HEALTH ECONOMICS
Health Econ. **15**: 653–664 (2006)
Published online 24 February 2006 in Wiley InterScience (www.interscience.wiley.com). **DOI:**10.1002/hec.1086

Is there a case for using visual analogue scale valuations in cost-utility analysis?

David Parkin* and Nancy Devlin
City Health Economics Centre, City University, London, UK

Summary

This paper critically reviews theoretical and empirical propositions regarding visual analogue scale (VAS) valuations of health states and their use in cost-utility analysis (CUA). A widely repeated assertion in the economic evaluation literature is the inferiority, on theoretical grounds, of VAS valuations. Five common criticisms are: VAS lacks a theoretical foundation; VAS values are not 'choice based'; VAS values are not consistent with utility-under-uncertainty requirements; context and range effects observed in VAS valuation data mean that they cannot even be considered to represent measurable value functions; and when completing a VAS, people are not trying to express values.

We address each of these points: the VAS does have a theoretical basis, being entirely consistent with the non-welfarist foundations of QALYs and CUA; the 'choiceless' nature of the VAS is incorrectly judged by stated preference criteria relevant to monetary rather than health state valuations, and VAS valuations do in any case involve an element of choice; because valuations are intended for use in social decision-making, it may be advantageous that VAS values are elicited under conditions of certainty; although there are measurement problems with the VAS, means such as better design and transformations of data can deal with these; and with any method of eliciting values, it is unrealistic to expect people consciously to think in terms of social science constructs such as utilities.

Moreover, there are problems, both theoretical and empirical, with alternative methods. Selection of the appropriate valuation method should be based on empirical performance, and in this the VAS has important advantages. We conclude that there are strong grounds for disputing the consensus view against the VAS and challenge those who hold it to deploy more convincing arguments and evidence in favour of alternative methods. However, we identify areas where further research is required to establish and consolidate the potential of the VAS as a valuation method. Copyright © 2006 John Wiley & Sons, Ltd.

Keywords visual analogue scale; cost-utility analysis; utility measurement; health state valuation

Introduction

Visual analogue scales (VAS) and other types of rating or category scales are a very common means of measuring both individuals' rating of their own health, and their preferences for other, hypothetical health states expressed as health state scenarios. A key issue regarding VAS, which is the focus of this paper, is the appropriateness of applying the resulting health state scenario values to the estimation of health gain in economic evaluation studies. An increasingly strongly asserted conclusion in the economic evaluation literature is the inferiority, on theoretical grounds, of direct methods of eliciting health state valuations, such as the VAS.

*Correspondence to: City Health Economics Centre, Department of Economics, City University, Northampton Square, London, EC1 V OHB, UK. E-mail: d.parkin@city.ac.uk

Received 25 March 2004
Accepted 25 October 2005

This paper critically assesses both theoretical and empirical propositions regarding VAS valuation. In the following section, we address a number of underlying theoretical issues regarding economic evaluation and the nature of health state valuations. We then set out the theoretical case against VAS and argue that there are grounds for challenging it. We also consider empirical issues and argue that both well-known and newly established properties of population VAS data confer important advantages that should not always be traded off against the alleged theoretical merit of its alternatives. Torrance *et al*. [1] have reported some similar analyses and conclusions; however, we place these in a rather more positive light and draw somewhat different conclusions about the role of the VAS in CUA.

To illustrate some of our arguments, we will refer to the EuroQol EQ-5D, which is a health related quality of life instrument that incorporates both a descriptive system and a VAS [2]. However, it should be emphasised that this is used only for exposition purposes, and our analysis and conclusions are relevant to other classifications and VAS scales. The EQ-5D classification, detailed in Figure 1, describes health states according to five dimensions. Within each dimension, the intensity of ill-health is classified according to three simple descriptors, which essentially record 'no problems', 'some problems' or 'severe problems'. The EQ-5D self-classification questionnaire incorporates an instrument for respondents to record their current health state according to this scheme and a VAS, known as the EQ-VAS, to record their current health state; however, there is also a standard EQ-5D valuation questionnaire that also includes an instrument, known as the EQ-5D VAS, of the same design as the EQ-VAS. This is described in the following section.

The use of VAS valuations in cost-effectiveness and cost-utility analysis

In this section, we discuss some key issues relevant to our argument – the nature of VAS; the inter-

Mobility
1. No problems in walking about
2. Some problems in walking about
3. Confined to bed

Self-care:
1. No problems with self-care
2. Some problems washing or dressing self
3. Unable to wash or dress self

Usual activities
1. No problems with performing usual activities (eg work, study, housework)
2. Some problems with performing usual activities
3. Unable to perform usual activities

Pain and discomfort
1. No pain or discomfort
2. Moderate pain or discomfort
3. Extreme pain or discomfort

Anxiety and depression
1. Not anxious or depressed
2. Moderately anxious or depressed
3. Extremely anxious or depressed

Each of the 243 possible health states can be uniquely identified by a five digit number, for example "12212" would mean:

No problems in walking about
Some problems washing or dressing self
Some problems with performing usual activities
No pain or discomfort
Moderately anxious or depressed

Figure 1. The EQ-5D health related quality of life classification

pretation, in terms of value, of measurements derived from VAS; the nature of health gain measures derived from health state valuations; and the use of health gain measures in economic evaluation.

Rating scales, category scales and visual analogue scales

The terms *rating scale*, *category scale* and VAS are often used in the literature interchangeably. It seems logical to call the VAS a category scale when it is used as a measurement instrument comparable to a Likert scale, with numbers replacing verbal descriptors; and to call the VAS a rating scale when it is used to derive preference weights, as an alternative to methods such as paired comparisons, magnitude estimation, time trade-off and standard gamble. However, this conclusion is not derived from the literature, which is unclear and inconsistent about terminology. In what follows, we are specifically concerned with the VAS, but some of the analysis and conclusions derive from and relate to the wider class of instruments represented by rating and category scales.

A VAS usually consists of a single line on a page with verbal and numerical descriptors at each end. Scale markers are often added to the line, and these are sometimes also numbered. The EQ-5D VAS, for example, is by convention a 20 cm thermometer-like vertical line with endpoints labelled 'best imaginable health state possible' and 'worst imaginable health state possible', denoted as 100 and 0 respectively, and is by convention demarcated in units of one and labelled in units of 10.

In an exercise to value health state scenarios, participants are presented with a set of health states and are asked to rate the desirability of each by placing it at some point on the line on or between these two endpoints. This procedure is generally considered to be capable of providing an interval scale measure of preferences such that '...if a state Q^* is placed mid-way between two states Q and Q^{**}, this is supposed to represent the fact that the respondent regards being in state Q as better than being in state Q^* to the same extent that being in state Q^* is better than being in state Q^{**}', [3] therefore capturing the strength of an individual's preferences over the set of states.

Measurable value functions, utility functions and uncertainty

The theoretical properties of rating scales as preference measures are based on the axiomatic approach outlined by Dyer and Sarin [4,5]. Rating scale valuations potentially belong to a class of value functions known as measurable value functions. Such functions describe values under certainty; they have properties of both correct ranking of preferences and measuring strength of preferences. However, except under conditions where people are risk-neutral, VAS valuations should not, according to this taxonomy, be called utilities. It should be noted that this also applies to TTO valuations, which are also elicited under conditions of certainty. Utility functions also correctly rank preferences and are cardinal but have the additional property of measuring values under uncertainty; in other words, they measure *certainty equivalent values* for uncertain outcomes.

This distinction is important in decision-making where outcomes are uncertain. For example, suppose that we have VAS or TTO valuations that a person agrees measure their relative values for EQ-5D health states. Compared with state 11111, fixed at a value of 1, they value state 21111 at 0.90 and state 23322 at 0.02. They are then faced with a decision between a certain outcome (their current health state) of state 21111 and an uncertain outcome (the result of a treatment) of state 23322 with a probability of 0.1, and 11111 with probability of 0.9. The expected value of the uncertain outcome is 0.902, which is higher than the certain outcome. However, it is quite possible that many people would regard a 10% probability of such a serious outcome as too high when a successful outcome is simply to remove problems in walking about. For such risk-averse people, the certainty equivalent of the uncertain outcome would be much lower and they would prefer the certain outcome. The standard gamble technique attempts to derive utilities that give the correct certainty equivalent values to uncertain outcomes.

The argument is, therefore, made that measurable values – and therefore the use of the VAS and TTO – are inappropriate in health care, which is characterised by a high degree of uncertainty. This argument is not compromised by positive considerations of whether people in fact behave as expected utility maximisers – although there is ample evidence that they do not. Neither is it

based on a normative proposition that they ought to maximise utility – although a recommendation that they should do so is valid only where the same individual will take the decision many times and the stakes are small, whereas in health the decisions are often one-off and have serious consequences [6]. The argument's basis is simply that people may not be risk-neutral about health care and that fact ought to be recognised in social decision making about it.

Health-adjusted life years and quality-adjusted life years

Gold *et al*. [7] recently coined the term 'Health Adjusted Life Years' to refer to a 'family' of health measures, the family members being quality adjusted life years (QALYs) and disability adjusted life years (DALYs). HALYs are 'summary measures of population health that allow the combined impact of death and morbidity to be considered simultaneously'. QALYs are specifically stated to be utility based, though the terms utility, value, preference and weights are used interchangeably. The distinction is a valuable one, though unfortunately their historical review omits reference to the key early 1970s European contribution to the development of QALYs by Culyer *et al*. [8].

Culyer *et al*. outlined a measure of health based on the product of 'intensity of ill-health' and 'duration'. Health gains from any health care intervention are measured by the change in this measure that the intervention generates. The units for this measure were not given any label. However, it is apparent that the units are in fact QALYs; this is consistent with one of the earliest writings on this topic; in 1968, Klarman *et al*. [9] calculated what they termed 'quality-adjusted life expectancy' based on quality adjustment weights.

Subsequently, in 1977, an influential article by Weinstein and Stason [10] connected QALYs with utilities, specifically expected utility, rather than the 'weights' of the earlier literature; and this connection has remained. However, we are reluctant to concede the term 'quality' to refer only to expected utility-based measures; perhaps we should instead have subsets such as utility-based QALYs (U-QALYs) and value based QALYS (V-QALYs).

Cost-effectiveness and cost-utility analysis

Leaving aside cost-benefit analysis and other, lesser-used, techniques, economic evaluation in health care consists of cost-effectiveness (CEA) and cost-utility (CUA). Both measure cost compared with output, but differ in how output is measured: 'physical' quantities or 'natural units' (CEA) and health related quality of life, particularly QALYs gained (CUA). A common, but less well-articulated view is that CEA provides evidence about technical efficiency and CUA about allocative efficiency. However, as normally practised in health economics, CEA does not provide evidence on technical efficiency – the relationship between physical inputs and outputs – and CUA does not provide full information on allocative efficiency, of the kind that cost-benefit analysis would provide. Instead, both usually produce estimates of the observed cost of achieving different levels of output. The difference between the two is that CEA relates to the output of particular types of health care and CUA to the output of health care as a whole.

This definition of CUA as the calculation of costs per QALY gained is the result of a set of influential articles in the 1980s, of which an important example is by Boyle *et al*. [11]. However, this was neither a neologism nor an inevitable choice. The term already existed; for example, in 1972 Fisher [12] suggested it as a generic term encompassing all kinds of economic evaluation rather than a specific term referring to the calculation of a ratio of cost to expected utility based measures of gain.

These authors, in particular Torrance [13], did not restrict CUA to include only what we termed above U-QALYs. However, inclusion of the word 'utility' has increasingly led critics to believe that any CUA must involve a strictly defined utility base. Although this semantic argument is reasonable, it would leave no distinctive term for a cost-per-QALY evaluation based on V-QALYs. This is important because a V-QALY analysis is closer to narrowly defined CUA than to CEA, where measures are in 'physical quantities' or 'natural units'. Kind [14] has argued that VAS valuations are in fact 'natural units', and this is a plausible argument for the use of patients' self-rated VAS scores of their own health states in CEA. However, when VAS valuations of health state scenarios are used, the resulting V-QALYs have a radically different interpretation that should accord them a

special status in CEA, equivalent to the label CUA used for U-QALYs.

The case against visual analogue scales and responses to it

Below, we present some quotations that exemplify the common objections to the use of VAS valuations in economic evaluation. We then put the case for the defence for VAS against the charges that it faces.

> 1 VAS 'lacks a theoretical foundation and cannot be related to the underlying theory of QALYs' (Johannesson *et al*. [15]).

VAS does of course have an underlying theoretical foundation. It has its foundations in psychological theories of response to sensory stimuli and has a long history in psychometric research. Nord [16] for example, suggests that one of the attractions of VAS is that it 'can be related to an established body of measurement theory'. What Johannesson *et al*. probably intended is not that VAS lacks theoretical foundations, but rather that these are not *economic* in origin. However, the theory of measurable value functions provides an adequate theoretical base for its use in economic evaluation. Whether or not it does in fact conform to such functions is discussed below.

The question of whether or not VAS relates to the underlying theory of QALYs depends on what we believe the underlying theory of QALYs to be. Of course, if QALYs are defined as U-QALYs, then VAS by definition may not provide the correct values; the same is true for all techniques apart from SG. If the suggestion is that the valuation method should measure QALYs directly, rather than health states, then VAS health state values are again by definition not appropriate; again, the same is true for all methods except perhaps TTO and HYEs.[1] However, if QALYs are a more general concept, then VAS values are indeed related to the theory of QALYs.

As described above, QALYs were initially developed as a pragmatic alternative to needs assessment for the purposes of health policy making, and as means of facilitating more extensive scope in comparisons of value for money than the more restrictive outcome measures typically used in cost effectiveness analysis. Early descriptions of QALY-like measures refer not to individuals' utilities, but to 'weights', which might be established in any number of ways, including by decision-makers themselves. The quality-adjusted life expectancy calculated by Klarman was based on quality adjustment weights that were not derived from utilities and elsewhere he explicitly stated that the use of QALYs was intended as a non-monetary numeraire for cost-effectiveness analysis [17]. Similarly, Culyer, Lavers and Williams stated that (italics in original)

> Since it is intended to use these numbers as *weights*, and not simply as *rankings*, it is important to stress that society's judgements concerning the relative importance of avoiding one state rather than another are represented by the actual numbers attached to each respectively...This implication must not be shirked, and must be regarded as a statement about *health policy* (and is to be made by whoever is entrusted with that responsibility – e.g. 'the Minister') [8].

In 1976, Culyer [18] added the suggestion that the 'value judgements' (note, still not 'utilities' or 'valuations') upon which QALYs might be estimated could be sought from patients and populations, but continued to note that '...this value judgement is also, essentially, a planning matter about policy in the NHS and is again appropriately taken by publicly accountable planners'.

Subsequently, the use of QALYs in economic evaluation has come to be grounded in the theory of extra welfarism. Beginning with Sen [19] and first articulated specifically in relation to the use of QALYs by Culyer [20], extra welfarism rejects the idea that the social welfare function should be based only on individuals' utilities, instead allowing for something other than utility to be maximised. The use of QALYs as the measure of benefit in the economic evaluation of health care programmes relies on an external ('decision-maker's') judgement that, in the allocation of limited health budgets, improvements in health, rather than utility, are the appropriate maximand. Viewed in this way, the QALY is no more than a convenient device to combine both length and quality of life into a single metric of health, which replaces utility as the objective function to be maximised. The theory of QALYs and CUA therefore does *not* require that the valuation of quality of life has its roots in utility theory: indeed, the very use of CUA could be argued to imply a rejection of measures of utility as the sole basis for

social choices. Thus, although the measurement of benefit in cost-benefit analysis, in the welfarist tradition, shares many of the same theoretical roots as attempts to measure utility in CUA [21], individual utility maximisation is *not* a theoretical requirement of the latter.

Moreover, while welfare economics' concern is with changes to affected individuals' utility arising from alternative 'states of the world', the application of extra welfarism via CUA is not individualistic. The values of specific individuals who gain or lose QALYs are not usually considered in CUA. They are replaced by mean or median population values, applied uniformly across all affected people, *regardless of their individual valuations* [22].

Thus, the theoretical foundations of QALYs provide extremely limited support indeed for the idea that the valuation of quality of life should conform to any particular measurement method, or that U-QALYs are superior to V-QALYs.

> 2 VAS 'involves no choice, so it is not possible to observe any trade-off' (Johannesson *et al*. [15]).
>
> VAS methods '...do not present a choice, and are therefore thought to be unable to measure strength of preference on a cardinal scale. Due to the lack of choice and the absence of opportunity cost in the VAS task, one common view is that they have no basis in either economic or decision theory' (Brazier *et al*. [23]).

A widespread view amongst health economists is that 'choice-based' valuation techniques are based on economic theory, but 'choiceless' techniques are not.[2] However, the justification for this is not clear. As a view based on scientific principles, rather than simple prejudice, it may derive from consideration of principles that are relevant to valuation of goods and services using stated preference (SP) techniques.

SP refers to valuations of goods and services derived from experimental or survey data. Its justification is the assumption that it accurately mimics what would happen if the valuations were derived from real choices. Its basis in economic theory is therefore derived from the theory of revealed preference (RP): real choices are based on comparing benefits with opportunity costs, or what has to be sacrificed to obtain them; therefore observing real choices reveals real preferences. However, the theoretical base for the assumption that SP mimics RP is psychological and the relevant measurement theory is psychometric.

Guidelines for SP studies (for example, [24]) therefore emphasise that valuation of goods and services should, amongst other things, be based either on *choice modelling*, if we are interested in people's preferences for attributes of goods and services, or *contingent valuation*, if we are interested in the good or service as a whole. Techniques that do not involve choice are regarded as doubtful with respect to their consistency with welfare economics theory.[3]

However, such guidelines refer to the attempt to obtain *monetary* values for *goods and services*. There is no obvious reason why they should apply to the attempt to obtain non-monetary values such as utilities; utility theory itself is not based on choice. A more useful way of distinguishing between 'choice-based' and 'choiceless' techniques is to regard them as 'indirect' and 'direct' measurement of utility, respectively. There is no *economic* theory that supports indirect rather than direct utility measurement; indeed it is arguable that there is no well-founded *economic* theory of utility *measurement* within the domain of SP. There may be good reasons for supporting one technique rather another, but the presence or absence of choice is not, of necessity, one of them. The SG technique, for example, was not motivated by the desire to force people to make choices, rather it is an attempt to ensure that inferred utilities conform to the set of axioms which underlie von-Neumann–Morgenstern utilities.

Moreover, the dichotomy between 'choice-based' and 'choiceless' techniques is not as clear-cut as it appears. The notion that VAS involves no choice is not true. In the most trivial sense, the person using the scale makes a choice as to the point chosen on the line. This is not a choice between two alternatives, but it is nevertheless a choice. Moreover, it could be argued that when VAS valuations are sought for *sets* of states, this is less restrictive and more natural than the choice between two alternatives, as is the case for the TTO and the SG.

Less trivially, VAS valuations involve respondents weighing up pairs of health state scenarios, each of which embodies specific variants of health related dimensions and levels. For example, the 'standard' set of EQ-5D health states typically valued in VAS valuation questionnaires includes the states 11211 and 11121. The varying dimensions are usual activities (*some problems* with performing usual activities in the former state, *no problems* in the latter) and pain/discomfort

(*moderate* pain/discomfort in the latter state, *no problems* in the former). Respondents are asked to value these two states compared to each other, the VAS 'anchors' (best and worst imaginable health), six other states, including the 'marker' states 11111 and 33333 and, subsequently, the state *dead*. The respondents' valuation of 11211 and 11121 involves an observable trade-off: does an improvement in one dimension, and a corresponding worsening in another, lead, *ceteris paribus*, to a scenario which is preferred, equivalent to, or less preferred than its comparator? Figure 2 illustrates this trade-off, using the form of exposition first used by Culyer *et al*. [8] and assuming that all other EQ-5D dimensions are held constant at level 1. The scores shown are VAS values estimated from a sample of the New Zealand population [25]. Contour lines connect the points that are considered equally bad; in this example, the scores for each contour line show that the valuations placed on each state are more strongly affected by decrements in pain than in usual activities (thus 11211 > 11121, and 11311 > 11131).

In welfare economics terms, the advantage claimed for SG and TTO is that the welfare change associated with a change in health status can be determined by identifying the compensating change in remaining arguments in the individuals' utility function – risk for SG, and longevity for TTO – that would be required to leave utility unchanged [26]. However, if the dimensions in the health state descriptive system can be treated as *separate* arguments in the individual's utility function, the act of valuing states using the VAS has a similar interpretation and could therefore be argued to involve both choice and trade-off. Moreover, these trade-offs across health state dimensions are far more transparent to participants in VAS tasks than in SG and TTO, where, other than for chronic states worse than dead, each state is valued relative not to other health states which have different combinations of these dimensions, but only to perfect health or perfect health and death.[4] For the same reason, the comparisons made using the VAS are more congruent with the purpose of the exercise, which is to value health states relative to each other.

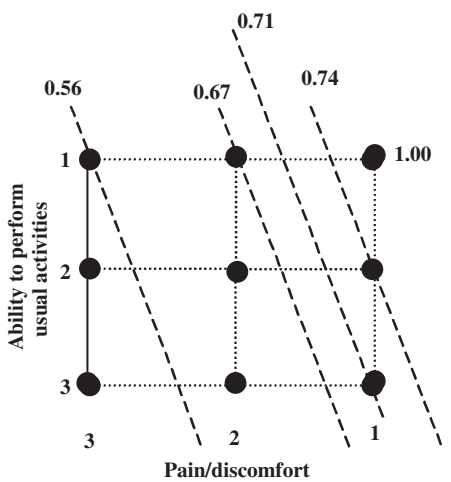

Figure 2. Trade-offs between states comprising two dimensions ('usual activities' and 'pain/discomfort')

Furthermore, any advantage conferred by indirect valuation approaches comes at the expense of the possibility that values may be contaminated by the *numeraire* against which trade-offs are sought. TTO, for example, is known to be confounded by time preference [27] and relies upon the assumption of constant proportionality of utility to time, violations of which include evidence of 'maximal endurable time' [28]. Similarly, SG is confounded by attitudes toward risk [29] and, as we argue in the next section of this paper, there may be decision-making contexts in which it is actually inappropriate to include risk.

3 VAS values '...are only appropriate for problems that involve certainty; thus, values are much more restricted in their applicability' (Drummond *et al*. [30]).

As already noted, this same criticism can be made of TTO, which also seeks valuations under conditions of certainty and, as with VAS, falls into the class of 'measurable value functions'. This is therefore an argument that favours SG only, as it alone satisfies the utility-under-uncertainty requirement of expected utility theory.

Notably, Drummond *et al*. go on to state that '...these theoretical arguments (in favour of SG) *are only valid at the individual level*. Von Neumann-Morgenstern utility theory only covers individual decision making, and once we aggregate the utilities across the respondents and use the results to inform societal decision-making, *the theory no longer directly applies*' (italics and clarification in parentheses added) [30].

A simple demonstration of this can be given using the earlier example of a treatment with an uncertain outcome, 11111 with 0.9 probability and 23332 with probability 0.1. Suppose that the

decision concerned whether or not to offer this treatment to 1000 people. The outcome, assuming that all took up the offer, would be 900 people at 11111 and 100 people at 23322. The problem changes from a *risk* of adverse outcomes, as viewed by individuals, to a *distribution* of adverse outcomes at the societal level. It is not obvious what a risk attitude means for a decision maker at this aggregate level; the certainty equivalent may therefore simply be the expected value of these two certain outcomes, and it may additionally be that on normative grounds society would wish it to be so.

A similar argument can be traced to Broome [31], whose starting point is that QALYs are intended to be a measure of 'benefit', or 'good'. He notes that 'an action whose results are uncertain should be valued by first fixing a value on each of its possible results, and then following the recommendations of expected utility theory' but that 'when it comes to 'social' valuations, involving the good of more than one person, expected utility theory is controversial. In particular, it prevents one from giving value to equality in the distribution of risk between people, and on the face of it that seems unreasonable'.

Reiterating the issue discussed above regarding the extra welfarist foundations of CUA, Broome notes:

> 'Using QALYs does not commit one to a narrow...conception of good. QALY analysis assigns values to states of health, and leaves it open whether these values are determined by how people feel when they are in these states, by their preferences about them, or perhaps by some objective principles. All of these possibilities are consistent with the general idea...that QALYs are aimed at assessing good or benefit'. [31]

Broome then goes on to argue that, given this is the case, 'if the adjustment factors are to be severed from preferences, then neither the time method nor the probability method can determine them'. Broome describes the VAS approach employed by Torrance *et al*. [32], and notes: 'If questions like this elicit sensible answers, they will do so whether or not the subject discounts risk or is risk neutral. In some ways, therefore, this could be a more reliable way of estimating the adjustment factors than either the time method or the probability method'. In essence, Broome is arguing that compliance with utility theory is not requisite to the estimation of QALYs and that the choice of valuation approach therefore rests on empirical performance.

Finally, the proof that VAS valuations produce values rather than utilities is not as certain as it is portrayed. There is no theory that proves that this must be the case. The evidence is largely based on observed differences between VAS valuations and those derived from other measurement methods, particularly standard gamble. However, there are so many differences between methods – and no homogeneity within them – that it is by no means certain that such differences represent the difference between values and utilities. As a simple example, asking people to contemplate 'being dead', which is what the EQ-5D VAS typically does, is not the same task as asking them to contemplate a lottery consisting of perfect health and 'immediate death', which is what the SG typically does [30]. The latter involves imagining the possibly unpleasant *process* of dying, whereas the former involves participants reflecting on what the *state* of 'being dead' is like.

> 4 'If the interpretation of VAS valuations as points on a measurable valuation function is rejected, there remains no theoretical justification for the use of VAS methods in CUA'. (Brazier *et al*. [23])

Two pieces of empirical evidence have been brought against the interpretation of VAS as a measurable value function: *context bias* and *end state aversion*. It is alleged that these problems result in VAS scores not having interval-scale properties.

Context bias is the allegation that VAS values are affected by the choice of comparators. Bleichrodt and Johannesson [33] report an empirical test of the theoretical properties of VAS values. Their results show that VAS values for a given state were not independent of other states included in the exercise, with the valuations dependent on the number of health states preferred or less preferred to the state for which valuations are sought. They quote similar findings by Loomes *et al*. [3] in support of their contention that context effects render VAS valuations inconsistent. These findings question whether VAS valuations *do* represent an underlying measurable valuation function; they form the rationale for the conclusions by Brazier *et al*. quoted above. Citing Parducci [34], Brazier *et al*. point to an explanation of context bias in 'response spreading', where 'the respondent seeks to place (spread) responses over the whole (or a specific portion) of an available scale' [23].

However, Schwartz [35] pointed out that Parducci's range-frequency theory provides not only a theoretical reason for the existence of context effects, but also a means of retrieving true preferences. Schwartz used a transformation of raw VAS scales that takes account of the VAS score, the minimum and maximum VAS scores and the rank. Applying this to the Bleichrodt and Johannesson data removed the observed inconsistencies. Subsequently, Robinson *et al*. [36] applied the same transformation to their data, with the same results. The conclusions by Brazier *et al*. concerning the theoretical justification for the use of VAS methods are therefore overturned, at least for the present. It does, however, imply that transformed rather than raw VAS valuations should be used.

End of scale aversion is the allegation that respondents avoid using the ends of the VAS.[5] Torrance *et al*. [1] found evidence of end aversion towards the top of the scale, but were able to correct for this. Again, the conclusion is that VAS valuations do have desirable properties, but should be used transformed rather than raw.

Psychometric issues should of course be a key means of assessing the VAS. However, that is also true for alternative techniques, and not only do these alternatives have known problems, other possible problems have not been investigated to the same extent as for the VAS. It is incorrect, in making judgements between alternatives, to make a partial comparison based on the alleged psychometric defects of one alternative without assessing the known and possible defects of the others. All methods – including VAS, SG and TTO – will suffer from framing effects to some extent. In the absence of a 'gold-standard' technique, it is not possible to prove that *any* of the alternative techniques produce scores that have interval properties. Concluding that scores derived from a VAS do not have interval properties simply because they do not behave in the same way as those from SG and TTO [30] is fallacious unless there is proof that SG and TTO scores in practice have interval properties.

5 '... we have no empirical basis for making assumptions as to what people *mean* by their placements (*on the VAS*) ... it is far from self evident that they are trying to express utility weights ... responses to the question on the meaning of valuations indicate that one should not put too much emphasis on the numerical values as such'. (Nord [16])

VAS is not the only health state valuation method that suffers from an apparent discord between theoretical proposition and observable choices and behaviours. There is considerable evidence that human behaviours and choices under experimental conditions violate the axioms of expected utility theory (EUT) [37–40]. Brazier *et al*. [23] cited the conclusion of Llewellyn-Thomas *et al*. that:

> ...because people's decision behaviours often are not congruent with the axioms of rational choice, the validity of using this prescriptive method (EUT) to describe an individual's actual decision-making, or to select the 'best' treatment strategy for that individual, has to be challenged [41].

Indeed there are fundamental questions about the extent to which the valuations generated by any method can be considered to *elicit* fully formed stable preferences held by individuals as mental entities, or whether such valuations are constructed in response to the particular method by which they are sought [42]. These concerns are pertinent to the validity of the theoretical foundations of *all* methods of health state valuation.

Above all, it should be remembered that with concepts such as utility functions, we are dealing with constructs of social science rather than directly observable entities. The appropriate way to regard them is that people behave *as if* they have utility functions and utility weights, rather than that people actually have them. Qualitative evidence is useful in assessing whether or not people do behave in conformity to social science constructs, but its role is not to provide evidence about whether or not such constructs exist in their own minds. It would be as logical to refute consumer choice theory on the basis that people do not in fact say they refer to their indifference maps and budget lines when they make decisions about the purchase of goods and services.

Conclusions and suggested research on visual analogue scale valuation

We have outlined a defence for the VAS, based largely on a refutation of its alleged theoretical defects and consequent inferiority to alternative methods such as SG and TTO. In some cases, we argue that the VAS may actually be superior on theoretical grounds to these alternatives. However,

there is considerable evidence that the VAS has other advantages, in particular in terms of feasibility and reliability (see [27] for a systematic review). The practical advantages of VAS lie principally in valuations being able to be elicited via postal surveys. This reduces cost, and increases the speed and frequency with which valuations may be sought in response to academic or clinical issues. There have been challenges to this. Respondents may find the VAS difficult because of the complexity of the task they are asked to complete, leading to problems with the data quality [43,44], although such problems can be addressed through improved instrument design [45]. It has been alleged that respondents find the VAS more difficult than other techniques; however, evidence on this is mixed. Moreover, using a less 'difficult' method does not unambiguously confer practical advantages, because participants who indicate no or little difficulty in completing valuation exercises are more likely to return low quality valuation data [25].

It should be said that this review and other assessments of the VAS do not assess the means by which VAS data are collected, for example interview *versus* postal survey, and the differences that this may cause in the characteristics of the data. For example, Greiner [46] emphasises the importance of differences between VAS valuations that have been preceded by a ranking exercise and those that have not. The documented advantages of VAS methods may not apply to all ways in which the VAS is used; such research should be systematised and gaps in it filled.

More generally, the empirical properties of VAS valuations should be fully explored. First, it is possible that some of sources of bias may be possible to 'design out', and effort should continue to be directed toward research to inform improvements to VAS instruments. Secondly, research could investigate the possibility of developing of standard transformation algorithms to remove context bias and end-aversion. Torrance *et al.* [1] concluded that there is a restricted use for the VAS, largely as an aid to producing pseudo-SG utilities. This is consistent with the aims of their research programme, which is to produce values that conform to expected utility theory, and uses techniques such as VAS to help approximate them. However, other research agendas are possible, and research programmes already exist which could facilitate the development of standard transformation algorithms to remove context bias and end-aversion.

For example, the EuroQol Group has a standard, widely used instrument and a set of routines for particular uses. This makes the EQ-5D particularly amenable to research on these issues. The instrument is well defined and if applied strictly according to the EuroQol Group's recommendations should produce comparable and replicable results. The actual values for the transformations to remove context bias and end-state aversion are specific to applications; however, these should not vary if the context and the end states do not vary. For example, the EQ-5D valuation exercises typically value the same 'core' set of states (so the context is identical across studies), and the endpoints of the VAS are well defined and standard in their use and application (so the end points do not vary across studies). A hypothesis is that there exist widely applicable, standard transformations that could be applied to the numerous sets of EQ-5D VAS valuation data that have been generated in Europe and elsewhere.

Our principal conclusion is that although it has become part of the orthodoxy of health economics that VAS values should not be used in CUA, the foundations of this position are actually quite weak. Whilst recent writing suggests near unanimity on the inappropriateness of the VAS, earlier papers were much more equivocal; the absence of new theoretical developments suggests that the current consensus reflects the evolution of beliefs rather than analysis. In fact, both theoretical and empirical evidence suggests that the VAS is a sound method that has many advantages over its rivals. There are strong grounds for challenging the prevailing prejudice against the VAS, and we suggest that those who advocate alternative methods need rather better arguments and evidence than they currently deploy. However, there are many areas in which empirical research is required to establish and consolidate the potential of VAS as a valuation method.

Acknowledgements

We are grateful for useful comments received on earlier drafts of this paper which were presented at the EuroQol Group conference in October 2003 and the CES-HESG conference in January 2004, in particular from the discussants Alan Williams and Pierre Lévy. We also thank two anonymous referees for *Health Economics* for their helpful suggestions.

Notes

1. It would be quite possible to value QALYs directly using a VAS scale, but that has not, to our knowledge, been undertaken.
2. This use of the term 'choiceless' refers to the means by which the value of a state is measured. This is different to 'choiceless utility', which refers to the utility derived from a state of the world that the individual experiences without having chosen it, in contrast to 'modified utility' [47]. The former use of the term 'choiceless' is relevant to *utility measurement*, the latter to *utility theory*.
3. In health economics, many 'willingness-to-pay' studies do appear to use 'choiceless' methods: respondents are merely asked to state money values for entities, which is a similar cognitive task to asking them to state values for them using some other metric. WTP methods are, in practice, mainly assessed on their psychometric properties rather than their adherence to economic theory, and are not subject to the same criticism on this basis as 'choiceless' utility measurement methods such as the VAS.
4. The VAS could be used to support an alternative approach that even more directly tackles the nature of these trade-offs between dimensions/levels. For a given EQ-5D health state with a given VAS placement, what improvement in one dimension would be required to exactly compensate for a worsening in another? This would enable the choices and trade-offs we posit above to be explored to establish points of indifference. The difficulty in implementing this approach is the limitations in the number of levels; an alternative metric (such as money) would be required to operationalise the procedure.
5. Normally, this causes a problem with the scores given to health states which are closest to the best or worst health states described by the VAS end points; such end points are *defined* as having values such as 100 and 0. A complication with the EQ-5D is that the best health state described by the descriptive system (11111) is *valued* relative to the best imaginable health state, which is given the value 100. The problem that arises if 11111 is given a score below 100 is conventionally dealt with by rescaling the scores for 11111 and all other valued states.

References

1. Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Med Decis Making* 2001; **21**(4): 329–334.
2. Brooks R, Rabin R, de Charro F (eds). *The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective*. Kluwer: Dordrecht, 2003.
3. Loomes G, Jones-Lee MW, Robinson A. What do visual analogue scales really measure? *Paper presented to HESG*, Newcastle, July 1994.
4. Dyer JS, Sarin RK. Measurable multiattribute utility functions. *Oper Res* 1979; **27**(4): 810–822.
5. Dyer JS, Sarin RK. Relative risk aversion. *Manage Sci* 1982; **28**(8): 875–886.
6. Cohen BJ. Is expected utility theory normative for medical decision making? *Med Decis Making* 1996; **16**(1): 1–6.
7. Gold M, Stevenson D, Fryback D. HALYs and QALYs and DALYs, oh my: similarities and differences in summary measures of population health. *Annu Rev Public Health* 2002; **23**: 115–134.
8. Culyer AJ, Lavers RJ, Williams A. Social indicators: health. *Soc Trends* 1971; **2**: 31–42.
9. Klarman HE, Francis JO, Rosenthal GD. Cost-effectiveness analysis applied to the treatment of renal disease. *Med Care* 1968; **6**: 48–54.
10. Weinstein MC, Stason WB. Foundations of cost effectiveness analysis for health and medical practice. *N Engl J Med* 1977; **296**(13): 716–721.
11. Boyle MH, Torrance GW, Sinclair JC, Horwood SP. Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *N Engl J Med* 1983; **308**: 1330–1337.
12. Fisher GH. The role of cost-utility analysis in program budgeting. In *Planning Programming Budgeting. A Systems Approach to Management*, Lyden FJ, Miller EG (eds). Markham Publishing Company: Chicago, 1972.
13. Torrance GW. Measurement of health state utilities for economic appraisal: a review. *J Health Econ* 1986; **5**(1): 1–30.
14. Kind P. Guidelines for value sets in economic and non-economic studies using EQ-5D. In *The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective*, Chapter 4, Brooks R, Rabin R, de Charro F (eds). Kluwer: Dordrecht, 2003.
15. Johannesson M, Jonsson B, Karlsson G. Outcome measurement in economic evaluation. *Health Econ* 1996; **5**: 279–296.
16. 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.
17. Klarman HE. Application of cost-benefit analysis to the health services and the special case of technologic innovation. *Int J Health Serv* 1974; **4**(2): 325–352.
18. Culyer AJ. *Need and the National Health Service*. York Studies in Economics. Martin Robertson, 1976.
19. Sen A. Social choice theory: a re-examination. *Econometrica* 1977; **45**: 53–90.

20. Culyer AJ. The normative economics of health care finance and provision. In *Providing Health Care*, McGuire A, Fenn P, Mayhew K (eds). Oxford University Press: Oxford, 1991.
21. Birch S, Donaldson C. Valuing the benefits and costs of health care programmes: where's the 'extra' in extra-welfarism? *Soc Sci Med* 2003; **56**(5): 1121–1133.
22. Tsuchiya A, Williams A. Welfare economics and economic evaluation. In *Economic Evaluation in Health Care: Merging Theory With Practice*, Chapter 2, McGuire A, Drummond MF (eds). Oxford University Press: Oxford, 2001.
23. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. NHS R&D HTA programme. *Health Technol Assess* 1999; **3**(9).
24. Pearce D, Özdemiroglu E. *Economic Valuation with Stated Preference Techniques*. Department for Transport, Local Government and the Regions: London, 2002.
25. Devlin N, Hansen P, Kind P, Williams A. Logical inconsistency in survey respondents' health state valuations – a methodological challenge for estimating social tariffs. *Health Econ* 2003; **12**: 529–544.
26. Dolan P. The measurement of health related quality of life. In *Handbook of Health Economics*, vol. 1b, Chapter 32, Culyer AJ, Newhouse JP (eds). North-Holland: Amsterdam, 2000.
27. Dolan P, Gudex C. Time preference, duration and health state valuations. *Health Econ* 1995; **4**(4): 289–299.
28. Buckingham K, Devlin N, Tabberer M. *A theoretical framework for TTO valuations and a taxonomy of TTO approaches: results from a pilot study*. Economics Discussion Paper 04/07, Department of Economics, City University, London, 2004.
29. Richardson J. Cost utility analysis: what should be measured? *Soc Sci Med* 1994; **39**(1): 7–21.
30. Drummond MF, O'Brien B, Stoddart G, Torrance G. *Methods for the Economic Evaluation of Health Care Programmes* (2nd edn). Oxford Medical Publications: Oxford, UK, 1997.
31. Broome J. QALYs. *J Public Econ* 1991; **50**(2): 150–167.
32. Torrance G, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Oper Res* 1982; **30**(6): 1043–1068.
33. Bleichrodt H, Johanesson M. An experimental test of a theoretical foundation for rating scale valuations. *Med Decis Making* 1997; **17**: 208–216.
34. Parducci A. Contextual effects. A range-frequency analysis. In *Handbook of Perception*, vol. III, Carterette E, Freidman M (eds). Academic Press: New York, 1974.
35. Schwartz A. Rating scales in context. *Med Decis Making* 1998; **18**: 236.
36. Robinson A, Loomes G, Jones-Lee M. Visual analogue scales, standard gamble and relative risk aversion. *Med Decis Making* 2001; **21**(1):17–27.
37. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica* 1979; **47**: 263–291.
38. Llewellyn-Thomas H, Sutherland H, Tibshirani A, Ciampi J, Till J, Boyd N. The measurement of patients' values in medicine. *Med Decis Making* 1982; **2**: 449–462.
39. Schoemaker P. The expected utility model: its variants, purposes evidence and limitations. *J Econ Lit* 1982; **20**: 529–563.
40. Camerer C. Individual decision-making. In: *Handbook of Experimental Economics*, Kagel J, Roth A (eds). Princeton University Press: Princeton, NJ, 1993.
41. Llewellyn-Thomas H, Williams JI, Levy L, Naylor CD. Using a trade-off technique to assess patients' treatment preferences for benign prostatic hyperplasia. *Med Decis Making* 1996; **16**: 262–282.
42. Baker R, Robinson A. Responses to standard gambles: are preferences 'well constructed'? *Health Econ* 2004; **13**(1): 37–48.
43. Bjork S, Norinder A. The weighting exercise for the Swedish version of the EuroQol. *Health Econ* 1999; **8**: 117–126.
44. Devlin N, Hansen P, Selai C. Understanding health state valuations: a qualitative analysis of respondents' comments. *Qual of Life Res* 2004; **13**(7): 1265–1277.
45. Devlin N, Hansen P, Macran S, Herbison P. A 'new and improved' EQ-5D valuation questionnaire? Results from a pilot study. *Eur J Health Econ* 2005; **6**(1): 73–82.
46. Greiner W. A European EQ-5D valuation set. In *The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective*, Chapter 8, Brooks R, Rabin R, de Charro F (eds). Kluwer: Dordrecht, 2003.
47. Loomes G, Sugden R. Regret theory: an alternative theory of rational choice under uncertainty. *Econ J* 1982; **92**(368): 805–824.

