HEALTH ECONOMICS
Health Econ. **13**: 281–290 (2004)
Published online 20 June 2003 in Wiley InterScience (www.interscience.wiley.com). **DOI:**10.1002/hec.834

OUTCOME VALUATION

A multi-method approach to measuring health-state valuations

Joshua A. Salomon[a,b,]* and Christopher J.L. Murray[a]
[a] *Evidence and Information for Policy, World Health Organization, Geneva, Switzerland*
[b] *Harvard Center for Population and Development Studies, Cambridge, MA, USA*

Summary

Existing techniques for eliciting health-state valuations incorporate both strength of preferences for health states and other values such as risk aversion or time preference. This paper presents a new methodological approach that allows estimation of a set of core underlying health-state values based on responses elicited through multiple measurement techniques. A study was undertaken in which respondents completed the visual analogue (VAS) scale, time trade-off (TTO), standard gamble (SG) and person trade-off (PTO) for a range of states. By specifying flexible parametric functions to explain responses on each measurement technique, we estimated both the underlying strength of preference values for the health states in the study and the values for a set of auxiliary parameters characterising risk attitudes, discount rates, distributional concerns and scale distortion effects in the group of respondents. This study demonstrates that it is possible to understand responses on these four different measurement techniques based on a consistent set of core values. The approach presented here can provide insights into different sources of observed variation in VAS, TTO, SG and PTO responses and facilitate appropriate adjustment of valuations elicited through different methods for use in summary health measures and economic analyses.

Keywords health-state valuations; standard gamble; visual analogue scale; time trade-off; person trade-off

Introduction

In summary measures of population health [1–3] and analyses of the cost effectiveness of health interventions [4], an essential data input is a set of weights assigned to time spent in different health states, which provide the critical link between information on mortality and information on the spectrum of non-fatal health experiences among the living. Several different techniques for the valuation of health states have been proposed and used widely in the derivation of these weights, including the visual analogue scale (VAS), standard gamble (SG), time trade-off (TTO) and person trade-off (PTO) [5–8]. Thus far, there has been little agreement as to which technique is most appropriate. Arguments for and against different methods have been based on economic theory [9], comparisons of psychometric properties [10] and ethical grounds [11]. Empirical results from studies using multiple methods have demonstrated differences in the weights produced by the various methods, but have not led to the emergence of a single preferred method [5,10,12–20].

Reviews of the techniques for eliciting valuations have emphasised the importance of reflecting carefully on the intended use of the

*Correspondence to: Harvard Center for Population and Development Studies, 9 Bow Street, Cambridge, MA 02138, USA. E-mail: jsalomon@hsph.harvard.edu

Received 20 September 2001
Accepted 17 March 2003

derived health-state weights in evaluating different methods [6,7,21]. Nord [7] has called for a 'reflexive equilibrium' to establish that the inferences drawn by researchers from responses to a particular measurement method are consistent with the actual preferences that are elicited. Richardson [21] has similarly defined two key criteria for choosing an elicitation method: a 'weak interval' criterion, which requires that units of measurement must allow for meaningful comparisons of differences in values; and a 'strong interval' criterion demanded by the use of time-based summary health measures, which requires that a given increase in the health-state weight must be equivalent 'in some meaningful way' (p.15) to the same proportional increase in longevity. The evaluative criteria suggested by Nord and Richardson both point to the need for measurement methods that enable meaningful statements about the strength of preferences for time lived in different health states.

For each of the commonly used elicitation techniques, researchers have challenged whether responses on these techniques may be interpreted directly as measures of strength of preference. The interval properties of the VAS have been questioned, and experiments in psychophysics have indicated that respondents may use VAS scales in ways that relate non-linearly to the stimuli they are evaluating, e.g. through a power function [22]. With regard to the standard gamble, it has been argued that the von Neumann–Morgenstern utilities inferred from this method do not correspond to strength of preferences under certainty because choices over lotteries are determined in part by attitudes toward risk [21,23]. Although Richardson suggests that the TTO and PTO appear to fulfil the strong and weak interval criteria [21], responses to both techniques may be influenced by factors other than strength of preference; TTO responses may depend on time preference or threshold effects [24,25], and proponents of the PTO acknowledge that 'distributive considerations become a serious confounding factor' that complicates interpretation of PTO responses [7, p. 562].

While it may be debated whether some of these factors (risk, time preference, distributional concerns) are relevant to certain clinical and policy decisions, there are applications – such as comparisons of average health levels in different populations [1] – for which it is clearly desirable to use health-state weights that relate more precisely to the relative values people attach to different health outcomes rather than weights that conflate these values with other considerations such as attitudes toward risk [3]. Ideally, continuing work on improving modes of eliciting health-state valuations will eventually yield methods that minimise the biases inherent in available techniques. Until then, the demand for comparative health analyses requires researchers to make the most of existing tools.

Amidst ongoing reflection on the relative merits of different elicitation methods, one common strategy has been to estimate mathematical functions to map from one method to another in studies where multiple methods are used [5,10,18,26,27]. If none of the methods elicits responses that are directly interpretable as strength of preference values, however, it is worth considering whether a different approach may be warranted. In this paper, we propose an alternative, which acknowledges that none of the available methods gives us the exact quantity of interest, but that each of them produces responses from which this quantity may be imputed.

Indeed, the standard transformations used to convert responses on each method into weights scaled from zero to unity already entail implicit assumptions about the relationships between these methods and some unobserved measure; i.e. 'utility' or 'value' in each case is a derivative construct inferred from observed preference responses. For example, translation of standard gamble responses to health state weights relies on a representation theorem that posits the existence of a latent utility scale, such that a ranking of lotteries by their expected payoffs on this scale is consistent with stated preferences over these lotteries [23]. Our proposal is to enrich the standard representation theorems that implicitly underlie each single method in order to present a coherent framework for understanding all of the methods – in a way that accounts explicitly and simultaneously for multiple factors in addition to strength of preference. By formalising our understanding of how each of the measurement techniques relates to strength of preferences for health states, we aim to abstract a core set of underlying health-state values from responses to multiple measurement techniques applied to a range of different states. This paper describes a first study illustrating this alternative approach.

Methods

Health-state valuation exercise

A multi-method health-state valuation exercise was implemented in a convenience sample of 69 public health professionals from 28 different countries. Twelve health states were selected to span a wide range of severity and capture many key domains of health including physical and mental function as well as pain. The states were described by brief labels and standardised descriptions of levels on six dimensions: mobility, self-care, usual activities, pain/discomfort, anxiety/depression and cognition. The standardised descriptions were based on a modification of the EuroQol EQ-5D classification system [28], adding cognition to the five dimensions included in EQ-5D [29], and increasing the number of levels in each dimension from the three used in EQ-5D (e.g. for mobility: 'no problems in walking about', 'some problems in walking about' and 'confined to bed') to five in order to offer greater resolution in describing intermediate health states [30].

Each participant undertook the following tasks for all 12 health states.

1. *Visual analogue scale* (*VAS*): Respondents were asked to imagine what it would be like to live in each health state, assuming the same life expectancy (10 years) in every state, and to rank the states from most desirable to least desirable with the aid of index cards. They were then asked to rate the 12 states on a visual analogue scale from 0 (a state comparable to death) to 100 (best imaginable health), with 100 equally spaced tick marks labelled at every even number. Respondents were encouraged to use the intervals on the scale meaningfully, such that similarly attractive states would be placed close together while very different states would be placed far apart.
2. *Time trade-off* (*TTO*): Respondents were asked to imagine a choice between: (a) living in a given health state with a life expectancy of 10 years; or (b) living in ideal health, but with shortened life expectancy. A worksheet for each health state guided respondents through a series of trade-offs in order to identify the number of years of ideal health ($\leqslant 10$) considered equivalent to 10 years in the given health state.
3. *Standard gamble* (*SG*): Respondents were asked to imagine a choice between: (a) living for 10 years in a given state with certainty; or (b) accepting a risky procedure that offered a chance at living the 10 years in ideal health but presented some risk of immediate death. A worksheet was used to help identify the level of risk at which the uncertain option would be equally attractive as the certain option.
4. *Person trade-off* (*PTO*): Respondents were asked to imagine that they were decision makers facing a choice between: (a) a programme that would prevent the deaths of 100 fully healthy individuals (thus extending their lives for 10 years); or (b) a programme that would prevent the onset of a given health problem in some number of healthy people (thus improving their health expectancy from 10 years in sub-optimal health to 10 years in ideal health). A worksheet was provided to help identify the number of averted health problems considered equivalent to the prevention of 100 deaths.

Before beginning each task, basic instructions were given, and two volunteers were led through examples. After the instruction, individuals were allowed to complete each task for the 12 states. For the TTO, SG and PTO, once respondents completed each exercise, they were presented with a summary of their responses for all 12 conditions (in the original units of response) and allowed to revise any of the values if they wished to do so. The exercise was conducted in two 90-minute sessions, with a 20-minute break between sessions.

Analysis

We assumed that each measurement technique produces responses that may be described by an increasing function of the strength of preferences for different health states. Parametric specifications of these functions were developed based on previous theoretical and empirical findings. For each technique, one auxiliary parameter was used to describe the relationship between strength of preference and responses on that type of question (mathematical details provided in Appendix A).

For the VAS, a long-standing result from psychophysics suggests that individual perceptions of sensory stimuli of varying intensities tend to follow a power function transformation of the true intensity levels [22]. We have based the model for the VAS responses on this finding.

For the TTO, we allowed for time preference in individual choices over health and longevity [31].

If respondents have non-zero discount rates, then the two streams of life that are compared in the TTO (e.g. 10 years in state X and 5 years in perfect health) must be translated into their equivalent present values in computing implied health-state weights [24]. An exponential discounting model was used, with a single parameter to capture the discount rate.

Based on the observation that choices involving lotteries reflect both strength of preferences and attitudes toward risk, we modelled SG responses as a function of the underlying health-state value and a risk aversion parameter. We examined several different formulations including exponential, logarithmic and power functions, based on the utility-based theoretical framework for characterising intrinsic risk aversion presented in [32].

For the PTO, proponents have recognised that responses depend both on the value attached to a particular health state and on distributional concerns [11]. In responding to PTO questions, some individuals may be reluctant to choose to prevent large numbers of non-fatal health outcomes when the option of preventing deaths is available. This so-called 'rule of rescue' [33] may be interpreted as a population analogue to risk aversion in individual choices. Given this parallel interpretation, and similarities in empirical results from the two methods, we have therefore modelled the PTO responses using similar functions as those used for the standard gamble, but allowing for a distinct parameter to capture distributional concerns.

Based on the specifications described above, the model included a total of 16 parameters of primary interest: 12 core health-state values, plus four auxiliary parameters. Maximum likelihood methods were used to estimate the parameters (see Appendix A). The stochastic component was modelled as a truncated normal distribution constrained between 0 and 1. Analysis of the data revealed wide disparities in the variance of responses across methods, with the variance for different states related strongly to the mean values. We therefore expressed the variance in the model as a linear function of the mean and allowed the slope and intercept of this function to differ by measurement method. Numerical simulation methods were used to compute confidence intervals around the estimated strength of preference values and auxiliary parameters in the model.

Results

A total of 3257 responses were collected from the 69 participants. Of these, there were 576 cases where respondents revised their initial assessments upon completing the TTO, SG or PTO exercises for all 12 states. TTO responses were revised more frequently (37%) than either SG (17%) or PTO (17%) responses. Results reported here are based on the revised values where applicable.

Responses from the four different measurement methods are summarised in Figure 1. The responses on all methods have been mapped to a scale on which 1 indicates the best imaginable health and 0 indicates a state equivalent to death. Overall, the PTO values are the highest, followed by SG values, then TTO values and then VAS values. In addition to being lower on average for a given health state, the VAS values have the smallest variance across respondents, but span the largest range across states. For severe states, the SG and PTO both produce considerably higher variance across respondents than either the TTO or VAS. Furthermore, the variances for different states show a strong inverse correlation to the mean health-state weights, especially for the TTO, SG and PTO, which is accommodated in the estimation of the model.

There is considerable agreement in the rank orderings of the health states implied by the different methods, both at the aggregate level (Figure 1) and at the individual level (Table 1). The high level of agreement on the rankings of states using different measurement methods supports the notion that the various methods may be related to a common set of underlying strength of preference values through a series of monotonic functions.

Table 2 lists the estimated strength of preference values for the 12 states in this study, along with the approximate 95% confidence interval for each estimate. The rank order of these estimated core values is consistent with the observed rankings from the four measurement methods. Because our method distinguishes the effects of risk aversion and equity concerns from the valuation of the health state itself, mild health states have lower ratings than in previous studies. This may have important implications for the economic analysis of preventive and curative health interventions.

The estimated coefficients for the auxiliary parameters (Table 3) imply that the respondents are strongly risk averse and have preferences

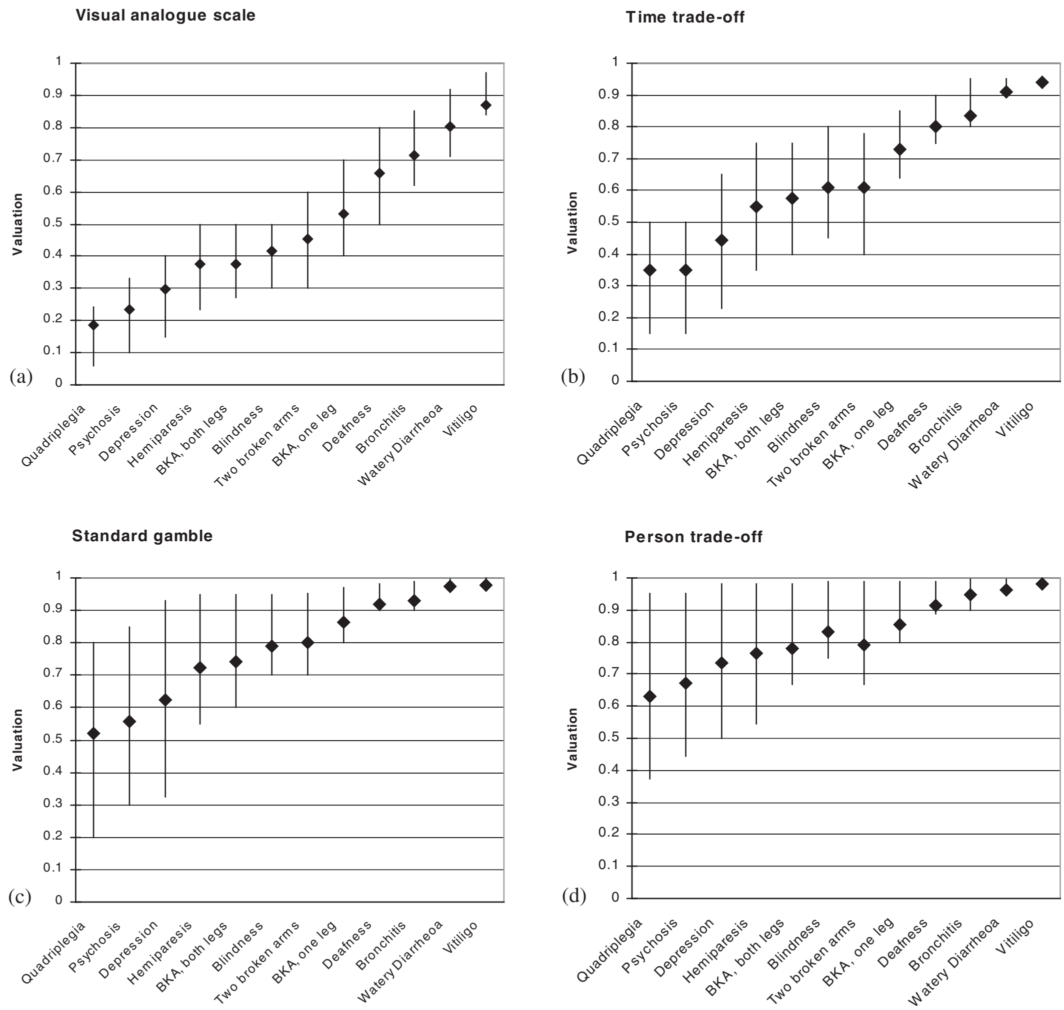

Figure 1. Mean response and interquartile range across 69 respondents for 12 states and four valuation methods (BKA = below the knee amputation)

Table 1. Individual Spearman's rank correlation coefficients for pairwise comparisons of different valuation methods

	Mean	Median	Interquartile range
VAS–TTO	0.94	0.84	0.83–0.97
VAS–SG	0.94	0.85	0.83–0.98
VAS–PTO	0.85	0.78	0.67–0.95
TTO–SG	0.92	0.86	0.81–0.97
TTO–PTO	0.84	0.80	0.70–0.95
SG–PTO	0.86	0.79	0.66–0.94

consistent with strong distributional concerns. The results point to a moderate degree of scale distortion in VAS responses. For the TTO, the model results indicate negative time preference, which runs against conventional health economics wisdom but has been found in other empirical studies of individual discount rates [31,34,35]. Figure 2 displays graphically the modelled relationships between responses on the four different measurement methods and the unobserved core strength of preference values for the 12 states in this study.

We may compare the predicted responses from the different methods based on the maximum

Table 2. Estimated strength of preference values and ranges based on multiple-method protocol using VAS, TTO, SG and PTO techniques

State	Value	Range[a]
Quadriplegia	0.29	(0.23–0.37)
Active psychosis	0.31	(0.24–0.39)
Major depression	0.36	(0.28–0.44)
Hemiparesis	0.43	(0.35–0.52)
Below the knee amputation, both legs	0.45	(0.36–0.54)
Two broken arms	0.49	(0.41–0.59)
Blindness	0.50	(0.40–0.59)
Below the knee amputation, one leg	0.60	(0.50–0.68)
Deafness	0.71	(0.62–0.79)
Chronic bronchitis	0.77	(0.68–0.84)
Watery diarrhoea	0.84	(0.76–0.89)
Vitiligo on face	0.89	(0.83–0.94)

[a] 95% confidence interval.

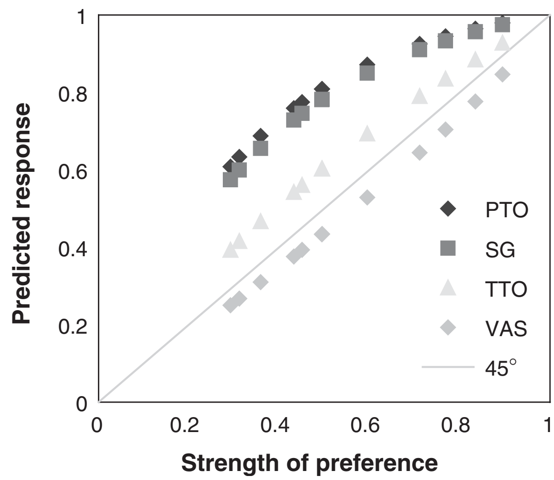

Figure 2. Modelled relationships between core strength of preference values and responses on four measurement methods, with predicted responses for 12 states

Table 3. Maximum likelihood estimates of auxiliary parameters[a]

Parameter	Estimate	Std. error
Scale distortion (VAS)[b]	0.83	0.10
Discount rate (TTO)	−0.089	0.039
Risk attitude (SG)[c]	2.6	0.43
Distributional concerns (PTO)	2.9	0.44

[a] See Appendix A for complete description of parameters.
[b] A number <1 indicates a convex curve.
[c] A number >0 indicates risk aversion. The parameter for distributional concerns in the PTO has a similar interpretation.

likelihood estimates of the core health-state values and auxiliary parameters (plotted in Figure 2) to the observed distributions of responses shown in Figure 1 to assess the fit of the model. The intraclass correlation coefficients [36] between the predicted responses and the mean observed responses by state are greater than 0.98 for all methods, demonstrating a very high level of agreement.

Discussion

In this paper, we have demonstrated that it is possible to explain responses to the SG, TTO, PTO and VAS based on a consistent set of core values for a range of health states. Although none of these elicitation methods provides a pure and direct measure of strength of preference, we may explicitly model the process by which individuals respond to different types of measurement techniques given an underlying valuation for a health state. Our finding that individuals are highly consistent in the orderings of health states via different elicitation methods supports the notion that each different method may be related monotonically to a common set of core values.

With wider use of summary measures of population health and economic appraisal of health interventions, there is considerable interest in the extent of cultural variation in valuations of health states [37–40]. Variation across individuals in responding to different types of questions may be due to at least three different factors: different interpretation of the health-state descriptions; differences in risk aversion, time preference, distributional concerns, VAS distortions or other factors that mediate the translation of strength of preference into responses on different methods; or differences in the underlying strength of preference values assigned to the same health state.

One important component of the current research agenda on health-state valuations is to improve the mode of description of health states as stimuli for valuation. In the study described here, each of the state descriptions included domain levels from a standardised descriptive system. There may be some doubts, however, as to how much of this information was actually reflected in the valuations; the extent to which individuals substitute their own preconceptions about health states for the descriptions that are provided is an important concern. In ongoing studies, we are experimenting with alternative modes of description, including the use of respondents' own ratings of each health state on a range of domains. There may also be some concern that the need to standardise the durations of different hypothetical health states produces unrealistic scenarios, for example, having two broken arms in stiff casts for 10 years. The finding in this study, however, that this state was rated as worse than blindness, deafness or amputation of one leg suggests that respondents were able to suspend disbelief about the realism of the hypothetical and use their imaginations to focus on the actual health impact implied by this state; a number of respondents remarked on the profound difficulties this state would create for most daily activities, and this perception is reflected in the low valuations for this state.

Using a multi-method approach as outlined here, it should be possible to disentangle cultural or individual variation in factors such as risk aversion or time preference from variation in strength of preferences for health states. This will require larger data sets and the elaboration of the statistical model used here to allow for the core health-state values and auxiliary parameters to be treated as random variables. In the meantime, observed cross-cultural variation in the results from one method such as the VAS or TTO should be interpreted with caution, as it does not necessarily indicate cultural variation in the (unobserved) health-state value itself. It will be useful to repeat this exercise among general population samples in order to examine whether results differ from those reported here among health professionals.

It is important to recognise that different results might be obtained depending on the functional form of the models that are used. While consistencies in ordinal rankings from the four methods encourages the conclusion that they all relate to a common set of core values, cardinal measurement of these values requires some particular parametric specification of these relationships. For this study, we have demonstrated our proposed approach using relatively simple parametric forms based on previous theoretical and empirical findings, but other plausible alternatives should be considered. For example, researchers have noted that the SG may be biased by deviations from expected utility due to probability weighting, loss aversion and scale compatibility [41–43]. Because we modelled the relationship between SG responses and strength of preference using a single parameter, this parameter may capture a combination of risk aversion with these additional factors. As Bleichrodt has observed [43], loss aversion and scale compatibility also influence the TTO, which may partially explain the upward bias in TTO responses that we have attributed to negative time preference due to our limited specification. Consideration of alternative functional forms for each of the different measurement methods remains an important area for further research. As the purpose of this paper was to demonstrate a new approach rather than to draw definitive conclusions about levels of risk aversion, time preference and other confounding factors in health-state valuations, we emphasise that interpretation of the coefficient values estimated here must be undertaken cautiously, and it will be important to investigate the sensitivity of the results to the choice of models in future applications.

Other avenues for methodological advances also warrant further attention, for example, the use of Bayesian statistical methods for incorporating additional prior information into the estimation framework, or the use of statistical models that account more appropriately for clustering in the data. We acknowledge that our results may underestimate the standard errors of the coefficient estimates by treating all observations as if they were independent. The nature of measurement error in the application of these methods also merits further examination. While the truncated normal distribution we used improves on the traditional assumption of normality by accounting for the natural constraints of the data, more work is required before the most appropriate choice of error distributions is clear.

Despite these limitations, the results of this study suggest that new approaches to health-state valuation may hold promise. We are hopeful that wider application of these methods can lead to significant improvements in the development of valid, reliable and comparable health-state valuations for use in summary measures of population health and evaluations of the benefits of health interventions.

Acknowledgements

We gratefully acknowledge Omar Ahmad, Steve Begg, Maria Bulzacchelli, Gabriella Covino, Alan Lopez, Meghna Majmudar, Colin Mathers, Catherine Michaud and Theo Vos for facilitating the data collection, and Erik Nord, Jeff Richardson, Daniel Hausman and Gary King for helpful discussions. This project was supported by the National Institute on Aging (Grant P01-AG17625).

Appendix A

Data

Each measurement technique produces responses from individuals (indexed by i) for health states (indexed by x) on a scale particular to that method (Table A1).

Responses from each of the four methods may be mapped onto a scale that ranges between 0 and 1, with 1 representing ideal health. Each response by individual i for health state x using method m is then expressed as $r_{i,x}^{m}$:

$$r_{i,x}^{\mathrm{VAS}} = \frac{s_{i,x}}{100} \quad \text{(A1)}$$

$$r_{i,x}^{\mathrm{TTO}} = \frac{y_{i,x}}{10} \quad \text{(A2)}$$

$$r_{i,x}^{\mathrm{SG}} = 1 - p_{i,x} \quad \text{(A3)}$$

$$r_{i,x}^{\mathrm{PTO}} = 1 - \frac{100}{n_{i,x}} \quad \text{(A4)}$$

Statistical model

The rescaled responses from each method are assumed to be distributed truncated normal, with parameters specific to state and method. For ease of interpretation, we may describe the truncated distribution in terms of the parameters of the corresponding untruncated normal distribution:

$$TN\left(r_{i,x}^{m} \middle| \mu_x^m, \sigma_x^m\right) = \frac{N\left(r_{i,x}^{m} \middle| \mu_x^m, \sigma_x^m\right)}{R\left(\mu_x^m, \sigma_x^m\right)} \quad \text{(A5)}$$

where TN stands for the truncated normal distribution with limits at 0 and 1, and $R\left(\mu_x^m, \sigma_x^m\right)$ is the area of the untruncated normal distribution falling within the unit interval, a normalising factor that keeps the area under the truncated distribution equal to 1.

The deterministic component of the model relates the mean values for the untruncated distributions for each state and method, μ_x^m, to a latent strength of preference value, v_x, which is state-specific but shared across all four methods. In each case, the relationship is described by a monotonically increasing transformation function that depends on one auxiliary parameter.

The VAS transformation is a power function with parameter θ_1.

$$\mu_x^{\mathrm{VAS}} = 1 - [1 - v_x]^{\theta_1} \quad \text{(A6)}$$

This formulation is based on results from psychophysics experiments [22] and has been suggested by Torrance [5] in modelling the functional relationship between VAS and TTO.

For the TTO, we allow for time preference by translating the two durations referenced in the TTO to their equivalent present values [24] using the formula for discounting a continuous stream

Table A1

Method	Response	Units and scale	Interpretation
VAS	$s_{i,x}$	0–100	Rating of health state x by respondent i
TTO	$y_{i,x}$	Years 0–10	Years of perfect health equivalent to 10 years in state x for respondent i
SG	$p_{i,x}$	Risk 0–100%	Risk of death at which treatment is equivalent to certainty in state x for respondent i
PTO	$n_{i,x}$	Persons[a] 100 to ∞	Number of averted cases of x equivalent to 100 deaths averted for respondent i

[a] In principle, n may be less than 100, which would imply that preventing a case of health state x is preferred to preventing the death of an individual in ideal health. In practice, all respondents indicated values greater than 100 for all states.

of life (with θ_2 representing the discount rate):

$$v_x = \frac{(1/\theta_2) - (1/\theta_2)(e^{-\theta_2 * 10 * \mu_x^{TTO}})}{(1/\theta_2) - (1/\theta_2)(e^{-\theta_2 * 10})} \quad (A7)$$

Solving for μ_x^{TTO},

$$\mu_x^{TTO} = -\frac{1}{10\theta_2}\ln\left[1 - (1 - e^{-10\theta_2})v_x\right] \quad (A8)$$

The standard gamble function has one parameter θ_3 that represents an individual's risk aversion. The formulation is derived from utility theory, as described by Bell and Raiffa [32].

$$\mu_x^{SG} = \frac{-e^{-\theta_3 v_x} + 1}{-e^{-\theta_3} + 1} \quad (A9)$$

The PTO formulation is parallel to the standard gamble formulation, but in this case the parameter θ_4 represents aversion to decisions resulting in loss of life, the so-called 'rule of rescue' [33].

$$\mu_x^{PTO} = \frac{-e^{-\theta_4 v_x} + 1}{-e^{-\theta_4} + 1} \quad (A10)$$

For each method, the variance is also allowed to vary across health states and is related to the mean of the distribution through a method-specific linear function:

$$\left(\sigma_x^m\right)^2 = \beta_0^m + \beta_1^m \mu_x^m \quad (A11)$$

Parameters of the model are estimated by maximising the likelihood

$$\prod_i \prod_x \prod_m TN\left(r_{i,x}^m | v, \theta, \beta\right) \quad (A12)$$

References

1. Mathers CD, Sadana R, Salomon JA, Murray CJL, Lopez AD. Healthy life expectancy in 191 countries, 1999. *Lancet* 2001; **357**(9269): 1685–1691.
2. Field JM, Gold MR (eds). *Summarizing population health: directions for the development and application of population metrics*. National Academy Press: Washington, DC, 1998.
3. Murray CJL, Salomon JA, Mathers C. A critical examination of summary measures of population health. *Bull WHO* 2000; **78**(8): 981–994.
4. Gold MR, Siegel JE, Russell LB, Weinstein MC (eds). *Cost-effectiveness in Health and Medicine*. Oxford University Press: New York, 2001.
5. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Econ Planning Sci* 1976; **10**(3): 129–136.
6. Froberg DG, Kane RL. Methodology for measuring health-state preferences–II: Scaling methods. *J Clin Epidemiol* 1989; **42**(5): 459-471.
7. Nord E. Methods for quality adjustment of life years. *Soc Sci Med* 1992; **34**(5): 559–569.
8. Nord E. The person-trade-off approach to valuing health care programs. *Med Decis Making* 1995; **15**(3): 201–208.
9. Torrance GW. Toward a utility theory foundation for health status index models. *Health Serv Res* 1976; **11**(4): 349–369.
10. Krabbe PF, Essink-Bot ML, Bonsel GJ. The comparability and reliability of five health-state valuation methods. *Soc Sci Med* 1997; **45**(11): 1641–1652.
11. Arnesen T, Nord E. The value of DALY life: problems with ethics and validity of disability adjusted life years. *BMJ* 1999; **319**(7222): 1423–1425.
12. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res* 1973; **8**(3): 228–245.
13. Kaplan RM, Bush JW, Berry CC. Health status index: category rating versus magnitude estimation for measuring levels of well-being. *Med Care* 1979; **17**(5): 501–525.
14. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes. Comparison of assessment methods. *Med Decis Making* 1984; **4**(3): 315–329.
15. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ* 1986; **5**(1): 1–30.
16. Bass EB, Steinberg EP, Pitt HA, Griffiths RI, Lillemoe KD, Saba GP *et al*. Comparison of the rating scale and the standard gamble in measuring patient preferences for outcomes of gallstone disease. *Med Decis Making* 1994; **14**(4): 307–314.
17. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. *J Health Econ* 1996; **15**(2): 209–231.
18. Dolan P, Sutton M. Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. *Soc Sci Med* 1997; **44**(10): 1519–1530.
19. Badia X, Monserrat S, Roset M, Herdman M. Feasibility, validity and test–retest reliability of scaling methods for health states: the visual analogue scale and the time trade-off. *Qual Life Res* 1999; **8**(4): 303–310.
20. Martin AJ, Glasziou PP, Simes RJ, Lumley T. A comparison of standard gamble, time trade-off, and adjusted time trade-off scores. *Int J Technol Assess Health Care* 2000; **16**(1): 137-147.
21. Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994; **39**(1): 7–21.
22. Stevens SS. On the psychophysical law. *Psychol Rev* 1957; **64**(3): 153–181.

23. Schoemaker PJH. The expected utility model: its variants, purposes, evidence and limitations. *J Econ Literature* 1982; **20**(2): 529–563.
24. Johannesson M, Pliskin JS, Weinstein MC. A note on QALYs, time tradeoff, and discounting. *Med Decis Making* 1994; **14**(2): 188–193.
25. Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? *Soc Sci Med* 1997; **45**(8): 1289–1297.
26. O'Leary JF, Fairclough DL, Jankowski MK, Weeks JC. Comparison of time trade-off utilities and rating scale values of cancer patients and their relatives: evidence for a possible plateau relationship. *Med Decis Making* 1995; **15**(2): 132–137.
27. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care* 1996; **34**(7): 702–722.
28. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med* 2001; **33**(5): 337–343.
29. Krabbe PF, Stouthard ME, Essink-Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. *J Clin Epidemiol* 1999; **52**(4): 293–301.
30. Mahapatra M, Nanda L, Rajashree KT. The 6D5L description system for health state valuation. In *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. Murray CJL, Salomon JA, Mathers CD, Lopez AD (eds). World Health Organization: Geneva, 2002.
31. Dolan P, Gudex C. Time preference, duration and health state valuations. *Health Econ* 1995; **4**(4): 289–299.
32. Bell DE, Raiffa H. Marginal value and intrinsic risk aversion. In *Decision Making: Descriptive, Normative and Prescriptive Interactions*, Bell DE, Raiffa H, Tversky A (eds). Cambridge University Press: Cambridge, MA, 1988.
33. Hadorn DC. Setting health care priorities in Oregon. Cost-effectiveness meets the rule of rescue. *JAMA* 1991; **265**(17): 2218–2225.
34. Van Der Pol MM, Cairns JA. Negative and zero time preference for health. *Health Econ* 2000; **9**(2): 171–175.
35. Ganiats TG, Carson RT, Hamm RM, Cantor SB, Sumner W, Spann SJ *et al*. Population-based time preferences for future health outcomes. *Med Decis Making* 2000; **20**(3): 263–270.
36. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health statistics measures: statistics and strategies for evaluation. *Controlled Clin Trials* 1991; **12**(Suppl. 4): 142S–158S.
37. Üstün TB, Rehm J, Chatterji S, Saxena S, Trotter R, Room R *et al*. Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *Lancet* 1999; **354**(9173): 111–115.
38. Jelsma J, Chivaura VG, Mhundwa K, De Weerdt W, de Cock P. The global burden of disease disability weights. *Lancet* 2000; **355**(9220): 2079–2080.
39. Üstün TB, Saxena S, Rehm J, Bickenbach J. Are disability weights universal? WHO/NIH Joint Project CAR Study Group. *Lancet* 1999; **354**(9186): 1306.
40. Baltussen RM, Sanon M, Sommerfeld J, Wurthwein R. Obtaining disability weights in rural Burkina Faso using a culturally adapted visual analogue scale. *Health Econ* 2002; **11**(2): 155–163.
41. Wakker P, Stiggelbout A. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Med Decis Making* 1995; **15**(2): 180–186.
42. Jansen SJ, Stiggelbout AM, Wakker PP, Vliet Vlieland TP, Leer JW, Nooy MA *et al*. Patients' utilities for cancer treatments: a study of the chained procedure for the standard gamble and time trade-off. *Med Decis Making* 1998; **18**(4): 391–399.
43. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Econ* 2002; **11**(5): 447–456.

