Current state of the art in preference-based measures of health and avenues for further research

Brazier, J
The University of Sheffield

December 2005

HEDS Discussion Paper 05/05


ScHARR
School of Health and
Related Research

The University of Sheffield

ScHARR
School of Health and Related Research

Health Economics and Decision Science

Discussion Paper Series

December 2005
Ref: 05/5

CURRENT STATE OF THE ART IN PREFERENCE-BASED MEASURES OF HEALTH AND AVENUES FOR FURTHER RESEARCH

John Brazier

Corresponding Author:
John Brazier
Health Economics and Decision Science
School of Health and Related Research
University of Sheffield
Regent Court, Sheffield, UK
S1 4DA
Email: j.brazier@sheffield.ac.uk

INTRODUCTION

Preference-based measures of health (PBMH) have been developed primarily for use in economic evaluation. They have two components, a standardized, multidimensional system for classifying health states and a set of preference weights or scores (1) that generate a single index score for each health state defined by the classification, where full health is one and zero is equivalent to death. A health state can have a score of less than zero if regarded as worse than being dead. These PMBH can be distinguished from non-preference-based measures by the way the scoring algorithms have been developed, in that they are estimated from the values people place on different aspects of health rather than a simple summative scoring procedure or weights obtained from techniques based on item response patterns (e.g., factor analysis or Rasch analysis).

The use of PBMH has grown considerably over the last decade with the increasing use of economic evaluation to inform health policy, for example through the establishment of bodies such as the National Institute for Clinical Excellence in England and Wales (2) and the Health Technology Board in Scotland (3) and similar agencies in Australia (4) and Canada (5). Preference-based measures have become a common means of generating health state values for calculating quality-adjusted life years (QALY). The status of PBMH was considerably enhanced by the recommendations of the U.S. Public Health Service Panel on Cost-Effectiveness in Health and Medicine to use them in economic evaluation (6). A key requirement for PBHM in economic evaluation is that they allow comparison across programs.

While PBMH have been developed primarily for use in economic evaluation, they have also been used to measure health in populations. PBHM provide a better means than a profile measure of determining whether there has been an overall improvement in self-perceived health. The preference-based nature of their scoring algorithms also offers an advantage over non-preference-based measures since the overall summary score reflects what is important to the general population. A non-preference-based measure does not provide an indication to policy makers of the overall importance of health differences between groups or of changes over time**.**

The purpose of this paper is to critically review methods of designing preference-based measures. The paper begins by reviewing approaches to deriving preference weights for PBMH, and this is followed by a brief description and comparison of five common PBMH. The main part of the paper then critically reviews the core components of these measures, namely the classifications for describing health states, the source of their values, and the methods for estimating the scoring algorithm. The final section proposes future research priorities for this field.

APPROACHES TO OBTAINING PREFERENCE WEIGHTS FOR MEASURES OF HEALTH

There are three empirical approaches to deriving preference weights for measures of health: i) empirical mapping onto an existing PBMH, ii) mapping onto a respondent's own health valuation, and iii) asking respondents to value states defined by the health measure. While existing preference-based measures are currently based on the third approach, it is important to understand the potential role of the alternatives.

Mapping between measures

The approach of empirically mapping a health measure onto a PBMH so as to obtain preference weights for the former has been used in numerous studies. This approach requires the PBMH and the non-preference-based measure to be administered to the same population. The approach was used by Fryback et al. (7) in the Beaver Dam study and Nichol et al. (8), both of whom mapped the generic SF-36 onto preference-based measures. Tsuchiya et al (9) and Brazier et al (10) have mapped PBHMs onto condition-specific measures. These studies have regressed the dimension scores of the non-preference-based measure onto the preference-based measure.

This approach can be pragmatically useful, but it makes several assumptions. First, it assumes that the items used to score the dimensions have equal importance, and second, that the intervals between the response choices are equally important to people. These assumptions can be relaxed by modelling each item response as a dummy variable (9). A more important limitation is the assumption that the preference-based measure covers all the important aspects of health covered by the non-preference-based measure. If it does not, important dimensions of health might not be valued appropriately. Mapping onto an existing PBMH should be viewed as a second best compared to the direct valuation of the health measure.

Valuing a health measure using direct values for health states

This approach involves administering the health measure alongside a valuation question about the respondent's own health. Such an approach was taken by Lundberg and colleagues (11) and involved administering the SF-12 health status questionnaire alongside a self-administered version of the time-trade-off (TTO) technique in a postal survey of 8,000 members of the general population. The TTO question asks respondents to consider the number of years of life they would be willing to sacrifice in order to live the remainder of their life in full health. SF-12 item responses (among other variables such as age) were regressed against their TTO value to provide a set of preference weights for the SF-12.

This study was limited because the health states were those of a random sample of the general population. The sample of states valued in the survey was therefore not determined by any statistical design but by their natural occurrence in the population. As many severe states are quite rare, this reduces the ability of the model to predict values for more severe states. These limitations could be partly overcome by careful sampling of people with a wide range of conditions, but there will always be people in some medical conditions that could not participate in such a survey, necessitating the use of proxies.

In conclusion, this approach captures only the values of those people in the states. This may be seen as a problem for those seeking to implement the Washington Panel's recommendations to use 'societal values'. The issue of *whose* values is examined latter in this paper.

Hypothetical valuation of health states

The approach adopted by PBMH developers to date involves asking respondents (typically adult members of the general population) to imagine being in a health state. They could be asked to imagine the state on someone else's behalf, perhaps as a proxy or by taking a third-party perspective (such as behind a 'veil of ignorance'), but most

applications ask respondents to imagine they are in the state. This has a logistical advantage over the previous direct approach because it allows a single respondent to value numerous states and the researcher can select the states being valued according to a proper statistical design. This approach conforms most closely to the recommendations of the Washington Panel.

EXISTING PREFERENCE-BASED MEASURES OF HEALTH

The five preference-based measures considered in this section are the Quality of Well-Being (QWB) scale (12), the Health Utility Index (HUI) versions two and three (HUI-2 and HUI-3) (13,14), EQ-5D (15,16), and the SF-6D - a derivative of the SF-36 and SF-12 (17,18). These instruments were chosen because they are the most widely used. In Table I, which summarizes their characteristics, we see that existing preference-based measures differ in the content of their descriptive systems, the valuation technique, and the method of extrapolation. And yet their descriptive systems share a common structure, as they all have multilevel dimensions. Furthermore, despite differences in methods of valuation, all preference data were obtained from a sample of the general population (although the HUI-2 was valued by a random sample of parents from Hamilton, Ontario).

All five instruments purport to be generic. Even so, they differ in defining health and in the dimensions they cover. The developers of the HUI instruments restrict health to 'beneath the skin' and exclude those consequences for quality of life (QOL) affecting the person's functioning in society, such as role and social functioning. By contrast, the QWB, SF-6D, and EQ-5D include 'out-of-skin' aspects of health. The HUI-2 was designed for children and the rest for adults. The instruments also differ greatly in size, with between five and eight dimensions and two to six levels on each dimension, resulting in the number of potential health states ranging from 243 for the EQ-5D to 972,000 for the HUI-3.

The instruments use different valuation techniques to elicit preferences, including the visual analogue scale (VAS), standard gamble (SG), and TTO (see definitions in next section). The HUI-2 and HUI-3 health states were not directly valued using SG but rather via a transformation of VAS values. The five measures were also valued using different variants of the valuation techniques. Finally, all five measures have too many states to be valued in one survey (with the possible exception of EQ-5D), and thus values must be extrapolated from a sample of states defined by their respective classifications. This was done for three of the measures by statistical modelling (EQ-5D, QWB, and SF-6D), and the HUI-2 and -3 use multi-attribute utility theory (MAUT). Valuation work has been replicated across a number of countries for the EQ-5D, SF-6D and HUI-2&3.

Comparison of measures

This section briefly compares the five PBMH in terms of practicality, reliability, and degree of agreement.

There is little to choose between the instruments on the basis of practicality, with each now having self-completed versions. The EQ-5D has just five questions, but it is closely followed by the 15 items for the HUIs and 36 (or just the 11 used to construct

the classification) for the SF-6D. Stability over time in those whose health has not changed has been demonstrated for most instruments, and there is no reason to suppose them to be worse than non-preference-based measures (19). The standard deviations of the preference scores, however, are larger for the EQ-5D and HUI-3 than the SF-6D (20-22), because the EQ-5D and HUI-3 cover a larger range of states.

Content validity depends on the aspects of health the user wishes to cover and also on the disease group and age of the patients. There are also issues about perspective and whether social health is relevant. It is also evident from the descriptions that the SF-6D may suffer from 'floor effects' on dimensions of physical functioning and role limitation, where many respondents choose the lowest response category, and this has been borne out in recent empirical comparisons (20-22). Conversely, concern has been expressed that the EQ-5D suffers from a ceiling effect because large numbers of patients are given states of full health (23).

Among those who use a choice-based valuation technique, the HUI-2 and -3 and SF-6D might be preferred to the EQ-5D by those who regard the SG as the 'gold standard' and EQ-5D by those who prefer TTO. The SG utilities for the HUIs have been derived from VAS values using a power transformation that has been criticized (see above). The valuation of the HUI has also been obtained from a smaller and less representative sample of the general population than the valuation survey of the EQ-5D in the UK and US. A further difference between them is the methods of modelling the values for health states, with the HUIs using MAUT and the rest using statistical inference. The review in the next section concludes there is little evidence on their relative predictive performance.

The measures seem to be moderately correlated at around 0.5 to 0.6, and the differences in mean values for patient groups are often just 0.05 on the zero-to-one utility scale, but this makes for significant systematic variation between instruments (20,21). Comparisons of the SF-6D with the EQ-5D and HUI-3 have shown the former to generate higher values for more severe states and lower values for the mildest states. Furthermore, these cross-sectional differences were found to translate into significant differences in the size of change measured over time in one patient group (20). The variation is a product of the differences between the instruments, but it is not possible to determine whether this variation is driven by differences in their descriptive systems, techniques of valuation, or methods of extrapolation.

The next section critically reviews the three components of PBMH: the descriptive systems, valuation techniques, and methods of extrapolation.

REVIEW OF METHODS

Descriptive systems

Two important issues in designing a descriptive system for a generic preference-based measure are (i) the definition of health and (ii) the construction of descriptive systems for PBMH.

Definition of health

The developers of the HUI advocate a 'beneath the skin' definition of health that excludes social activities and work (e.g., social and role dimensions on the SF-6D and

usual activities in the EQ-5D). Social and role activities are deemed the result of the personal preferences of the respondent as well as his/her state of health and it has been argued should be excluded from the descriptive system. For the HUI measures, this also helps to achieve orthogonality or independence between attributes in the classification of health state, which is important for the application of MAUT and, to a lesser extent, the statistical approaches.

There is a long history of having 'out of skin' consequences in measures of health, due in part to the original WHO (World Health Organization) definition that included social health (24). It could also be argued that the impact on role and social activities is important in helping respondents in a valuation survey to fully understand the impact of a health state on their QOL. It effectively reduces the imaginative workload being demanded of respondents.

This debate has interesting parallels in the QOL literature in general through the notion of 'response shift' (25), where the impact of a health state on a person's QOL is not stable. A sudden change in health may initially have a substantial impact, but gradually people learn to cope and adapt to their limitations in a number of ways, and over time the impact lessens. Those measures having role and social dimensions are likely to be even more prone to this 'response shift'. Whether adaptation should be excluded from the final values given to states is a normative question addressed in a later section.

Most preference-based measures of health have a generic descriptive system (14,17,26). These have been found, however, to be inappropriate or insensitive for many medical conditions (e.g., 27-29). Condition-specific descriptions may be more sensitive to changes in the condition than generic measures and more relevant to the concerns of patients (30). There is a concern, however, that condition-specific PBMH fail to achieve comparability. Differences in scores between measures (whether generic or condition specific) result from differences in the methods of valuation as well as the descriptive system. In principle, if the descriptive system is valued on the same full health–dead scale using the same variant of the same valuation technique employing a comparable population sample, the valuations should be comparable. Any remaining differences in values should be a legitimate consequence of the descriptive system. This, however, assumes that the value of a dimension is independent of those dimensions outside of the descriptive system, and this requires empirical testing.

Construction and validation of descriptive systems for PMBH

Although there are difficulties in establishing the validity of the values generated by a preference-based measure (as discussed below), it is important to show that a descriptive system accurately describes the health state. Little has been written in the economics literature about this aspect of validity (31). Published economic evaluations rarely address the issue even though small differences in descriptions of health states can substantially alter the results. There has been some criticism expressed of specific generic preference-based measures but little systematic evaluation of their descriptive systems. The assessment of the validity of the descriptive system of a preference-based measure should be undertaken with the same

rigor as would be applied to non-preference-based measures. There are some important differences of principle, however, and these are explained below.

Content and face validity

The psychometric criteria of content and face validity, although subjective, are important in assessing the comprehensiveness, relevance, and sensitivity of the dimensions of preference-based measures. Content, in terms of dimensions and items, limits the attributes being covered by the measure. Economists, who are concerned with ensuring that the measure correctly reflects a person's utility function, will prefer an approach which generates items and dimensions from patients. This has been pursued to a limited extent in the development of some existing instruments, but most preference-based measures were constructed primarily by teams of experts.

Item Response theory

IRT provides a powerful technique for understanding the relationship between items. It is based on the assumption that item responses are determined by their degree of difficulty along some spectrum of a unidimensional construct. This has been very helpful in understanding the relationship between items and has been used, among other purposes, to assist in selecting items for an instrument and in scoring the items. Recent years have seen a rapid expansion of the application of IRT to the construction and testing of non-preference-based measures.

IRT has potential in helping to construct the descriptive systems of preference-based measures. It was used in the selection of items for the physical functioning dimension of the SF-6D from the 10 items of this dimension in SF-36 (17). By pooling items from different PBMH and non-preference-based measures, IRT can assist in understanding the severity range covered by the descriptive systems of the preference-based measures. IRT may identify significant gaps in the descriptive classifications and provide candidate items for improving them.

It is important to understand the limitations of the IRT, however. It would make no sense to apply IRT across dimensions for a preference-based measure, anymore than it would for a non-preference-based measure. Furthermore, while the advocates of IRT claim it generates an interval scale and this might be true for the health concept being measured, it does not provide a measure of strength of preference with the required interval properties.

Construct validity

Having an empirically based means of testing the descriptive validity of an instrument is important. Construct validation is appropriate for testing the validity of the description of health or health change underlying a PBMH. The ability of an instrument to reflect known or expected differences in health is an essential precursor to its ability to reflect preferences. Such tests should be undertaken on the unscored descriptions of an instrument, however. Otherwise, there is a danger that the failure of a score to detect a difference found by a condition-specific QOL measure is incorrectly interpreted to imply that the descriptive component of the preference-based instrument is insensitive. The score of a preference-based measure may fail to detect the difference simply because this difference is not valued by patients.

Valuation

The methods of valuation underpin preference-based measures and have been a topic of considerable debate in the health economics literature. The key areas of concern have been the choice of technique, the variant of the technique, the problem of states worse than death, and the appropriate sources of values.

Valuation Technique

The five preference-based measures reviewed in this paper have used VAS, SG, or TTO, and much has been written on the relative virtues and flaws in each of these techniques. This section attempts to summarize this well-trodden path.

Several informative reviews have compared the VAS, SG, and TTO techniques in terms of practicality, reliability, and validity (32-36). Generally, all three techniques have been reported to be practical and acceptable for most populations, but VAS is considered marginally better in terms of response rate, cost, and consistency of responses (33).. On the other hand, the validity of VAS as a measure of the strength of preference has been challenged (37-39). The main criticism is that a rating scale does not confront the respondent with the notion of opportunity cost and so does not reflect the economist's notion of strength of preference. Interviews with respondents indicate that they did not intend it to reflect their preferences (37,39,40). When asked, respondents talk about concepts of fitness or the natural history of illness and not the value of health.

SG and TTO present respondents with a choice and offer a more theoretically appealing measure of strength of preference (i.e., involving opportunity cost). SG asks respondents to make a choice between alternative outcomes where one of them involves uncertainty. Respondents are asked how much risk in terms of probability of death or some other bad outcome they are willing to accept to avoid living in the certainty of the health state being valued. This technique is based on the Expected Utility Theory of decision making under uncertainty developed by Von Neumann and Morgenstern (41), which rests on a set of axioms about the nature of individual preferences when prospects are uncertain. The TTO technique, developed as an alternative to SG (32), was designed to overcome the problems of explaining probabilities to respondents. The TTO technique asks the respondent to choose between two alternatives with certain prospects, i.e., shorter years (x) in full health and longer years (t) in the health state being valued. Respondents are asked to consider trading a reduction in their length of life (t-x) for an improvement in health. The health state valuation is the fraction of healthy years equivalent to a year in a given health state, i.e., x/t.

SG has the most rigorous foundation in theory in the form of the Expected Utility Theory of decision making under uncertainty. There are theoretical arguments against SG in valuing health states, however (34), and little empirical support for Expected Utility Theory (42,43). There are also concerns about the empirical basis of the TTO technique. Some evidence suggests that the time spent in a health state and the time at which it occurs affect TTO values (44,45). A key review by Bleichrodt (46) summarized the different sources of bias in each of these techniques and concluded that SG is subject to mainly upward bias, whereas TTO is subject to upward and downward bias, and so he concludes that, overall, TTO may be preferred. There is

currently no consensus regarding the best technique. One solution might be to try to correct for these biases (47).

A further consideration is that the SG 'utilities' of the HUI-2 and HUI-3 are derived from VAS valuations on the basis of an estimated power function where the difference between VAS ratings and SG utilities is assumed to be a person's attitude toward risk. This conclusion was based on the suggestion of Dyer and Sarin (48), but the validity of the power transformation has been questioned in the literature. Dolan and Sutton (49) and Stevens et al (50) have all demonstrated that other specifications fit the data as well as and in some cases better than a power function.

More recently, there has been interest in basing valuations on ordinal data from ranking and discrete choice experiments. The use of ordinal individual data to generate cardinal health state values draws on random utility theory. It may prove to be a promising alternative to SG and TTO, particularly in more vulnerable populations, because empirical work has found that it can produces similar values for the EQ-5D, SF-6D, and HUI-2 (51,52).

Variant of the technique

Differences in variant may prove more important than choice of technique. It has been shown in numerous studies that SG responses are subject to framing effects, such as whether the probabilities are expressed in terms of success or failure. SG has been found to generate inconsistent valuations with changes to the lower anchor (53).

There are a wide range of variants, including (i) mode of administration (interview or self-completion, computer or paper administration), (ii) the use of props, (iii) presentation of probabilities, (iv) time allowed for reflection, and (v) individual versus group interviews. Few publications in the health economics literature have compared these alternatives, although several researchers have undertaken considerable efforts to try to improve the quality of the data from SG and TTO (e.g., 35,54). The evidence is that values for health states vary considerably between variants of the same technique (49,55). Indeed, it has been found that between-variant differences can be more important than between-technique differences.

Evidence that the way people are asked about their preferences has a major impact on the results raises questions about the nature of people's preferences for health states. It suggests that people do not have well-defined preferences about health before the interview; rather, their preferences are constructed during the interview. This would account for the apparent willingness of respondents to be influenced by the precise framing of the question.

A common criticism of the current methods for eliciting preference is that the tasks are cognitively complex, with respondents being asked to consider variations in up to eight health dimensions alongside a life-and-death scenario involving probabilities of survival. Evidence from psychology suggests that respondents faced with such complex problems would tend to adopt simple heuristic strategies (56). This would be particularly true where respondents have little time to consider their real underlying values.

The foregoing underscores the need to develop respondent-friendly methods. A well-conducted study involving the elicitation of preferences should fully explain the task to the respondent and undertake a practice question. Unfortunately, respondents are typically expected to evaluate health states in one sitting, with little time for reflection. Respondents need more time and support to reflect on their values in order to process such complex information. It has been suggested that respondents could be re-interviewed after they have had time to deliberate on the health state in question. Shiell and colleagues (57) found significant differences between interviews for the same states, with values tending to be higher at subsequent sittings. An implication may be to move away from the current large-scale surveys of members of the general public involving a single sitting to smaller-scale studies of panels from the general public who are better trained and more experienced in the techniques and who are given time to fully reflect on their valuations.

States worse than dead

For states deemed worse than dead, the SG technique asks the respondent to choose between the prospect of death for certain and the uncertain prospect of full health or the state being valued. The probability of full health is varied until the point of indifference where the value of the state worse than death is $-P/(1-P)$. The analogous TTO question asks respondents to choose between the first alternative of dying immediately and the second alternative of some number of years in the state being valued followed by a number of years in full health (where the two periods sum to t). The time in full health x is varied until the person is indifferent between these alternatives. The value for the state worse than death is then given by $-x/(x-t)$. These two formulas, together with the method for calculating the value of states better than dead, produce a range of values between +1 and $-\infty$, which gives greater weight to negative values in the calculation of mean health state scores and presents problems in statistical analysis. To reduce the influence of extreme negative outliers, Patrick and colleagues (58) proposed transforming the values to limit the range from +1 to –1.

It is important to consider states worse than being dead because they are common among preference-based measures. The UK TTO valuation of the EQ-5D, for example, produced mean scores below zero for one-third of all states. Evidence from the UK valuation of the EQ-5D suggests a discontinuity around zero that appears to indicate a special significance is attached to this value. Once people regard a health state as worse than death, they are willing to give quite a low value. Doubts must exist about whether the scale has the same interval properties either side of zero. More research is needed into the valuation of states worse than death.

Source of values

There is evidence of significant variation in values by disease experience, age, and education. In general the evidence points to patients giving health states a higher value than do members of the general population (59-61).

The Washington Panel argued that using the values of the general population favors patients because general population values give a larger value to treatments that restore patients to full health**.** Lenert and colleagues, however, have demonstrated that values of the general population may be less sensitive to movements between points at the lower end of health because of a reference point effect (62). Furthermore, the lower health state valuations of the general population work against

the interests of patients for life-saving interventions since these lower values result in a smaller gain in QALY.

The main normative argument for using general population values seems to hinge on the view that in a publicly funded health care system it is society's resources that are being allocated, and therefore it is the views of the general population that are relevant. In a similar way, it can be argued that it is enrollees of insurance schemes who should be asked to provide values in the context of decisions within a private insurance program rather than patients. By contrast, it has been suggested that the values of patients should be used because they are in the best position to know their own state (63).

The choice of viewpoint is ultimately a value judgement, but it may depend on the reasons for the discrepancy. The main cause of the difference between the values of patients and the general population is adaptation to the condition. Patients experiencing long-term conditions might also change their life goals and expectations. These are aspects of the 'response shift' recognized in the QOL research literature mentioned earlier in this paper. These adaptations will be related to the length of time a patient experiences the condition.

Members of the general population know little about such adaptation. The choice between patient and general population values really comes down to the extent to which these changes should be taken into account. Menzel et al (64) tried to distinguish between 'laudable' adaptations, such as enhancing one's skills, adjusting activities, and even altering perceptions of health, and less desirable changes such as cognitive denial of functional health and suppressed recognition of full health. There is also a genuine concern that many of the changes listed are the result of laudable effort, and incorporating them into health state values in resource allocation may work against patients' interests, which seems in some sense unfair.

It seems difficult to justify using just patient values or uninformed members of the general population to obtain preferences for health measures. Menzel and colleagues (64) suggest that more research is required into the causes of adaptation. They also suggest that patients should be consulted on the extent to which they want their adapted values to be used in decision making. This empirical research would not address the normative question of what aspects of adaptation should ultimately be used. Menzel et al. suggest rather ambitiously that perhaps the general population may be able to disentangle appropriate from inappropriate adaptation. This is consistent with the recommendations of the Washington Panel, which advocate the use of *informed* general population values. A middle way could be to provide members of the general population with patients' values before asking them for their own values. There is important empirical work to be done to develop methods for conveying such information to the general population and to measure its impact on health state values.

The question of whether values from one country or culture can be used in another is also an important one. The emerging evidence suggests that VAS valuations do not vary much between countries, but there are significant differences between countries in TTO values for EQ-5D (9,16,65) and SG valuations of the SF-6D (27) and HUI-2 (52). It seems there are significant differences between countries in health state values

and important variations by sociodemographic characteristics and ethnic group (16,66).

Method of extrapolation

There have been two approaches to estimating a function for valuing states from a health state classification system, the decomposed and composite approaches (33).

The decomposed approach employs MAUT to determine the functional form and the sample of states to be valued. MAUT substantially reduces the valuation task by making simplifying assumptions about the relationship between dimensions. The most commonly used specifications are the additive and multiplicative functional forms. The simple additive functional form assumes dimensions to be independent and hence permits no interaction. This was found by Torrance et al. (13) to be invalid, and the multiplicative function has been used to value the HUI-2 and HUI-3. The multiplicative function permits a very limited form of interaction between dimensions by assuming the interdependency is the same between all dimensions and for all levels of each dimension.

The application of MAUT decomposes the valuation task into three parts. First, each dimension is valued separately to estimate single-attribute utility functions. Second, 'corner states' are valued; these are states where one dimension is at one extreme (usually the worst level) and the rest are set at the other (usually the best) level. Such 'corner states' may represent infeasible combinations of dimension levels unless the descriptive system ensures the dimensions are truly independent. A failure to achieve this with the HUI-2 resulted in a complex 'backing-off' procedure (13). Third, a set of multi-attribute states determined by the model specification is valued. A single respondent undertakes all tasks for the HUI-2 and 3.

The composite approach requires a rather larger sample of states to estimate a function by regression. A common method for sampling states is to use an orthogonal design for estimating an additive model. There are problems, however, with determining the states required to estimate a model with interaction terms. To date, researchers have typically added extra states at random. An important piece of research will be to develop more sophisticated algorithms for sampling health states in order to value key interactions. The statistical models developed for the EQ-5D and SF-6D have estimated crude summary terms for interactions, such as dummy variables taking a value of one for states containing at least one dimension at its worst level (15,17).

The composite approach requires more states than can be valued by a single respondent. The sample states are allocated between respondents, and thus it is necessary to disentangle variation between respondents from variation between states. The statistical modelling has to cope with this hierarchical structure to the data set, and researchers have done this using random effects techniques or by modelling mean health state values (15,17). Modelling also has to cope with a highly skewed data set. Work in this area has explored a range of transformations to overcome this problem, but none has been found to improve the models. Recently, work has been undertaken to use a Bayesian approach that applies a nonparametric method to estimating posterior mean health state values with variances, and the first application to the SF-6D has proved very successful (67).

There has been little written comparing statistical and MAUT approaches. The MAUT multiplicative models are (in a limited sense) more sophisticated than the additive EQ-5D, but they are based on the valuation of a far smaller number of health states. The MAUT approach uses deterministic models and does not allow for the pattern of the error structure. For HUI-2 and HUI-3, VAS was used to value the health states, which means the transformation may introduce another source of error. In principle, however, it is possible to apply the MAUT approach using SG (or TTO) directly.

The choice between these approaches must rest on their ability to predict health state values in an independent sample. A comparison of the MAUT and statistical approaches was undertaken two decades ago in a study of job choice by Currim and Sarin (68). The authors found the statistical approach outperformed a multiplicative algebraic model: the correlation between actual and predicted choices across jobs using SG utility values was 0.64 and 0.16, respectively. This was a very limited study, however, and not in the context of health. A recent study by McCabe (69) applied the MAUT and statistical approaches to the UK valuation of the HUI-3 and found that the statistical approach was marginally better in terms of absolute mean error and percent within the range of plus or minus 0.1 and 0.05 of actual values. This evidence is also not conclusive because it is also influenced by the VAS-SG mapping.

The ensuing debate between MAUT and the statistical approach requires a head-to-head comparison where the two are used optimally, with the statistical approach being based on a better sampling procedure and the MAUT using SG or TTO in a direct fashion.

FUTURE AVENUES FOR RESEARCH

Descriptive systems of existing PBMH

The psychometric properties of existing PBMH need to be better understood through head-to-head comparisons with each other and with non-preference-based measures of health. These data would permit the application of IRT and classical psychometric assessments of the validity of the descriptive systems of these instruments.

Comparison of existing measures

There is currently research under way to compare the PBMH in different patient groups. This will provide a better understanding of the relationship between existing measures and should contribute to an understanding of the reasons for the differences between these measures.

Source of values

There are two related pieces of research into this component. The first will be to elicit patient health values alongside PBMH across a wide range of patient groups. This will help us understand the differences between patient and general population values and how the differences vary between medical condition and other background variables. The second would be to examine the use of informed general population values. This would require research into using patient values in valuing health states among the general population.

Methods for eliciting preferences
There is increasing interest in using ordinal tasks, like ranking or pairwise comparison, to derive the values of health states, and these could prove particularly valuable in enfranchising the more vulnerable groups. Second, the potential role of reflection and deliberation in the valuation of health states needs to be explored in future valuation surveys of PBMH.

States worse than being dead
Research is needed into developing ways to value states worse than being dead that lie on the same scale as states better than being dead.

Estimation
Further research is needed into estimating preference-based index measures from a sample of health state valuations. MAUT should be applied directly using SG and TTO, and statistical modelling can be improved including the use of Bayesian approaches; then, the different approaches need to be compared.

Compare existing measures or develop a new one?
For any major program of research an important question is whether to use existing measures. In the short to medium term the use of existing measures, such as the recently completed valuation of the EQ-5D and SF-36, makes good use of existing data sets. In the longer term, however, there could be a case for developing new measures drawing on more recent psychometric literature (including IRT) and valuation (such as the use of ordinal methods). This research is particularly important in the more vulnerable groups such as children, the very elderly, and people with major mental health problems, where existing measures are often inappropriate.

CONCLUSION
PBHM have come a long way over the last 20 years and offer an important set of tools for economic evaluation and other uses of summary health measures. Existing instruments differ in their descriptive systems and methods of valuation, and so they often generate different scores. Recent developments in assessing and valuing health status provide an important basis for improving preference-based health measurement and for developing more appropriate instruments for special groups, including vulnerable groups such as the very young, the very elderly, and people with serious mental health problems. The challenge is to ensure that new instruments provide a means of generating standardized and comparable scores across populations.

Acknowledgements
I am grateful to colleagues at HEDS who contributed to a 'brainstorm' session on an outline draft of this paper and to subsequent drafts. I would particularly like to acknowledge Professors Chris McCabe, Paul Dolan, and Jennifer Roberts and Dr. Aki Tsuchiya and Katherine Stevens. The views expressed in this paper and all remaining errors are of course mine. The author gratefully acknowledges funding from the UK MRC HSRC.

References

1. Drummond MF, O'Brien B, Stoddart GL, et al. Methods for the Economic Evaluation of Health Care Programmes. 2nd ed. Oxford, UK: Oxford Medical Publications; 1997.

2. National Institute for Clinical Excellence. Guide to the Technology Appraisal Process. London, United Kingdom; National Institute for Clinical Excellence; 2001.

3. Health Technology Board for Scotland. Guidance for Manufacturers on Submission of Evidence Relating to Clinical and Cost Effectiveness in Health Technology Assessment. Glasgow, Health Technology Board for Scotland; 2002.

4. Commonwealth Department of Health, Housing and Community Service. Guidelines for the Pharmaceutical Industry on the Submission to the Pharmaceutical Benefits Advisory Committee. Canberra: Australian Government Publishing Service; 1992.

5. Ministry of Health (Ontario). Ontario Guidelines for the Economic Evaluation of Pharmaceutical Products. Toronto, Ontario, Canada: Ministry of Health; 1994.

6. Gold MR, Siegel JE, Russell LB, et al. Cost-Effectiveness in Health and Medicine. Oxford, United Kingdom: Oxford University Press; 1996.

7. Fryback DG, Dasbach ED, Klein R, et al. Health assessment by SF-36, Quality of Well-Being index and time trade-offs: predicting one measure from another. Med Decis Making. 1992;12:348-356.

8. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted adjusted life years: estimation of the health utility index (HUI2) from the SF-36. Med Decis Making. 2001,21:105-112.

9. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. Health Econ. 2002;11:341-353.

10. Brazier JE, Kolotkin RL, Crosby RD, et al. Estimating a preference-based single index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) instrument from the SF-6D. Value Health 2004;7:490-498.

11. Lundberg L, Johannesson M, Isacson DG, et al. The relationship between health- state utilities and the SF-12 in a general population. Med Decis Making. 1999;19:128-140.

12. Kaplan RM, Anderson JP. A general health policy model: update and applications. Health Serv Res. 1988;23:203-235.

13. Torrance GW, Feeny DH, Furlong WJ, et al. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Med Care. 1996;34:702-722.

14. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002;40:113-128.

15. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095-1108.

16. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005;43:203-220.

17. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21:271-292.

18. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004;42:851-859.

19. Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. J Health Serv Res Policy. 1999;4:174-184.

20. Hatoum HT, Brazier JE, Akhras KS. Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical setting. Value Health. 2004;7:602-609

21. Brazier J, Roberts J, Tsuchiya A, et al. A comparison of the EQ-5D and the SF-6D across seven patient groups. Health Econ. 2004;13:873-884.

22. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. Health Econ. 2003;12:1061-1067.

23. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. Oxford, United Kingdom: Oxford University Press; 1996.

24. World Health Organization. Constitution of the World Health Organization. Basic documents. Geneva, Switzerland:World Health Organization,1948.

25. Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med. 1999;48:1531-1548.

26. Brooks RG. Euroqol: the current state of play. Health Policy. 1996;37:53-72.

27. Brazier J, Fukuhara S, Ikeda S, et al. The Japanese valuation of the SF-6D and comparison to the UK values. HEDS DP 01/05. Sheffield, UK: University of Sheffield, UK; 2005.

28. Barton GR, Bankart J, Davis AC, et al. Comparing utility scores before and after hearing-aid provision: results according to the EQ-5D, HUI3 and SF-6D. Appl Health Econ Health Policy. 2004;3:103-105.

29. Kobelt G, Kirchberger I, Malone-Lee J. Review. Quality-of-life aspects of the overactive bladder and the effect of treatment with tolterodine. BJU Int. 1999;83:583-590.

30. Guyatt G. Commentary on Jack Dowie, "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions". Health Econ. 2002;11:9-12.

31. Brazier JE**,** Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. Health Econ. 1999;8:41-51.

32. Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ. 1986;5:1-30.

33. Froberg DG, Kane RL. Methodology for measuring health-state preferences--II: Scaling methods. J Clin Epidemiol. 1989;42:459-471.

34. Richardson J.. Cost-utility analysis: what should be measured? Soc Sci Med. 1994;39:7-21.

35. Dolan P. Gudex C, Kind P, et al. Valuing health states: a comparison of methods. J Health Econ. 1996;15:209-231.

36. Green C, Brazier J, Deverill M. Valuing health-related quality of life. A review of health state valuation techniques. Pharmacoeconomics. 2000;17:151-165.

37. Robinson A, Loomes G, Jones-Lee M. Visual analogue scales, standard gambles, and relative risk aversion. Med Decis Making. 2001; 21:17-27.

38. Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. Med Decis Making. 1997;17:208-216.

39. Nord E. The validity of a visual analogue scale in determining social utility weights for health states. Int J Health Plann Manage. 1991;6;234-242.

40. Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? Soc Sci Med. 1997;45:1289-1297.

41. Von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton, NJ: Princeton University Press; 1944.

42. Camerer C. Individual decision-making. In:Kagel J, Roth A (eds) Handbook of Experimental Economics (Princeton University Press)

43. Schoemaker PJH. The expected utility model: its variants, purposes, evidence and limitations. J Econ Lit. 1982;20:529-563.

44. Sutherland HJ, Llewellyn-Thomas H, Boyd NF, et al. Attitudes toward quality of survival. The concept of "maximal endurable time". Med Decis Making. 1982;2:299-309.

45. Dolan P, Gudex C. Time preference, duration and health state valuations. Health Econ. 1995;4:289-299.

46. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 2002;11:447-456

47. Oliver A. The internal consistency of the standard gamble: tests after adjusting for prospect theory. J Health Econ. 2003;22:659-674.

48. Dyer JS, Sarin RK. Relative risk aversion. Manage Sci. 1982;28:875-886.

49. Dolan P, Sutton M. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. Soc Sci Med. 1997;44:1519-1530.

50. Stevens K, McCabe C, Brazier J. Mapping between Visual Analogue Scale and Standard gamble data: Results from the UK study using the Health Utilities Index II Framework. Health Eonomics Study Group January 2003, Leeds, UK.

51. Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul Health Metr. 2003;1:12.

52. McCabe C, Brazier J, Gilks P, et al. Estimating population cardinal health state valuation models from individual ordinal (rank) health state preference data. Sheffield Health Economics Group Discussion Paper 04/02; 2004.

53. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, et al. The measurement of patients' values in medicine. Med Decis Making. 1982;2:449-462.

54. Furlong W, Feeny D, Torrance GW, et al. Guide to design and development of health state utility instrumentation. Hamilton, Ontario, Canada: McMaster University; 1990. Centre for Health Economics and Policy Analysis Paper 90-9.

55. Brazier JE, Dolan P. Evidence of preference construction in a comparison of SG methods. Sheffield, UK: Health Economics and Decision Science Department, University of Sheffield, UK; 2004.

56. Lloyd AJ, Hutton J. Do decision making heuristics distort efforts to elicit preferences? Paper presented to Developing Economic Evaluation Methods workshop. York, United Kingdom; 2002.

57. Shiell A, Seymour J, Hawe P, et al. Are preferences over health states complete? Health Econ. 2000;9:47-55.

58. Patrick DL, Starks HE, Cain KC, et al. Measuring preferences for health states worse than death. Med Decis Making. 1994;14:9-18.

59. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. J Chron Dis. 1978;31:697-704.

60. Boyd NF, Sutherland HJ, Heasman ZK, et al. Whose utilities for decision analysis? Med Decis Making. 1990;10:58-67.

61. Hurst NP, Jobanputra P, Hunter M, et al. Validity of Euroqol - a generic health status instrument - in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. Br J Rheumatol. 1994;33:655-662.

62. Lenert LA, Treadwell JR, Schwartz CE. Associations between health status and utilities: implications for policy. Med Care. 1999;37:479-489.

63. Buckingham K. A note on HYE (healthy years equivalent). J Health Econ. 1993;12:301-309.

64. Menzel P, Dolan O, Richardson J, et al. The role of adaptation to disability and disease in health state valuation: a preliminary normative analysis. Soc Sci Med. 2002;55:2149-2158.

65. Badia X, Roset M, Herdman M, et al. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making.2001;21:7-16.

66. Roberts J, Dolan P. To what extent can we explain time trade-off values from other information about respondents? Soc Sci Med. 2002;54:919-929.

67. Kharroubi SA, O'Hagan A, Brazier JE. Estimating utilities from individual health preference data: a nonparametric Bayesian approach. Appl Stat. (in press).

68. Currim IS, Sarin RK. A comparative evaluation of multi-attribute consumer preference models. Manage Sci. 1984;30:543-561.

69. McCabe C. Estimating preference weights for a paediatric health state classification (HUI-2) and a comparison of methods. PhD thesis, University of Sheffield, UK; 2003.

Table 1: Characteristics of multi-attribute utility scales

MAUS	Descriptive characteristics			Valuation characteristics			
	Dimension	**Levels**	**Health states**	**Valuation technique**	**Method of extrapolation**	**Sample**	**Country**
QWB	Mobility, physical activity, social functioning 27 symptoms/problems	 3 2	1170	VAS	Statistical	866 (General population)	USA (San Diego)
HUI-2	Sensory, mobility, emotion, cognitive, self-care, pain, fertility	4-5 3	24,000	VAS tranformed into SG	MAUT	203 (parents)	Canada (Hamilton), UK
HUI-3	Vision, hearing, speech, ambulation, dexterity, emotion, cognition, pain	5-6	972,000		MAUT	504 (General population)	Canada (Hamilton), France
EQ-5D	Mobility, self-care, usual activities, pain/discomfort, anxiety/depression	3	243	TTO and VAS	Statistical	3395 (General population)	UK, Japan, Spain, USA (among others)
SF-6D	Physical functioning, role limitation, social functioning, pain, energy, mental health	4-6	18,000	SG	Statistical	611 (General population)	UK, Japan, Hong Kong

Note: VAS – visual analogue scale, TTO – Time trade-off, SG – standard gamble, MAUT – multi-attribute utility theory