Pergamon

0277-9536(95)00161-1

Soc. Sci. Med. Vol. 42, No. 4, pp. 609–615, 1996
Elsevier Science Ltd. Printed in Great Britain

INCONSISTENCY AND HEALTH STATE VALUATIONS

PAUL DOLAN[1] and PAUL KIND[2]

[1]Department of Economics, University of Newcastle, Newcastle-Upon-Tyne NE2 7RU and
[2]Centre for Health Economics, University of York, Heslington, York YO1 5DD, England

Abstract—The comparison of scaling methods used to value health states sometimes rests upon an analysis of aggregate scores. This analysis is usually undertaken once 'inconsistent' respondents have been excluded from the data. However, it is important to be able to judge the extent to which respondents as a whole are logically consistent when assigning values to health states. The degree of inconsistency will depend on how the health states are described, how the questionnaire is administered and who the respondents are. This paper analyses the inconsistency rates from two studies in which valuations for EuroQol health states were elicited using the visual analogue scale (VAS) method. The studies differed in design and incorporated several different variants of the standard EuroQol questionnaire, thus providing an opportunity to examine the relative importance of the different factors that were thought to affect inconsistency rates. Our general conclusions are that inconsistency rates are higher for interviewer-based than for postal surveys, possibly due to response bias, and that inconsistency rates are positively related to age and negatively related to educational attainment.

INTRODUCTION

It is becoming increasingly common for preference-based measures of health status to be used in clinical decision-making and economic analysis. A central task in the field of health status measurement involves eliciting valuations for health states using one or more different scaling methods, for example category rating, magnitude estimation, equivalence of numbers, graphical rating scales [1]. Individuals generally hold differing views and opinions on a wide range of issues relating to their everyday experiences, so that it might be expected that valuations for health states derived from different respondents would naturally vary also. It is not surprising, then, that within any given method there may be significant variability across population subgroups [2].

However, different scales of values have also been observed from the different scaling methods from the *same* respondents. There are many aspects of the measurement process which may be implicated in the degree of 'inconsistency' found in empirical studies, including the way in which health states are described and the valuation method used. For example, there is evidence to suggest that the valuations given to health states when described in terms of a narrative scenario are lower than those given to the same health state described in a standardised outline form [3]. Also, it has been found that the standard gamble generates internally inconsistent valuations; when the certain alternative remains unchanged, changes in the outcomes of the risky alternative significantly influence the values elicited [4].

Both of these examples refer to inconsistencies that result as a consequence of some aspect of the measurement procedure. Kahneman and Tversky (1981) refer to such inconsistencies, or situations where "changes in perspective often reverse the relative apparent size of objects and the relative desirability of options", as *framing effects* [5]. Much of the literature has focused on inconsistencies of this kind. For example, McNeil *et al.* (1982) showed that the preferences of respondents were different when a clinical decision was framed in terms of the probability of surviving as opposed to the probability of dying [6].

Inconsistencies that result from framing effects are to some extent to be expected given that most respondents, be they the general public, patients, or health professionals, are rarely asked to express their preferences on health or health care in quite the way they are asked to in surveys such as those cited above. Moreover, the recent proliferation of both positive and normative theories attempting to explain why preferences may change as the framing of the question changes, suggests that preference reversals may to some extent be considered 'consistent' [7,8]. More serious are inconsistencies that arise as a result of respondent confusion with the central task being asked of them, or a lack of understanding or misinterpretation of the instructions. Inconsistencies of this kind will, in situations where respondents are asked to value a number of different health states, be evidenced by preference rankings that fail to conform to *a priori* expectations. It is therefore important to distinguish between 'primary' inconsistencies that arise largely from the intrinsic limitations of respondents and 'secondary' inconsistencies that result largely as a consequence of some aspect of the measurement procedure.

The extent of inconsistencies due to 'primary' factors is largely unknown. Much of the potential evidence on this matter is 'lost' through the censoring of empirical data so that analysis is carried out only *after* respondents whose rankings fail to conform with those expected by the researcher have been excluded. For example, Bush *et al.* (1977) arbitrarily excluded respondents who rated health states inconsistently with the instructions given to them, resulting in the loss of data from 11 out of 65 respondents [9]. Sackett and Torrance (1978) considered data to be 'useable' only if respondents recorded valuations for *all* scenarios and these were internally consistent. Of the 112 completed interviews, only 87 (78%) yielded 'useable' data, the remaining 22% of respondents "were eliminated from the data" [10]. Patrick *et al.* (1985) go further and exclude respondents "with scores that were widely disparate from the group mean" and others were excluded because "their instructor rated their understanding of the task as inadequate" [11].

Using such criteria to exclude respondents may have important implications for the choice of descriptive system, valuation method, mode of administration, and ultimately, for the valuations themselves. Only by looking more closely at those respondents whose implied ranking of health states violates *a priori* expectations and analysing why these 'primary' inconsistencies occur in the first place will our understanding of health status measurement be enhanced. This provides the motivation of this paper. It focuses on *uncensored* data and in so doing addresses issues that have been largely ignored by researchers in the health status measurement field.

METHODS

The EuroQol instrument

The analysis in this paper uses data from two studies undertaken by the Measurement and Valuation of Health (MVH) Group at the Centre for Health Economics, University of York. The Group has been actively involved in the development of the EuroQol classification, a generic instrument for describing and valuing health states [12]. It is intended to be used alongside condition-specific measures of health-related quality-of-life (HRQoL) to provide information for use in clinical and economic evaluation. As part of that developmental work, the MVH Group set up a series of studies to examine specific methodological issues. Although the primary objective of each study differed, valuations for EuroQoL health states were elicited in all of them. Since it is the analysis of these valuations data which forms the basis of the current paper, the study protocols are described in outline only.

The LC study

The first data set (LC) was obtained as part of a wider study of how the general public view health and illness and what to them constitute the salient features of HRQoL. Self-completion questionnaires were administered during face-to-face interviews conducted with a quota sample of disabled adults aged 18 to 25 and their carers, recruited through the Physically Handicapped and Able Bodied (PHAB) organisation. For the purposes of this study, there was no operational definition of disabled: membership of PHAB was a sufficient condition for recruitment. A corresponding group of individuals matched by age and sex, together with a sample drawn randomly from the community (identified using electoral registers), were also selected — thus yielding a total of 5 subgroups. The fieldwork was carried out in Dudley, Walsall and Wolverhampton during 1988. The questionnaire utilised the original 6-dimensional form of the EuroQol classification which describes health in six dimensions (mobility, self-care, main activities, leisure activities, pain/discomfort and anxiety/depression). Each dimension could be experienced at one of either two or three levels: broadly speaking 'no problems' (level 1), 'some problems' (level 2) or 'extreme problems' (level 3). Health states are described by combining one statement from each of these dimensions to produce a state of health identified by a 6-digit code, e.g. 211111 describes the state in which a person is unable to walk without a stick, crutch or walking frame but has no problems on any of the other dimensions.

The F4 study

The second data set (F4) consisted of responses to a postal survey of patients registered with a large general practice in Frome, Somerset. The study, carried out in 1991, was designed to address a number of methodological issues including the relative importance attached to energy and tiredness and the effect of changing the way in which EuroQol states are described, hence a number of different questionnaire variants were administered. The computerised practice list only provided a means of identifying individuals by name and address and no information bearing on patients' clinical record was sought or collected. Each selected patient, chosen at random from the practice list, received a copy of one of the questionnaire variants together with an introductory letter from the senior partner in the practice. A reply-paid envelope was also included to facilitate the return of completed questionnaires. The questionnaire was similar to that used in the LC study but utilised a revised 5-dimensional version of the EuroQol classification which combines main and leisure activities into usual activities and makes all dimensions comprise three levels. The differences *within* F4 were that two of the questionnaires (D and E) had an additional energy/tiredness dimension and level 2 of the mobility dimension was described as "unable to walk without a stick, crutch or walking frame" in questionnaires A,B and D whilst in C and E it was described as "some problems in walking about".

The questionnaires

A standard format is used in *all* forms of the EuroQol questionnaire and is based on two main sections. The first of these is designed to record self-rated health status. This is achieved by asking respondents to describe their own health by indicating their level of problem on each of the EuroQol dimensions, and to rate their own health on a 20cm visual analogue scale (VAS) with 100 ("best imaginable health state") and 0 ("worst imaginable health state") as endpoints. The second section of the questionnaire is concerned with the valuation of a standard set of 16 composite EuroQol states using an identical VAS. This set covers a broad range of health states and always contains the logically 'best' health state (with no problems on any dimension) and logically 'worst' health state (with extreme problems on all dimensions). The order in which states are presented is constant within each variant of the questionnaire. Respondents are asked to imagine themselves being in each health state for one year (they are told that what happens thereafter is not known and should not be taken into account) and to draw a line from the box containing the state to the point on the VAS which indicates how good or bad they consider the state is. Finally, respondents are asked a number of background questions relating to their age, sex, experience of illness and educational attainment.

The analysis

The two data sets present the opportunity to assess whether inconsistency rates are a function of (a) the descriptive system used, (b) the mode of administration, and/or (c) different respondent characteristics. Each group of respondents in the LC data was analysed separately since the five subsamples differed in respect of their background characteristics, not least in their experience of illness. The F4 study used five variants of the EuroQol questionnaire and since each questionnaire differed slightly, responses were not pooled. Because the distribution of inconsistency rates was highly skewed (as evidenced by the results of Kolmogorov-Smirnov tests), the non-parametric Mann-Whitney U Test was used to test for differences in inconsistency rates between questionnaires and respondents (significance level $P<0.01$).

Defining inconsistency

By definition discrete dimensions in both variants of the EuroQol descriptive system constitute ordinal scales in which $\text{level}_{i+1} < \text{level}_i$; for example having some problems with self-care (level 2) is worse than having no problems with self care (level 1). Health states are formed by combining elements from each dimension and these 'composite' descriptions may also be ordered according to their inherent ordinality. For any subset of such composite states it follows that there is an expected ordinal relationship between some, but not all, pairs of states. For example, if state A is formed by combining levels 13221 respectively on each of the 5 (revised) dimensions, and state B is similarly formed by combining levels 12221, then it follows that state A is logically worse than state B, since for each dimension in state A the level is equal to or worse than the corresponding level for state B. Contrariwise, the relationship between state A and some other state, C, say, 22132, cannot be logically predetermined since some levels in C are worse and some are better than for the corresponding dimensions in A. For a respondent to meet perfectly the assumption of ordinality, the value they give to state A should be lower than the value given to state B when state B is 'logically' better on at least one dimension and no worse on the other dimensions.

Only certain pairs of the EuroQol states used in the two surveys stand in a logically defined ordinal relationship. There are 83 such pairs for the LC data set and 75 for the F4 data set. For each respondent it is possible to calculate the number of times an expected logically inconsistent ranking occurs, and hence to calculate an inconsistency rate (expressed as a percentage) using as the denominator the maximum possible number of such logical pairings. This statistic is comparable to the coefficient of inconsistency described by Kendall with regard to paired comparisons data [13].

RESULTS

Response rates

Table 1 shows the response rates to the LC and F4 studies together with the key background characteristics of respondents. Of the 591 addresses issued to interviewers as part of the LC study, 488 (83%) yielded an interview. As might be expected the response rate to the postal F4 study was somewhat lower; 462 questionnaires were returned out of an initial mailing of 1320, giving a response rate of 35%. Overall, respondents in the LC study are younger than those in the F4 study and are predominantly female. A potential source of response bias is seen in the general population sub-group of the LC study and in the F4 study, where 17% and 22% respectively have gained higher education

Table 1. Response rates and background characteristics

Study	Total sample	Number achieved	Response rate (%)	Mean age (%)	Female (%)	Higher education (%)	Personal experience of illness
LC							
General population	196	163	83	42	49	17	23
Disabled group	95	81	95	20	35	10	67
Disabled control	92	69	75	20	51	15	11
Carer group	98	80	82	47	93	8	24
Carer control	110	95	86	44	94	8	23
Overall	591	488	83	37	62	12	28
F4							
A	270	95	35	48	34	26	26
B	270	96	36	52	36	22	29
C	270	96	36	50	48	23	30
D	240	80	33	49	57	22	26
E	270	95	35	52	58	17	32
Overall	1320	462	35	50	46	22	29

qualifications (compared to 8% of the population as a whole). As was to be expected, the disabled sub-group in the LC study has the highest proportion of respondents with personal experience of illness.

Distribution of inconsistency rates

Table 2 shows the mean and median inconsistency rates for each subsample in the two studies. Although both these measures of central tendency are similar for the LC subsamples, mean inconsistency rates are much higher than median inconsistency rates for all of the F4 subsamples. This suggests that in the F4 study a few respondents with very high inconsistency rates are biasing the mean upwards and that the distribution of inconsistency rates are much more skewed in the F4 subsamples than in the LC ones. Whilst no respondent is completely consistent in the LC study, one-quarter of all respondents to F4 have no logically inconsistent rankings and three-quarters have an inconsistency rate below 10% (this compares with 43% of respondents in the LC study). However, there are a greater number of respondents in the F4 study with inconsistency rates above 50%; 22 (4.8%) compared to 7 (1.4%) in the LC study.

Differences in inconsistency rates between LC and F4

Table 2 shows that median inconsistency rates for the LC subsamples are significantly higher than those for the F4 subsamples. This includes the random sample of the general population in the LC study who, despite being drawn from a very different geographical location, are similar in terms of key background characteristics to the respondents to the F4 study (see Table 1).

The influence of background characteristics

From Table 2 it can be seen that the disabled control group of respondents in the LC study (i.e. young, fit people who may not care or think much about illness) have higher inconsistency rates than the other groups, particularly the disabled group who are comparable in terms of age, sex and level of education. This suggests that those without experience of illness may be more inconsistent than those with such experience. However, when *all*

Table 2. Percentage inconsistency rates for the LC and F4 studies

	n	Mean	(SD)	Median	(IQR)
LC					
General population	163	13.2	(9.3)	10.8	(8.4–14.5)
Disabled group	81	11.6	(6.6)	9.6	(7.2–12.7)
Disabled control	69	19.9	(16.6)	15.6	(9.6–19.1)
Carer group	80	11.4	(4.7)	10.4	(8.4–14.2)
Carer control	95	13.0	(6.7)	12.0	(8.4–16.9)
Overall	488	13.5	(9.6)	10.8	(8.4–15.7)
F4					
A	95	10.9	(19.9)	2.7	(1.3–9.3)
B	96	8.6	(17.1)	2.7	(0–9.3)
C	96	9.8	(19.1)	2.7	(0–7.7)
D	80	9.8	(17.2)	2.7	(1.3–10.3)
E	95	6.7	(12.3)	2.7	(0–6.7)
Overall	462	9.3	(12.2)	2.7	(0–8)

All LC subsamples have significantly higher inconsistency rates than all F4 subsamples ($P < 0.01$).
The disabled control group have significantly higher inconsistency rates than all other groups in the LC study ($P < 0.01$).

respondents are categorised according to their current and past illness experience, Mann-Whitney U tests reveal no significant differences, suggesting that experience of illness is *not* closely linked to inconsistency rates. There are also no significant differences in inconsistency rates relating to the sex or the smoking behaviour of the respondents.

However, inconsistency rates *do* appear to be affected by the age and educational attainment of the respondent, although for the general population subsample of LC there is no significant difference between those aged under 60 and those aged 60 or over in terms of their inconsistency rates. Older respondents in every F4 subsample have significantly higher rates of inconsistency. With respect to educational attainment, those with minimum education have significantly higher inconsistency rates than those with further education in all F4 subsamples. The LC data follows a similar pattern but, although median inconsistency rates are generally higher for the less educated, this difference is not statistically significant.

DISCUSSION

The 83% response rate to the interview-based LC study is very acceptable whilst a 35% response to the postal F4 study might be considered disappointing. However, it must be borne in mind that the F4 sample population was the general public who were sent no follow-up reminders. Local Government elections tend to attract small numbers of participants, so that though these rates are low, they are not altogether out of keeping with the proportion of voters who actively express their opinions on local issues. However, the fact that respondents to postal studies appear to be much better educated than the general population suggests that researchers using this mode of administration to elicit health state valuations should exercise caution in the interpretation of their results.

The starting point for this paper relies upon a definition of inconsistency that is predicated on a principal assumption that an ordinal structure exists within discrete EuroQol dimensions. This results in a unique logical ordering of *some* health states, but leaves indeterminate the ordering of other states. The definition of inconsistency adopted in this paper gives the same weight to *all* pairs of states in the logically predetermined group. For example, an inconsistency involving states 11121 and 11122 is accorded the same weight as an inconsistency involving states 11121 and 33333. This approach could be modified to take account of the proximity of health states using an appropriate 'distance' function where greater weight would be given to inconsistencies involving states that are more clearly separated by the ordinal structure. In the LC and F4 studies, the descriptive systems differed both in terms of the number of dimensions *and* the number of levels within dimensions. Thus, an approach along these lines proved untenable. However, studies using a single standard descriptive system could present results in which 'distances' between states are accounted for.

Overall, the inconsistency rates associated with using the visual analogue scale to value EuroQol health states are encouraging. Median inconsistency rates of around 10% are obtained from the LC study whilst every subsample of the F4 study produces median rates below 3%. Closer analysis of the health state scores given by the 22 respondents in the F4 study who had inconsistency rates greater than 50% shows that 11 give valuations that are clustered around the two endpoints of the scale (i.e. close to, or at, 100 and 0). This finding may indicate a misinterpretation of the labels "best imaginable health state" and "worst imaginable health state" since it is possible that these 11 respondents thought they were to locate the states on the scale according to how well they could actually *imagine* being in them [14]. In contrast, none of the 7 respondents in the LC study with inconsistency rates above 50% appear to be interpreting the instructions in this way.

That inconsistency rates were found to be higher in the LC study than in the F4 one suggests that the 6-D descriptive system used in the LC study may be more prone to rankings that violate the logical ordering than the revised 5-D one used in the F4 study. However, two of the F4 questionnaire variants contained an extra dimension, effectively making these 6-D descriptive systems too, and there are no significant differences in the inconsistency rates of respondents to these questionnaires compared to the others in F4. Therefore, observed differences in the median inconsistency rates observed appear to be explained by factors other than the descriptive system alone.

The different median inconsistency rates of the LC and F4 subsamples seem more likely to be explained in terms of the different ways in which the studies were conducted; the LC study was interview-based whilst F4 was a postal survey. In the F4 study, if potential respondents, for whatever reasons, experienced difficulties with the questionnaire, they were not obliged to reply. Furthermore, no reminders were sent out after the original mailing. Thus, there must be an inevitable response bias with those returning their questionnaires being a self-selected group of respondents who understood (or at least thought they understood) the questionnaire. This is evidenced by response bias in favour of more educated respondents in the F4 study. Of course, there is the possibility of response bias to interview-based studies such as LC but there would seem to be less scope for refusal under such circumstances, borne out by the much higher response rate in the LC study than the F4 one.

More generally, it seems entirely plausible that the motivation, attention to detail, and level of performance of respondents may be influenced by the settings in which they complete the questionnaire. However, there is little evidence on whether valuations of own health or of hypothetical states differ according to how the questionnaire is administered, or, as suggested by the results presented here, whether inconsistency rates differ across different modes of administration. This is an area that would benefit greatly from systematic investigation in the future.

With respect to the factors influencing inconsistency, it appears that rates are positively related to age and negatively related to educational attainment. These results are intuitively appealing since older and/or less educated respondents may be more likely to encounter difficulties in interpreting the composite health states and in comprehending the task required of them. However, whilst they confirm *a priori* expectations, the results have implications for studies that exclude respondents with high rates of 'primary' inconsistency from subsequent analysis; they suggest that these studies are biasing further a sample which may already be biased in terms of socio-demographic characteristics.

Of course, such issues are less important if it is known that older or less educated respondents have the same preference structure as the rest of the population. Although the limited evidence that does exist points in this direction, it is far from being an established fact. Indeed, the very notion that the preference weights of those most affected by a policy or intervention should be incorporated into the decision-making process may be compromised if the views of such people are excluded because it is known they have high rate of 'primary' inconsistency. Moreover, the existence of a high rate of 'primary' inconsistency may be used by some to discount the views of such people or as justification for using the preferences of surrogates free of such 'limitations' as proxies for the values of the people whose views *really* count. In turn, this raises the issue of the relative weights to be placed on the *need* for consistent responses and the objective of including the preferences of *all* subgroups.

The presence of inconsistent valuations data in an individual's responses is entirely compatible with the concept of imperfect judgements as described in psychological and measurement theories [15,16]. The existence of inconsistency with such valuations data might too be interpreted as evidence of the multidimensional character of the health states which people are asked to judge. Although respondents may have been able to arrive at judgements of complex health state descriptions, they may not be able to represent them consistently in a unidimensional form.

In this paper, valuations for states that fail to satisfy the ordinality conditions have been considered to be inconsistent. However, if a respondent *genuinely* prefers one (logically worse) state to another then it may be wrong to class this as an inconsistent judgement. Such apparent dissonance between the underlying logical structure and individual preferences could also result from other factors, such as ambiguity in the descriptive system. Whilst it is difficult to see how 'some' problems could be interpreted as worse than 'extreme' problems in the EuroQol classification system, ambiguity in description (for example, between terms such as 'mild' and 'moderate') or a complex choice of language (as in some health scenarios) may explain some of the inconsistencies found in other studies. Of course, excluding this group of 'inconsistent' respondents from subsequent analysis may be even more problematic, particularly if ambiguity of meaning is itself a function of one or more background characteristics. Future studies designed to address the issue of inconsistent judgements could mitigate against such problems by examining the *internal* structure of individual preferences. This could be achieved, for example, by comparing the rankings of health states implied from the valuations from different methods; something not possible given the design of the LC and F4 studies.

A recent study which used the same visual analogue scale method to value composite health states reported significantly higher inconsistency rates for those with a low self-rated health status and those with personal past experience of illness [17]. Such results were not reproduced in the study reported here although the results are not strictly comparable because the Kind *et al.* [17] study examined valuations based on the Rosser Classification of Illness States. This suggests that the factors that influence respondent inconsistency may be in part a function of the descriptive system.

The lowest median level of inconsistency found in the LC and F4 datasets was about 3% which suggests that an intrinsic residual inconsistency of this order could be anticipated in any study which uses the EuroQol descriptive system and/or the visual analogue scale valuation method. Of course, it is impossible to determine whether the descriptive system or the valuation technique has the most important role to play here but a residual level of inconsistency of this order appears to be quite robust; a recent study using the same 5-dimension EuroQol descriptive system reported comparable levels of inconsistency using the visual analogue scale [18]. The same study reported significantly higher rates (10–15%) for both the standard gamble and time trade-off techniques. Such findings may suggest that a trade-off exists between the theoretical advantages of 'choice-based' methods and their practical disadvantages in terms of ease of completion. Such a discussion lies outside the scope of this paper but the essential point is that the lack of unambiguous results leaves open the question of the origin of inconsistency in health state valuation. It seems likely to be determined by the 'primary' and 'secondary' factors alluded to in the introduction but the balance between the two remains indeterminate.

'Inconsistent' respondents have been reported in other studies but often only as a justification for their exclusion from subsequent analysis. Only by a more detailed analysis of what causes some respondents to rank health states in ways which violate *a priori* expectations will our understanding of health status measurement be enhanced. Whilst it has been difficult given the design of the LC and F4 studies to separate out the effects of descriptive system, valuation method, mode of administration and respondent characteristics on inconsistency rates, this paper has provided a framework in which other studies can report inconsistency, and provides a benchmark against which inconsistency rates can be compared. Given the increasingly widespread use of health state valuations in clinical and economic evaluation, this is a challenge to which other researchers *must* respond.

Acknowledgements—The support of the Department of Health, via the Centre for Health Economics at the University of York, is gratefully acknowledged. The authors would like to thank Professor Alan Williams and Dr Claire Gudex at the Centre for Health Economics for their helpful comments on earlier drafts of this paper. The comments of two anonymous referees are also gratefully acknowledged.

REFERENCES

1. Torrance G. W. Measurement of health state utilities for economic appraisal. *J. hlth Econ.* **5**, 1, 1986.
2. Froberg D. G. and Kane R. L. Methodology for measuring health state preferences III: population and context effects. *J. clin. Epidemiol.* **6**, 585, 1989.
3. Llewellyn-Thomas H., Sutherland H. J., Tibshirani R., Ciampi A., Till J. E. and Boyd N. F. Describing health states: methodologic issues in obtaining values for health states. *Med. Care* **22**, 543, 1984.
4. Llewellyn-Thomas H., Sutherland H. J., Tibshirani R., Ciampi A., Till J. E. and Boyd N. F. The measurement of patients' values in medicine. *Med. Care* **17**, 501, 1982.
5. Kahneman D. and Tversky A. The framing of decisions and the psychology of choice. *Science* **211**, 453, 1981.
6. McNeil B. J., Pauker S. G., Sox H. C. and Tversky A. On the elicitation of preferences for alternative therapies. *N. Engl. J. Med.* **306**, 1259, 1982.
7. Kahneman D. and Tversky A. The psychology of preferences. *Scient. Am.* **246**, 160, 1982.
8. Machina M. J. 'Expected utility' analysis without the independence axiom. *Econometrica* **50**, 277, 1982.
9. Bush J. W., Kaplan R. M. and Berry C. C. Comparison of methods for measuring social preferences for a health state index. *American Statistical Association Annual Meeting*, p. 682, 1977.
10. Sackett D. L. and Torrance G. W. The utility of different health states as perceived by the general public. *J. chron. Dis.* **31**, 697, 1978.
11. Patrick D. L., Sittampalam Y., Somerville S. M., Carter W. B. and Bergner M. A cross-cultural comparison of health status values. *Am. J. publ. hlth* **75**, 1402, 1985.
12. The EuroQol Group. EuroQol — a new facility for the measurement of health-related quality of life. *Hlth Pol.* **16**, 199, 1990.
13. Kendall M. G. *Rank correlation methods*, 3rd edn, p. 146. Griffin, London, 1962.
14. van Busscbach J., Hessing D. and de Charro F. Observations on one hundred students filling in the EuroQol questionnaire. *EuroQol Conference Proceedings* **2**, 55, Helsinki, October 1992.
15. Thurstone L. L. A law of comparative judgement. *Psychol. Rev.* **34**, 273, 1927.
16. Coombs P. Psychological scaling without a unit of measurement. *Psychol. Rev.* **57**, 147, 1950.
17. Kind P., Dolan P., Gudex C. M. and Williams A. H. Inconsistency and the judgement of health state valuation: a comparison of three scaling methods. IRSS workshop, University of York, October 1992.
18. Dolan P., Gudex C. M., Kind P. and Williams A. H. Valuing health states: a comparison of methods. HESG conference, Strathclyde, July 1993