Chapter 9.4

Estimating health state valuations using a multiple-method protocol

Joshua A. Salomon and Christopher J.L. Murray

Introduction

In any summary measure of population health such as healthy life expectancy (WHO 2000), disability-adjusted life years (Murray 1996), quality-adjusted life expectancy (Kaplan and Erickson 2000) or health-adjusted life expectancy (Rosenberg et al. 1999), one essential data input is a set of valuations that assign weights to health states that are worse than ideal health. These health state valuations provide the critical link between information on mortality and information on non-fatal health outcomes. In order to serve as the basis for the combination of these two types of information, health state valuations must provide a cardinal measure of the value placed on time spent in a particular health state, relative to time spent in ideal or full health. It is important that any discussion of health state valuations in summary measures begins with a clear understanding of this requirement. As Essink-Bot and Bonsel point out in chapter 9.1, much of the confusion in the research on health state valuations has resulted from the failure to recognize that strategies for measuring these valuations may differ depending on their intended use.

A range of different methods for eliciting health state valuations have been proposed and used widely, including the standard gamble, time trade-off, visual analogue scale and person trade-off (Froberg and Kane 1989; Nord 1992). Thus far, there has been little agreement as to which method is most appropriate, stemming in part from the lack of clarity regarding the need to tailor the approach to the application. Arguments for and against different methods have been based on ethical grounds (Arnesen and Nord 1999), economic theory (Torrance 1976), and comparisons of psychometric properties (Krabbe et al. 1997). Many of the proponents of different approaches, however, have implicitly acknowledged that perhaps none of the available approaches is ideal.

In a number of empirical studies in which multiple methods have been used, the different methods have yielded different valuations for the same set of states (Dolan et al. 1996; Torrance 1986), although high correlations between the different measures have been observed. In some studies, mathematical relationships between responses on different types of valuation questions have been estimated (Dolan and Sutton 1997). This strategy is motivated by the assumption that one of the available methods must be selected for practical purposes, and it may be useful to develop a convenient function that would allow the transformation of valuations obtained using another method to the equivalent valuations that would be expected from the chosen method. Yet, if in fact none of the available measurement techniques provides the cardinal valuation measure that is required for summary measures of population health, it is worth considering whether transformations from one to another of these techniques is the most appropriate way to proceed.

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. By formalizing our understanding of how each of the valuation techniques captures other values or sources of bias in addition to the value of the health state itself, we aim to recover the underlying health value function from responses to four different types of valuation questions applied to a range of different health states.

This paper presents our approach to estimating health state values using a multiple-method exercise and describes the design, implementation and analysis of a first study of this approach.

Methods

Health state valuation exercise

A multiple-method health state valuation exercise was implemented among a convenience sample of 69 public health professionals from 28 different countries.

Twelve health states were selected to span a range of different severity levels. The states were described by brief labels and standardized descriptions of levels on six dimensions of health (mobility, self-care, usual activities, pain/discomfort, anxiety/depression and cognition) based on the modification of the EuroQol EQ-5D classification system that increases the number of levels in each domain from three to five and includes cognition as an additional domain (Brooks 1996; Krabbe et al. 1999).

The exercise consisted of 5 different tasks:

- ordinal ranking of the 12 states with the aid of index cards;
- valuation of the 12 states using a visual analogue scale (VAS) anchored by the best imaginable health state at 100 and death at 0, and with 100 equally-spaced tick marks, labelled at every even number;

- valuation of the 12 states using a time trade-off (TTO) question and self-completed worksheets, followed by the opportunity to examine and revise all 12 TTO valuations;
- valuation of the 12 states using a standard gamble (SG) question and self-completed worksheets, followed by the opportunity to examine and revise all 12 SG valuations; and
- valuation of the 12 health states using a person trade-off (PTO) question and self-completed worksheets, followed by the opportunity to examine and revise all 12 PTO valuations.

The format of each of the valuation tasks was as follows:

1. *Ordinal ranking*. Respondents were asked to consider each health state, imagining what it would be like for them to live in that health state. They were asked to assume that the life expectancy in each health state would be the same (10 years). They were then asked to rank them from the most desirable to the least desirable state.
2. *Visual analogue scale*. Respondents again were asked to imagine what it would be like for them to live in each of the health states for a duration of 10 years. The instructions reminded them to try to use the actual distances on the scale in a meaningful way, such that states that are similarly attractive would be placed close together while states that are very different would be placed far apart. They were asked to indicate the exact point on the scale where they would place each state, relative to the best imaginable health, death and all of the other states.
3. *Time trade-off*. Respondents were asked to imagine that they were living in the health state with a life expectancy of ten years, and faced a choice between 2 options: (1) to remain in that health state for the 10 remaining years of life; or (2) to be restored to perfect health but live for a shorter period of time. A worksheet for each health state guided the respondents through a series of trade-offs in which the number of years of shortened life expectancy in return for improved health was varied. The worksheet was designed to help respondents determine their indifference point, which consisted of the number of years in perfect health that would make the two options equally attractive.
4. *Standard gamble*. Respondents were again asked to imagine that they were living in the health state with a life expectancy of 10 years, but this time the choice was between: (1) remaining in that health state with 100% certainty for the 10 remaining years of life; or (2) accepting a risky procedure that offers some probability of being raised to perfect health for the remaining 10 years but also carries some risk of immediate death. A similar worksheet was used in which the respondents answered a number of different trade-offs with varying levels of risk,

in order to identify the indifference point where the risky option would be equally attractive as the certain option.

5. *Person trade-off*. Respondents were asked to imagine that they were decision-makers facing a difficult choice between two different programmes, with only enough money to fund one of them. The first programme would prevent the deaths of 100 perfectly healthy individuals, thereby extending their lives for 10 years, while the second programme would prevent the onset of some health problem in a certain number of healthy people, thereby improving their health expectancy from 10 years in a state worse than perfect health to 10 years lived in perfect health. A worksheet was provided to help respondents identify their indifference point, where the number of cases of prevented health problems would balance 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 time trade-off, standard gamble and person trade-off, once respondents had completed the exercise for all 12 states, they were presented with their responses for all 12 conditions and allowed to revise any of the values if they wished to do so.

Analysis

The goal of the analysis was to use the entire collection of responses for each individual to impute the health state values required in the construction of summary measures. In order to achieve this goal, the first step was to formalize the relationship between responses on each type of measurement method and the underlying values of interest using flexible parametric forms.

For each of the four measurement methods—visual analogue scale, time trade-off, standard gamble and person trade-off—we assumed that responses were described by an increasing function of the underlying value for the state (i.e. more severe states would be ranked as such by all methods), with specification of the functional form guided by relevant theoretical and empirical findings. We describe these functions briefly in this section and include complete mathematical details in the Technical Appendix.

For each measurement method, at least one auxiliary parameter was needed to describe the transformation from the underlying value function to the response function:

- Based on the common observation that the standard gamble reflects both strength of preference for a health state and an individual's attitude towards risk, we modelled standard gamble responses as a function of the underlying valuation and a risk aversion parameter. We examined several different formulations including exponential, logarithmic and power functions, based on the theoretical framework presented in Bell and Raiffa (1998).

- For the person trade-off, proponents have recognized that responses depend both on the level of health in a particular state and on distributional concerns (Arnesen and Nord 1999). In responding to person trade-off 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. We have therefore modelled the person trade-off responses using similar functions as those used for the standard gamble, but allowing for a distinct parameter to capture this "rule of rescue" (Hadorn 1991).
- For the visual analogue scale, 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 (Stevens 1957). We have based the model for the VAS responses on this finding, including a coefficient that determines the amount of curvature in the power function.
- For the time trade-off, responses will vary depending on the degree of time preference that individuals exhibit. If individuals have non-zero discount rates, then the two streams of life that are compared in the time trade-off (e.g. 10 years in state *X* and 5 years in perfect health) must first be translated into their equivalent present values in order to compute the implied health state value. We have assumed an exponential discounting model, with a single parameter to capture the discount rate.

Given the specification described above, the model included a total of 16 parameters of interest: 12 health state values, plus 4 auxiliary parameters.

We used maximum likelihood methods to estimate the parameters in the model. It was assumed that the stochastic component of the model followed a truncated normal distribution constrained between 0 and 1. Inspection of the distributions of responses on the different measures suggested strong heteroskedasticity, which was confirmed in regressions of the standard deviation of responses by the mean values for each method. We therefore specified the variance of the truncated normal distribution as a linear function of the expectation, and allowed the slope and intercept of the function to differ by valuation method.

In order to represent the uncertainty around the model estimates, we undertook numerical simulations of the results by sampling from the joint distribution of the estimated parameters obtained from the maximum likelihood estimation, and recalculating the quantities of interest for each set of sampled parameters. This approach allowed us to develop bounds around the estimated strength of preference values and auxiliary parameters in the model that reflected important estimation uncertainties.

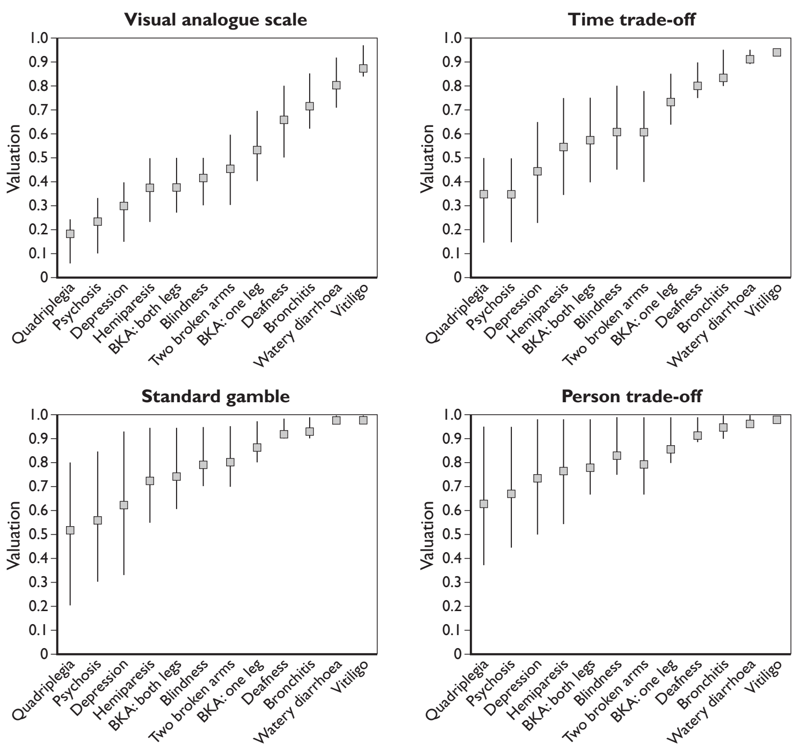
Results

Responses from the four different measurement methods are summarized in Figure 1.

Overall, the visual analogue scale tends to give the lowest values for a given health state, with the smallest variance across respondents. Time trade-off values tend to be slightly higher than VAS values, followed by standard gamble and finally, person trade-off. For severe states, the standard gamble and person trade-off methods both produce considerably higher variance across respondents than either the time trade-off or visual analogue scale.

Figure 2 plots the mean valuations against the standard deviations for each of the different methods. As this figure illustrates, there are systematic differences in the standard deviations for the four different methods. Furthermore, the standard deviations appear to be strongly associated with the mean level, especially for the time trade-off, standard gamble and

Figure 1 Mean response and interquartile range across 69 respondents for 12 states and 4 valuation methods



BKA: Below the knee amputation.

person trade-off. This relationship formed the basis for the heteroskedastic formulation of the stochastic model in the estimation procedure.

Table 1 lists the estimated severities for the 12 states in this study, along with the approximate 95% confidence interval for each estimate. The rank order of the estimated severities is consistent with the rankings from the four measurement methods. Because the methods have distinguished the effects of risk aversion and equity concerns from the severity of the health state, mild health states have a lower rating than the valuations in previ-

Figure 2 Mean valuations and standard deviations by method and state

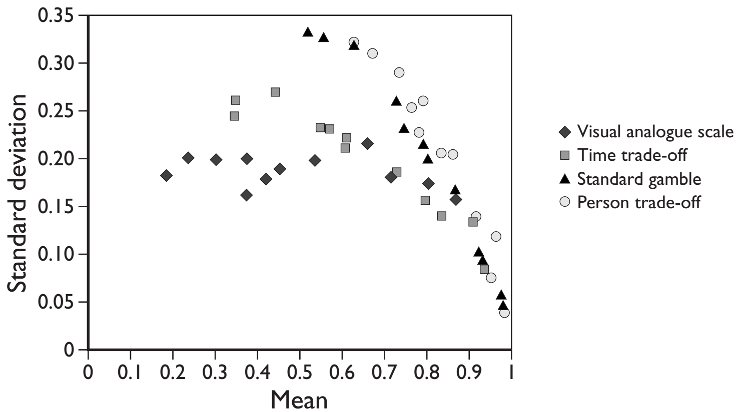


Table 1 Estimated health state valuations and ranges based on multiple-method protocol using visual analogue scale, time trade-off, standard gamble and person trade-off techniques

State	*Severity*	*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
Blindness	0.50	0.40–0.59
Two broken arms	0.49	0.41–0.59
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.

ous studies (Murray 1996). This has important implications for the economic analysis of preventive and curative health interventions.

The estimated values for the auxiliary parameters (Table 2) imply that the respondents are strongly risk averse and have preferences 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. The unusual finding of a negative discount rate would imply that individuals consider a unit of health in the future as more valuable than a unit of health today. While this finding has been demonstrated in other empirical studies of individual discount rates (Dolan and Gudex 1995; Ganiats et al. 2000; van der Pol and Cairns 2000), it runs against conventional health economics wisdom. It would be possible to add further constraints to the model, for example that only non-negative time preference would be allowed, but the results presented here are for the unconstrained model. Consideration of alternative functional forms for each of the different measurement methods remains an important area for further research.

Based on the maximum likelihood estimates of the underlying health state valuations for the 12 states and the auxiliary parameters, we have estimated predicted responses for each of the 12 states using the four different methods. The predicted responses fit the observed distributions of responses quite closely (Figure 3).

Discussion

In this paper, we have demonstrated that it is possible to explain responses to the standard gamble, time trade-off, person trade-off and visual analogue scale based on a consistent set of health state valuations for a range of states. None of these methods provides a pure measure of strength of preference, but we may explicitly model the process by which individuals respond to different types of measurement techniques given the underlying valuation of a health state. The multiple-methods approach presented here also allows the measurement of levels of risk aversion, time prefer-

Table 2 Maximum likelihood estimates of auxiliary parameters[a]

State	*Estimate*	*Standard error*
Risk attitude (SG)[b]	2.6	0.43
Distributional concerns (PTO)	2.9	0.44
Discount rate (TTO)	–0.089	0.039
Scale distortion (VAS)[c]	0.83	0.10

a. See Appendix for complete description of parameters.

b. A number > 0 indicates risk aversion. The parameter for distributional concerns in the PTO has a similar interpretation.

c. A number < 1 indicates a convex curve.

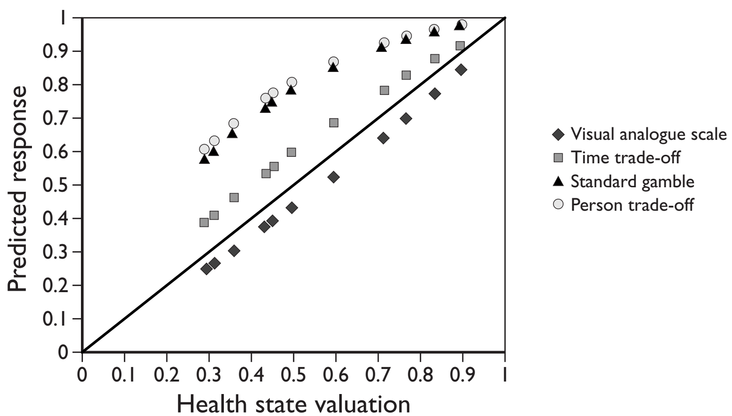
ence, visual analogue scale distortion and distributional concerns in a group of respondents. One important benefit of this approach is that it will facilitate the mapping between different measurements and the underlying health state valuations that may be needed for comparison of different studies.

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 (e.g. Üstün et al. 1999). 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 description; differences in risk aversion, time preference, distributional concerns or visual analogue scale distortion; or differences in the valuation 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 standardized 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.

Using a multiple-methods 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 the value assigned to a health state. This will require larger datasets and the elaboration of the statisti-

Figure 3 Underlying health state valuations and predicted responses for 12 states using 4 different valuation methods



cal model used here to allow for variation in the health state values and auxiliary parameters. One implication is that 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 health state valuation itself.

A number of limitations in this preliminary study are worth noting. First, it is important to recognize that different results might be obtained depending on the functional form of the models that are used. We have selected models based on previous theoretical and empirical work, but other plausible alternatives should be considered. As other model formulations are explored, it will be necessary to examine the sensitivity of the results to the choice of models. The model used for the time trade-off may require particularly careful inspection. Other methodological advances may be fruitful, for example, in using Bayesian statistical methods for incorporating additional prior information into the estimation framework. The nature of measurement error in the application of these methods also merits further examination. While the truncated normal distribution 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 valuations 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.

Technical appendix

Each measurement technique produces a response for each state, x, on a scale particular to that method:

Method	*Response*	*Units and Scale*	*Interpretation*
VAS	s	0 to 100	Rating of health state x
TTO	y	years, 0 to 10	Years of perfect health equivalent to 10 years in state x
SG	p	risk, 0 to 100%	Risk of death at which treatment is equivalent to certainty in state x
PTO	n	persons, 100 to ∞[a]	Number of averted cases of x equivalent to 100 deaths averted

a. In principle, n may be less than 100, which would imply that preventing cases of health state x is preferred to preventing deaths of individuals in ideal health. In practice, all respondents indicated values greater than 100 for all states.

These responses may be rescaled such that each one ranges from 0 to 1, with 1 being the highest valuation:

$$SG_x = 1 - p$$

$$PTO_x = 1 - \frac{100}{n}$$

$$VAS_x = \frac{s}{100}$$

$$TTO_x = \frac{y}{10}$$

We then assume that each one of these rescaled responses is an increasing function of the underlying value for the health state v_x and an auxiliary parameter.

The standard gamble has one parameter θ_1 that represents an individual's risk aversion. The formulation is derived from utility theory, as described by Bell and Raiffa (1998).

$$SG_x = \frac{-e^{-\theta_1 v_x} + 1}{-e^{-\theta_1} + 1}$$

The person trade-off formulation is parallel to the standard gamble formulation, but in this case the parameter θ_2 represents aversion to decisions resulting in loss of life, the so-called "rule of rescue" (Hadorn 1991).

$$PTO_x = \frac{-e^{-\theta_2 v_x} + 1}{-e^{-\theta_2} + 1}$$

The visual analogue scale is a power function with one parameter θ_3. This formulation is based on results from psychophysics experiments (Stevens 1957) and has been suggested by Torrance (1976) in modelling the functional relationship between VAS and SG.

$$VAS_x = 1 - [1 - v_x]^{\theta_3}$$

The function for the time trade-off is derived from the following relation, in which θ_4 characterizes an individual's rate of time preference:

$$v_x = \frac{\frac{1}{\theta_4} - \frac{1}{\theta_4}\left(e^{-10\theta_4 TTO_x}\right)}{\frac{1}{\theta_4} - \frac{1}{\theta_4}\left(e^{-10\theta_4}\right)}$$

If people have some rate of time preference, then both alternatives in the TTO should be converted to their present values, so the formula for discounting a continuous stream of life is applied. The function for the time trade-off simply solves the above equation for TTO_x.

$$TTO_x = -\frac{1}{10\theta_4} \ln\left[1 - (1 - e^{-10\theta_4})v_x\right]$$

References

Arnesen T, Nord E (1999) The value of DALY life: problems with ethics and validity of disability adjusted life years. *British Medical Journal,* **319**(7222):1423–1425.

Bell DE, Raiffa H (1998) 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.

Brooks R (1996) EuroQol: the current state of play. *Health Policy,* **37**(1):53–72.

Dolan P, Gudex C (1995) Time preference, duration and health state valuations. *Journal of Health Economics,* **4**(4):289–299.

Dolan P, Gudex C, Kind P, Williams A (1996) Valuing health states: a comparison of methods. *Journal of Health Economics,* **15**(2):209–231.

Dolan P, Sutton M (1997) Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. *Social Science and Medicine,* **44**(10):1519–1530.

Froberg DG, Kane RL (1989) Methodology for measuring health-state preferences, part – II: scaling methods. *Journal of Clinical Epidemiology,* **42**(5):459–471.

Ganiats TG, Carson RT, Hamm RM, Cantor SB, et al. (2000) Population-based time preferences for future health outcomes. *Medical Decision Making,* **20**(3):263–270.

Hadorn DC (1991) Setting health care priorities in Oregon. Cost-effectiveness meets the rule of rescue. *Journal of the American Medical Association,* **265**(17):2218–2225.

Kaplan RM, Erickson P (2000) Gender differences in quality-adjusted survival using a health utilities index. *American Journal of Preventive Medicine,* **18**(1):77–82.

Krabbe PF, Stouthard MEA, Essink-Bot ML, Bonsel GJ (1999) The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system. *Journal of Clinical Epidemiology,* **52**(4):293–301.

Krabbe PFM, Essink-Bot ML, Bonsel GJ (1997) The comparability and reliability of five health-state valuation methods. *Social Science and Medicine,* **45**(11):1641–1652.

Murray CJL (1996) Rethinking DALYs. In: *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. The Global Burden of Disease and Injury, Vol. 1. Murray CJL, Lopez AD, eds. Harvard School of Public Health on behalf of WHO, Cambridge, MA.

Murray CJL, Salomon JA, Mathers CD (2000) A critical examination of summary measures of population health. *Bulletin of the World Health Organization,* **78**(8):981–994.

Nord EM (1992) Methods for quality adjustment of life years. *Social Science and Medicine,* **34**(5):559–569.

Read JL, Quinn RJ, Berrick DM, Fineberg HV, Weinstein MC (1984) Preferences for health outcomes: comparison of assessment methods. *Medical Decision Making,* **4**(3):315–329.

Rosenberg MA, Fryback DG, Lawrence WF (1999) Computing population-based estimates of health-adjusted life expectancy. *Medical Decision Making,* **19**(1):90–97.

Stevens SS (1957) On the psychophysical law. *Psychological Review,* **64**(3):153–181.

Torrance GW (1976) Toward a utility theory foundation for health status index models. *Health Services Research,* **11**(4):349–369.

Torrance GW (1986) Measurement of health state utilities for economic appraisal: a review. *Journal of Health Economics,* **5**(1):1–30.

Üstün TB, Rehm J, Chatterji S, et al. (1999) Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH joint project CAR study group. *Lancet,* **354**(9173):111–115.

van der Pol MM, Cairns JA (2000) Negative and zero time preference for health. *Health Economics,* **9**(2):171–175.

Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russel LB (1996) Recommendations of the panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association,* **276**(15):1253–1258.

WHO (2000) *The World Health Report 2000. Health systems: improving performance*. World Health Organization, Geneva.