Department of Economics
School of Social Sciences

Is there a case for using Visual Analogue Scale valuations in Cost-Utility Analysis?

David Parkin *
City University London

Nancy Devlin
City University London

**Department of Economics
Discussion Paper Series**

No. 04/03

* Corresponding author, address: City Health Economics Centre, City University London, EC1V 0HB, Phone: 020 7040 0171, Fax: 020 7040 8580, e-mail: d.parkin@city.ac.uk

Abstract

This paper critically reviews theoretical and empirical propositions regarding visual analogue scale (VAS) valuations of health states and their use in Cost Utility Analysis. An oft-repeated conclusion in the economic evaluation literature is the inferiority, on theoretical grounds, of VAS valuations. Common criticisms are that VAS lacks a theoretical foundation; that VAS values are not 'choice based'; that VAS values are not consistent with utility-under-uncertainty requirements; and that context and range effects observed in VAS valuation data mean that they cannot even be considered to represent measurable value functions.

We address each of the above points, critically reviewing the economic and psychometric literature relating to theories of utility and theories of utility measurement, and the welfarist and non-welfarist literature relating to social choices and QALYs.

We conclude that there are strong grounds, both theoretical and empirical, for challenging the apparently emerging consensus that VAS valuations should not be used in economic assessments. The theoretical appeal of alternatives such as the standard gamble is valid only at the level of individuals, rather than social decision-making. Further, the non-welfarist foundations of CUA do not require health state valuations to be grounded in any particular theory of utility, suggesting that the selection of the appropriate valuation method should be based on empirical performance. The VAS has important advantages over rival techniques such as standard gamble and time trade-off. However, we identify a number of areas in which further research is required to establish and consolidate the potential of VAS as a valuation method.

1. 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. A key issue regarding VAS, which is the focus of this paper, is the appropriateness of applying the resulting values for health state scenarios 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.

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. In Section 3 we 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, Feeny and Furlong (2001) have reported some similar analysis 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 (Brooks *et al*, 2003). 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 health state 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 VAS component of the EQ-5D, known as the EQ-VAS, is described in the following section.

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

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

2. 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 visual analogue scales; the interpretation, in terms of value, of measurements derived from visual analogue scales; 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 *visual analogue scale* 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-VAS, for example, is by convention[1] a 20cm thermometer-like vertical line which has endpoints labelled 'best imaginable health state possible' and 'worst imaginable health state possible', denoted as 100 and 0 respectively, and is, again by convention[2], demarcated in units of one and labelled in units of ten.

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**" (Loomes, Jones-Lee and Robinson, 1994), therefore capturing the strength of an individual's preferences over the set of states.

[1] An exception is Parkin *et al* (2004), where the VAS was depicted as a 14cm format.
[2] An exception is Devlin *et al* (2002) where the VAS was demarcated in units of two.

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 (1979, 1982). 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, the valuations should not, according to this taxonomy, be called utilities. 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 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, using the VAS, state 21111 and state 23322 at 0.90 and 0.02 respectively, and agree that 21111 is 45 times better than 23322. 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 are inappropriate in health care, which is characterised by a high degree of uncertainty. The basis of this is neither a positive behavioural finding that people are expected utility maximisers, nor a normative proposition that they ought to be. There is ample evidence that people do not in fact attempt to maximise expected utility. 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; however in health the decisions are often one-off and

have serious consequences. The argument's basis is simply that people may not be risk-neutral and that fact ought to be recognised.

Health-adjusted life years and Quality-adjusted life-years

Gold, Stevenson and Fryback (2002) recently coined the term "Health Adjusted Life Years" to refer to a "family" of health measures, the family members being QALYs and 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 a key European contribution to the development of QALYs, by Culyer, Lavers and Williams (1971)[3].

Culyer, Lavers and Williams 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 product that is produced. 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[4] writings on this topic; Klarman, Francis and Rosenthal (1968) calculated what they termed "quality-adjusted life expectancy" based on quality adjustment weights.

Subsequently, an influential article by Weinstein and Stason (1977) 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[5], economic evaluation in health care consists of Cost-Effectiveness (CEA) and Cost-Utility (CUA). Both measure cost

[3] These authors have published the outline of their ideas in a number of other publications.
[4] This is in fact the earliest that we have found.
[5] For example, cost-minimisation analysis (Drummond *et al*, 1997) and Cost-Value Analysis (Nord, 1993).

compared with output, but differ in how output is measured: "physical" quantities or "natural units" (CEA) and health related quality of life, particularly Quality Adjusted Life Years (QALYs) gained (CUA)[6]. A common, but less well-articulated view is that CEA provides evidence about technical efficiency and CUA about allocative efficiency; however, it is arguable that in practice both aim to provide information relevant to decisions about allocative efficiency.[7] As normally practised in health economics, both produce estimates of the observed cost of achieving different levels of output[8]. The difference 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 Boyle, Torrance, Sinclair, and Horwood (1983). However, this was neither a neologism nor an inevitable choice. The term already existed; for example Fisher (1972) 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 (1986), 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 such an analysis is closer to narrowly defined CUA than to CEA, where measures are in "physical quantities" or "natural units". Kind (2003) has argued that "…valuing changes in EQ-5D health states using VAS methods is entirely consistent with the use of any naturally occurring, convenient outcome parameter as the measure of benefit in a cost-effectiveness study. ... If cost per unit change in blood pressure, then why not cost per unit change in EQ-5D?"[9]. Patients' self-rated VAS scores of their own health states might indeed be argued to better fit the

[6] They should, in theory, also differ in how costs are measured, for example the range of costs included and use of market or shadow prices.
[7] They do not examine pure technical efficiency; instead they assume it, since allocative efficiency cannot exist unless technical efficiency exists.
[8] Except in special cases, for example where observed costs may be reduced with no change or an increase in output, or output may be increased with no change or a reduction in observed costs.
[9] It is assumed that it is intended that such changes are not in VAS scores alone, but in V-QALYs, unless the duration of change is not important.

idea of a 'naturally occurring unit of measurement' than society's valuations of patients' EQ-5D states. However, although this does suggest legitimacy for V-QALY based evaluations, it does not accord them the special place in CEA that they should have.

3. 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, Jönsson and Karlsson, 1996).

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 (1991) 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[10]. However, if QALYs are a more general concept, then VAS values are indeed related to the theory of QALYs.

[10] Of course, it would be quite possible to value QALYs directly using a VAS scale, though that has not, to our knowledge, been undertaken.

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 he explicitly stated that the use of QALYs was intended as a non-monetary numeraire for cost-effectiveness analysis (Klarman, 1974). Similarly, Culyer, Lavers and Williams (1971) state 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')".

Culyer (1976) 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" (Culyer 1976, p. 44).

Subsequently, the use of QALYs in economic evaluation has come to be grounded in the theory of extra welfarism. Beginning with Sen (1977) and first articulated specifically in relation to the use of QALYs by Culyer (1991), 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 cost utility analysis (Birch and Donaldson, 2003), 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* (Tsuchiya and Williams, 2001).

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, Jönsson and Karlsson, 1996).*

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 *(1999).*

A widespread view amongst health economists is that "choice-based", or "indirect" valuation techniques are based on economic theory, but "choiceless", or "direct" techniques are not[11]. However, the justification for this is not clear. As a view based on scientific principles, rather than simple prejudice, it may derive from another terminological confusion, concerning *revealed*

[11] 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' (e.g. Loomes and Sugden 1982), 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'. The fromer use of the term 'choiceless' is relevant to *utility measurement*, the latter to *utility theory.*

preference (RP) and *stated preference* (SP). RP refers to valuations of goods and services that can be inferred from real choices that are made in the everyday world. It has a clear basis in economic theory: real choices are based on comparing benefits with opportunity costs, or what has to be sacrificed to obtain them; observing real choices therefore reveals real preferences. Such valuations may either be direct, where the good or service has a market price, or indirect, where hedonic prices, such as time prices or risk premiums, can be observed. SP refers to valuations derived from experimental or survey data. Its main theoretical base is psychology and the relevant measurement theory is psychometrics. An extreme view, associated with Austrian economics, is that only choices in the market have relevance to economic analysis; valuations derived from psychology may be of interest, but refer to an entirely different phenomenon than that which is relevant to economics. However, most economists (for a UK example, see Pearce and Özdemiroglu, 2002) seem to agree that it is acceptable to use SP techniques, as long as they are consistent with the aims of economic analysis.

It is clear from this definition that *all* utility measurement techniques, as well as techniques such as willingness to pay, contingent valuation and conjoint analysis are SP techniques and are largely based on psychology and psychometrics[12]. However, some economists have incorrectly conflated these concepts by suggesting that RP refers to "choice-based" techniques, which are backed by economic theory, and SP to "choiceless" techniques, which are not[13].

There is no justification for this. Preferences are not *revealed* in TTO or SG; they are simply stated. Utilities are then inferred from those stated preferences. A more correct distinction would be between direct and indirect measurement of utility. However, 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, based on economic reasoning, for supporting one technique rather another, but the presence or absence of choice is not of necessity one of them[14]. The SG technique, for example, was not

[12] The mainstream view that endorses SP strictly only concerns monetary valuation rather than utility measurement. This narrow view would exclude *all* utility measurement techniques from respectability within economic evaluation.

[13] For example, Abdellaoui et al (2002) state that choice-based utilities, made under uncertainty are 'derived from revealed preferences' whereas choiceless utilities are 'derived from direct judgements'.

[14] A curious feature of many "choice-based" techniques is that they are based on finding points of indifference, or equivalence between alternatives. This is not at all consistent with what is observed in the

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. Many willingness-to-pay studies appear to use equally "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, but are not subject to the same criticism on this basis as the VAS.

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.

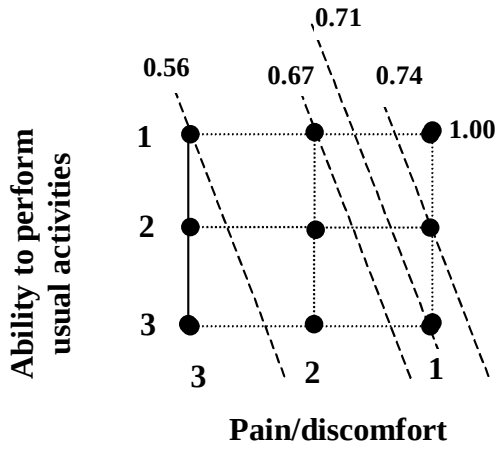
Less trivially, VAS valuations involve respondents weighing up various health state scenarios, each of which embodies specific variants of health related dimensions and levels. For example, the 'standard' set of EQ-5D health states typically valued in VAS valuation questionnaires includes the states 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, Lavers and Williams (1971) and assuming that all other EQ-5D dimensions are held constant at level 1. Contour lines

---

everyday world, where although choices are common and observable phenomena, points of indifference are rarely if ever observed.

connect the points which are considered 'equally bad'; in this example, the scores[15] for each contour line show that the valuations placed on each state are more strongly effected by decrements in pain than in usual activities (thus 11211 > 11121, and 11311 > 11131).

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



If one accepts that the dimensions in the health state descriptive system can be treated as *separate* arguments in the individual's utility function, then this potentially overcomes the advantage claimed of SG and TTO i.e., that 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 (Dolan, 2000).

The act of valuing EQ-5D states using the VAS could therefore be argued to involve both choice and tradeoff. While nothing is 'given up' to obtain the valuation[16], the purpose of the exercise (the preference attached to one health state *vis a vis* others, and thus the value attached to improved health) is probably more transparent in a VAS approach than in TTO or SG where (other than for

[15] The scores shown in Figure 2 are New Zealand VAS values estimated from the 'full population tariff' (Devlin, Hansen, Kind and Williams 2000).

[16] It is possible to conceive of the VAS being used to support an alternative approach that even more directly tackles the nature of these tradeoffs 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 tradeoffs 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.

chronic states worse than dead) the valuation of each state is sought relative only to time in perfect health or the probability of perfect health and death.

3 VAS values "…are only appropriate for problems that involve certainty; thus, values are much more restricted in their applicability" (Drummond et al*, 1997)*

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[17].

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).

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[18].

[17] It is well known that TTO was originally devised as an approximation to the values that would be produced by SG, because SG valuation methods were deemed impracticable.

[18] This does not suggest that the Person Trade-Off (PTO) is an appropriate method for measuring health state values for individuals. That method is largely an opinion poll about what decision-makers' values should be.

A similar argument can be traced to Broome (1991). Broome's 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".

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 or the probability method can determine them". Broome describes the VAS approach employed by Torrance *et al* (1982), 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 may do, is not the same task as asking them to contemplate "dying", which is what the standard gamble may do.

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*, 1999).*

Two pieces of empirical evidence have been brought against the interpretation of VAS as a measurable value function: *context bias* and *end state aversion*.

Context bias is the allegation that VAS values are affected by the choice of comparators. Bleichrodt and Johannesson (1997) 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, Jones-Lee and Robinson (1994) 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 (1974), 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".

However, Schwartz (1998) 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, Loomes and Jones-Lee (2001) 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-state aversion is the allegation that respondents avoid using the ends of the VAS; presumably they are reluctant to admit to described states being best = 1 and worst = 0. Torrance, Feeny and Furlong (2001) found evidence of end-state 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.

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, 1991).*

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) (Kahneman and Tversky, 1979, Llewellyn-Thomas *et al*, 1982, Schoemaker, 1982, Camerer, 1993). Llewellyn-Thomas *et al* conclude 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" (cited by Brazier *et al*, 1999)

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 (Read *et al*, 1984). 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 indifference curve 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.

4. Conclusions and suggested research on visual analogue scale valuation

Although recent writing suggests near unanimity on the inappropriateness of VAS valuation in economic evaluation, earlier papers are much more equivocal. The absence of new theoretical developments suggests that current consensus reflects the evolution of beliefs rather than analysis. In fact, both empirical and theoretical evidence suggests that the VAS is a sound method that has many advantages over its rivals. However, there are many areas in which empirical research is required to establish and consolidate this potential.

We have outlined a defence for use of the VAS, in which some positive aspects emerged, for example the inappropriateness of using expected utility based measures in a social decision-making context and the consequent superiority of the VAS in that regard. However, the VAS has other advantages. In particular, there is considerable evidence that the VAS has advantages over other methods in terms of feasibility and reliability (see Brazier *et al*, 1999, for a systematic review). 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 (2003) 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. Torrance, Feeny and Furlong (2001) 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. A hypothesis is that there exist widely applicable, standard transformations which could be applied to the numerous sets of EQ-5D VAS valuation data that have been generated in Europe and elsewhere.

Finally, if utilities similar to those estimated using the SG method are, despite the arguments we advance in this paper, deemed to be required, then the evidence provided by Torrance, Feeny and Furlong (2001) also suggests that VAS data can be transformed to achieve this. Although there is some controversy about the applicability of a transformation from VAS to SG scores, the evidence suggests that if aggregate data are used and the VAS data are transformed to remove context bias and end-aversion, a power transformation between the two can be found. This power transformation will be context specific, but the context in this case means a particular instrument applied to particular health states among a particular population. Again, the EuroQol Group's use of a standard VAS approach for the valuation of EQ-5D states permits investigation of the hypothesis that there exists a widely applicable and standard power transformation that can be applied to these VAS valuation data.

Acknowledgements

The authors are grateful for helpful comments received on earlier drafts of this paper presented at the EuroQol Group conference, October 2003 and the CES-HESG conference, January 2004, in particular the discussants Alan Williams and Pierre Lévy.

References

Abdellaoui, M., Barrios, C. and Wakker, P. (2002) Reconciling introspective utility with revealed preference:experimental arguments based on prospect theory. CREED, Department of Economics, University of Amsterdam, The Netherlands. http://www1.fee.uva.nl/creed/wakker/pdf/mocawa.pdf

Birch, S. and Donaldson, C. (2003) Valuing the benefits and costs of health care programmes: where's the 'extra' in extra-welfarism? *Social Science and Medicine* 56(5): 1121-33.

Bleichrodt, H., Johanesson, M. (1997) An experimental test of a theoretical foundation for rating scale valuations. *Medical Decision Making* 17: 208-216.

Boyle, M.H., Torrance, G.W., Sinclair, J.C. and Horwood, S.P. (1983) Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *The New England Journal of Medicine*, 308: 1330-1337.

Brazier, J., Deverill, M., Green, C., Harper, R. and Booth, A. (1999) A review of the use of health status measures in economic evaluation. NHS R&D HTA programme, *Health Technology Assessment* 3(9).

Brooks, R., Rabin, R. and de Charro, F (eds) (2003) *The measurement and valuation of health status using EQ-5D: a European perspective*. Kluwer.

Broome, J. (1991) QALYs. *Journal of Public Economics* 50(2) 150-167.

Camerer, C. (1993) Individual decision-making in: Kagel, J., Roth, A. (eds) *Handbook of Experimental Economics*. Princeton University Press.

Culyer, A.J. Lavers, R.J., Williams, A. (1971) Social indicators: health. *Social Trends* No. 2.

Culyer, A.J. (1976) *Need and the National Health Service*. York Studies in Economics. Martin Robertson.

Culyer, A.J. (1991) *The normative economics of health care finance and provision.* In: McGuire, A., Fenn, P., Mayhew, K (eds) Providing health care. Oxford University Press.

Devlin, N., Hansen, P., Kind, P. and Williams, A. (2000) *The health state preferences and logical inconsistencies of New Zealanders: a tale of two tariffs.* CHE Discussion paper 180, University of York.

Devlin, N., Hansen, P. and Macran, S. (2002) A 'new and improved' EQ-5D valuation questionnaire? Results from a pilot study. In: Kind, P. and Macran, S. (eds) *Proceedings of the 19th Plenary Meeting of the EuroQol Group,* Centre for Health Economics, University of York.

Dolan, P. (2000) The measurement of health related quality of life. Chapter 32 in: Culyer, A.J., Newhouse, J.P (eds) *Handbook of health economics* Volume 1b. North Holland.

Drummond, M.F., O'Brien, B., Stoddart, G. and Torrance, G. (1997) *Methods for the economic evaluation of health care programmes.* Oxford Medical Publications, 2nd edition.

Dyer, J.S. and Sarin, R.K. (1979) Measurable multiattribute utility functions. *Operations Research* 27(4):810-822.

Dyer, J.S. and Sarin, R.K. (1982) Relative risk aversion. *Management Science* 28(8):875-886.

Gold, M., Stevenson, D. and Fryback, D. (2002) HALYs and QALYs and DALYs, Oh My: similarities and differences in summary measures of population health. *Annu. Rev. Public Health*, 23: 115-34.

Greiner, W. (2003) A European EQ-5D valuation set. Chapter 8 in: Brooks, R., Rabin, R., de Charro, F (eds) *The measurement and valuation of health status using EQ-5D: a European perspective*. Kluwer.

Fisher, G.H., 1972. The Role of Cost-Utility Analysis in Program Budgeting. In: Lyden, F.J. and Miller, E.G. (editors) *Planning Programming Budgeting. A Systems Approach to Management*. Markham Publishing Company, Chicago.

Johannesson, M, Jonsson, B and Karlsson, G. (1996) Outcome measurement in economic evaluation. *Health Economics* 5: 279-96.

Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47:263-91.

Kind, P. (2003) Guidelines for value sets in economic and non-economic studies using EQ-5D. Chapter 4 in: Chapter 8 in: Brooks, R., Rabin, R. and de Charro, F (eds) *The measurement and valuation of health status using EQ-5D: a European perspective*. Kluwer.

Klarman, H.E., Francis, J.O. and Rosenthal, G.D. (1968) Cost-effectiveness analysis applied to the treatment of renal disease. *Medical Care,* 6: 48-54.

Klarman, H.E. (1974) Application of cost-benefit analysis to the health services and the special case of technologic innovation. *International Journal of Health Services*, 4(2): 325-352.

Loomes, G., Jones-Lee, M.W. and Robinson, A. (1994) *What do visual analogue scales really measure?* Paper presented to HESG, Newcastle, July.

Loomes, G. and Sugden, R. (1982) Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal* 92 (368):805-824.

Llewellyn-Thomas, H., Sutherland, H., Tibshirani, A., Ciampi, J., Till, J. and Boyd, N. (1982) The measurement of patients' values in medicine, *Medical Decision Making* 2: 449-462.

Nord, E. (1991) The validity of a visual analogue scale in determining social utility weights for health states. *International Journal of Health Planning and Management* 6: 234-242.

Nord, E. (1993) The trade-off between severity of illness and treatment effect in cost-value analysis of health care. *Health Policy*, 24, 227-238.

Parducci, A. (1974) Contextual effects. A range-frequency analysis. In: Carterette, E., Freidman, M. (eds) *Handbook of perception* volume III. New York: Academic Press.

Parkin, D, Rice, N., Jacoby A. and Doughty, J. (2004) Use of a visual analogue scale in a daily patient diary: modelling cross-sectional time-series data on health-related quality of life. *Social Science and Medicine* (forthcoming, currently available on the journal website at http://authors.elsevier.com/sd/article/S0277953603005525)

Pearce, D and Özdemiroglu, E. *Economic Valuation with Stated Preference Techniques*. Department for Transport, Local Government and the Regions: London, 2002.

Read, J.L., Quinn, R., Berwick, D., Fineberg, H. and Weinstein, M.C. (1984) Preferences for health outcomes: comparison of assessment methods. *Medical Decision Making* 4(3):315-242.

Robinson, A., Loomes, G. and Jones-Lee, M. (2001) Visual Analogue Scales, Standard Gamble and Relative risk aversion. *Medical Decision Making* 21(1):17-27.

Schoemaker, P. (1982) The expected utility model: its variants, purposes evidence and limitations. *Journal of Economic Literature* 20: 529-563.

Schwartz, A. (1998) Rating scales in context. *Medical Decision Making* 18: 236.

Sen, A. (1977) Social choice theory: a re-examination. *Econometrica* 45: 53-90.

Torrance, G.W., Boyle, M.H., Horwood, S.P (1982) Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research* 30(6):1043-68.

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

Torrance, G.W., Feeny, D. and Furlong, W. (2001) Visual Analog Scales: do they have a role in the measurement of preferences for health states? *Medical Decision Making* 21(4): 329-334.

Tsuchiya, A. and Williams, A. (2001) *Welfare economics and economic evaluation*. Ch. 2 in: Mc Guire, A., Drummond, M.F (2001) *Economic evaluation in health care: merging theory with practice*. Oxford University Press.

Weinstein, M.C. and Stason, W.B. (1977) Foundations in cost effectiveness analysis for health and medical practice. *New England Journal of Medicine.* 296(13):716-21.