8).