

(How) Can We Value Health, Safety and the Environment?

Graham Loomes

Centre for the Economic and Behavioural Analysis of Risk and Decision,
University of East Anglia, Norwich NR4 7TJ, UK.

Abstract

The assumptions that underpin the conventional economic model of ‘rational agents’ tend to be substantially violated by data from surveys that try to elicit people’s values for health, safety and environmental goods. Psychological research suggests that there may be a large affective component in people’s responses to such surveys, with the result that those data are not amenable to the ‘logic’ of economic rationality. This raises questions both about the way we model human judgment and decision processes, and also about the use of survey data to guide public policy in these and other areas.

Introduction

Most people would agree that measures which reduce the risks to human health and safety and protect the environment are desirable activities to which at least some scarce resources should be allocated. However, they are not the only things people want: food, clothing, shelter, transportation, education and entertainment are among the other goods and services we desire. So the question is: how do individuals and societies strike the balance and allocate the appropriate levels of resources to health, safety and the environment as compared with all the other goods and services which are valued?

The structure of this paper is as follows. The next section outlines the conventional economist's answer to that question. There is then an indication of the main kinds of practical problems encountered when researchers attempt to act on the basis of that conventional model. This is followed by a discussion of some of the related and relevant research by psychologists and behavioural economists, with particular reference to the influence of the work of Daniel Kahneman and his peers and collaborators. The concluding section considers some possible issues for policy and future research.

1. The Conventional Economic Approach

It is a fundamental premise of conventional welfare economics that public policy decisions should, as far as possible, reflect the preferences of those who will be affected by them. For example, if it is proposed to introduce some health or safety innovation – some new technology, perhaps, or some change in legislation or regulation – then public policy makers may need to consider how the costs that will fall on members of the population compare with the benefits they may expect to receive. In effect, that requires some (monetary) value to be attached to each component of the costs and benefits, including any changes in expected quality and/or length of life that may be entailed by the innovation.

In this paradigm, the welfare of each individual is the basic unit of analysis. Each individual is characterised as having some level of current and expected future income and wealth and some set of personal values and preferences. On the assumption that he is well informed about the prices and qualities of the large array of goods and services available, it is supposed that he will choose a pattern of present and planned future consumption which will bring him the greatest overall level of satisfaction (or, in

economists' terminology, *utility*) that his wealth will allow. This is, of course, a highly stylised portrait of an individual; however, the model appeals to the assumption that *on average* it is *as if* the population at large operates in this way.

Now suppose we are considering some new safety device or policy which offers some overall reduction in the numbers of injuries and/or deaths, and which, at the level of the individual, translates into some reduction in the personal probability of loss of quality and/or length of life. If this benefit has positive value for an individual, it is assumed that she will be willing to adjust her other (present and/or future) consumption in order to release some resources from current income and/or savings to pay for the safety innovation. It is supposed that she will be willing to make such adjustments up to, but not beyond, the point at which the anticipated benefit is exactly offset by the loss of utility from the other consumption she would have to forego. This is her *willingness to pay* (WTP) for the benefit. If we can elicit such a figure from each member of a representative cross-section of the population, we can derive a measure *in monetary form* of the aggregate value of the health/safety benefit to the population which can then be added to any other (non-health/safety) benefits, so that the total value of all benefits from some innovation can be compared with the total costs the population will have to bear as a result of its introduction¹.

For example, consider some proposed road safety innovation which, if implemented, would be expected to reduce the number of premature deaths on the roads each year by 1 for every 100,000 members of a particular population. Suppose that a random sample is drawn from that population and each member of the sample is asked to say how much such a reduction in the annual risk of premature death is worth to them, expressed in terms of the maximum amount of money they would be willing to pay for a

¹ Some policies may have the potential to produce adverse effects on the health/safety of at least some members of a population – for example, building a new by-pass may improve the safety of many residents of a town, but expose some others who live near, or use, the new road to higher risks. In theory, such adverse effects can be valued by measuring the minimum sum of monetary compensation which, if spent on other consumption, would just make up for the loss of utility entailed by the increased risk to health and safety. This is the *willingness to accept* (WTA) amount. For small increases in risks, standard theory normally expects the WTA figure to be a little higher than the WTP figure for a decrease of the same magnitude, with the two figures tending to converge as the magnitude diminishes. In practice, however, a much bigger disparity between WTP and WTA has often been observed than standard theory can readily accommodate. This raises issues for theory and practice, some of which are considered below, while others are discussed in more detail elsewhere: see for example Sugden (1999) and Guria *et al.* (2004).

reduction of that magnitude. Suppose that the average answer is £1 per month. If the sample is representative, we can infer that every 100,000 members of the population would, between them, be prepared to pay each year a total of £12 x 100,000 for a safety measure which would, on average, prevent 1 premature death on the roads. On this basis, the appropriate 'value of preventing a fatality' (VPF) for road safety project appraisal would simply be set at £12 x 100,000: that is, £1.2m. Moreover, such an approach is not restricted to valuing the prevention of fatalities: in principle, exactly the same method can be used to obtain values for preventing all kinds of different injuries and illnesses. The issue then is to find some robust and reliable way of eliciting the appropriate figures.

However, this has turned out to be a great deal more difficult to implement in practice than standard economic theory would suggest. The essence of the problem appears to be that although the model individual is assumed to have a complete set of values and preferences which she can access and process quite readily, the typical member of the population is not like that. Rather, most people have only somewhat imprecise and partly-formed values for such goods, so that when confronted with questions of the kind indicated above, they cannot simply pull the answer 'off the shelf'.

Moreover, under the sorts of conditions that typically apply in surveys – where people are asked to give answers in a rather limited period of time and with limited opportunity for careful reflection – respondents may be liable to reach for whatever simplifying strategies come most readily to hand, picking up on available (even if unintended) cues, paying more attention to some features of the question than to others, and using simple rules of thumb to come up with an answer. This may result in patterns of responses which do not conform with what would be expected under standard assumptions. The next section gives examples of such patterns of response and discusses the problems they present for policy and for conventional models of the rational economic individual.

2. Problems in Practice

From the many contingent valuation (CV) studies that have now been undertaken², it has become apparent that patterns of response are liable to deviate from the theoretical model in various predictable ways. These may be broadly classified under two headings: where responses show excessive sensitivity to factors that conventional theory considers irrelevant; and where responses show insufficient sensitivity to factors that conventional theory regards as crucial. Let us consider examples of each category in turn.

Excessive Sensitivity to Theoretically Irrelevant Factors

Implicit in most economic models of rational behaviour is a principle of *procedural invariance*: that is, the value or preference an individual expresses should not be affected by the way a decision is ‘framed’ or by the particular method of eliciting the value/preference. If respondents have a reasonably clear idea of their values and preferences and are trying to answer truthfully and accurately, we should expect that different methods of eliciting those preferences should all yield much the same central tendencies. In practice, however, the way decisions are framed and/or features of the elicitation method may exert strong systematic effects on responses which raise doubts about their validity as measures of people’s ‘true’ preferences.

Of course, expecting people to be able to give very firm and exact values, even for more familiar goods, is unrealistic: virtually all human judgments are susceptible to *some* degree of imprecision. To allow for this, a number of survey instruments ask for some indication of what might be thought of as individuals’ confidence intervals around their values. That is, rather than asking respondents to go straight to a single exact WTP (or WTA) amount, they may first be asked to identify some amounts they are quite sure they *would* be prepared to pay and other amounts they are quite sure they *would not* be prepared to pay. The aim is to home in on the range of amounts where they think their value is most likely to lie. If we take the lower bound of this range to be the largest amount they are sure they *would* pay (call this the *min* of their value interval) and take the upper bound (or *max*) to be the smallest amount they are sure they *would not* pay, we might then try to obtain some estimate of the point inside the interval which the respondent thinks is closest to what they think the benefit is worth to them (their *best*

² The term ‘contingent valuation’ really only covers a subset of ‘stated preference’ approaches to obtaining monetary valuations, but will be used in this paper more loosely as shorthand for such studies in general.

estimate). While most economists would not be greatly surprised to find that particular features of framing or elicitation have some effect on the position of the *best* estimate within the band, the standard economic model of preferences would expect that the *min* and *max* would be reasonably stable and robust. However, even this more modest expectation has proved to be overly optimistic.

For example, during developmental work for a major study for the UK Department of Transport to elicit values for preventing non-fatal road injuries, tests were conducted which revealed uncomfortably large *starting point* and *range* effects.

In that study, respondents were asked for their willingness to pay for various specified reductions in the risks of four non-fatal injuries of different severities as well as their risk of death³. To test for starting point effects, respondents were allocated at random to one of two subsamples. For one subsample, the first amount they were asked about was always £25. That is, for each risk reduction in turn they were initially asked whether they were sure they would be willing to pay £25, or whether they were sure they would not pay that amount, or whether they were unsure; thereafter, the values presented to them depended on their responses to this and subsequent amounts, with an iterative procedure designed to identify their *min*, *max* and *best* estimates. For members of the other subsample, the starting point was set at £75; but apart from that difference, the iterative algorithm and all other features of the procedure were the same.

This manipulation was used with five levels of injury and two different questionnaire scenarios⁴. This allowed ten separate pairwise comparisons to be made. The null hypothesis of no significant starting point effect was rejected at the 1% level in six cases, at the 5% level in three more cases and at the 10% level in the tenth case – in all cases in favour of the alternative hypothesis that the £75 starting point resulted in higher responses than the £25 starting point. The mean *best* responses elicited with the £75 starting point ranged from 1.89 to 2.87 times as large as those elicited with the £25 starting point. However, this was not just a matter of the position of the best estimate being influenced within otherwise reasonably stable min-max intervals. Rather, the whole

³ Further details of the injury descriptions and the procedures used can be found in Dubourg *et al.* (1997).

⁴ In one questionnaire, respondents were asked about a lump sum payment for a safety feature which would last for the entire life of the respondent's vehicle; in the other questionnaire, the safety feature had a one year life-span, so the aim was to elicit the amount the respondent would be willing to pay to achieve a reduction in risk for the following 12 months.

interval was liable to be influenced by the starting point: in *every one* of the ten comparisons the average *min* for those initially presented with £75 was between 35% and 114% higher than the average *max* associated with the £25 starting point. There was not a single case where those intervals overlapped at all.

A subsequent round of piloting replaced the iterative procedure with a ‘payment card’ on which a list of possible amounts was displayed, with respondents asked to tick each amount they were sure they would pay, put a cross against each amount they were sure they would not pay, and put an asterisk against their best estimate. In this case, the issue was whether the range of values presented on the payment cards – from £0 to £500 for one subsample, and from £0 to £1500 for the other subsample – would affect responses in the way that the starting point had done. Although the effect here was somewhat less dramatic, it still represented a striking departure from what might be expected from respondents with reasonably robust preferences. In the ten possible comparisons between subsample means, the £0 - 1500 payment cards produced higher mean *best* estimates in nine cases, while the mean *min* values for the £0 - £1500 subsamples exceeded mean *max* values from the £0 - £500 subsamples in five of the ten possible comparisons.

Later attempts to attenuate such effects have met with only limited success. For example, in a study undertaken in 1997-8 for the New Zealand Land Transport Safety Authority, respondents’ values for reducing the risks of various injuries and death were elicited by a computerised iterative procedure where the computer presented an initial value which was then adjusted up or down according to an inbuilt algorithm until the *min*, *max* and *best* estimates were established. The hope was that automating the procedure might cause respondents to attach less significance to the initial value ‘chosen’ by the machine. But although there was some evidence that the starting point effect became weaker in later questions, after respondents had become more accustomed to the procedure, a number of significant influences still persisted (see Guria *et al.*, 1999).

More recently, in a study for the UK Department of the Environment, Food and Rural Affairs (DEFRA) to estimate the value of the health benefits of reducing air pollution, an attempt was made to disarm the effects described above by using a *random card sorting* procedure. When the interview reached the point of eliciting a money response (in this case,

how much a household would pay each year for a specified package of health benefits), the interviewer took a small pack of cards, each one of which had a different sum of money printed on it, and visibly shuffled the pack, explaining that this was to ensure that the cards appeared in no particular order. The respondent was then asked to turn over each card separately and place each in turn into one of three piles: certainly *would* pay; certainly *would not* pay; and *unsure*. The thought was that if the respondent did not see in advance the range of values contained in the pack, she could not be influenced by that range; and that, having seen the pack shuffled, she would have no reason to attach any particular significance to the amount on the first card she turned over. But once again an effect showed through, with a significant positive linear correlation between the amount on that first card and the respondent's stated willingness to pay (see Chilton *et al.*, 2004).

The cases cited above are examples of how theoretically irrelevant factors *within* particular elicitation procedures are liable to have a disconcertingly large impact upon whole distributions of responses. The problems are further exacerbated by differences *between* procedures. In the paragraphs above, there was mention of three somewhat different procedures: iterative bidding, payment cards and random card sorting. Other widely used procedures include open-ended (OE) questions and several variants of dichotomous choice (DC)⁵. The evidence shows that using different procedures can have pronounced and systematic effects on the responses obtained: for example, DC questions invariably produce (substantially) higher mean values than OE questions asked about the same goods.

What is more, such systematic variability is not confined to comparisons between different methods of eliciting *monetary* expressions of value. Following the results of the piloting in the study of non-fatal road injuries referred to above, concern was expressed about whether it was just the *absolute* values for preventing the various injuries that were vulnerable to procedural influences: might it not also be the case that the *relative* values attached to the different injuries were affected by the procedures used?

⁵ A detailed description of all these procedures and examples of their implementation can be found in Chapter 4 of Bateman *et al.* (2002). However, put briefly, OE questions simply ask respondents what they are willing to pay for some good, usually without any prompts such as a starting point or payment card. By contrast, DC questions present respondents with an amount and ask a yes-no question – would they pay this or not? By putting different amounts to different subsamples, DC questions aim to build up a picture of the aggregate demand for the good in terms of the numbers of respondents saying 'yes' at different 'prices' – on which basis an estimate of the mean value can be estimated.

To check on this, the main study randomised respondents between two elicitation methods. One was a money payment card of the kind described above. The other was a ‘standard gamble’ (SG) procedure where respondents were asked to consider the certainty of a particular injury state as compared with an alternative treatment which offered a chance of being restored to normal health but also involved some risk of ending up in an even worse state of health. The SG responses were elicited using a card analogous to the payment card, and in theory both the money elicitation procedure and the SG method should have generated approximately the same relativities between the various non-fatal injuries⁶. But the patterns were radically different: relative to death, the values attached to each of the two more serious injuries were about four times higher when elicited through WTP than when elicited via SG, while the least serious injury was valued more than ten times higher via WTP than via SG. Moreover, in every case the average *max* value elicited by the SG procedure was clearly lower than the corresponding *min* figure elicited via CV.

As mentioned in footnote 1 earlier, substantial and systematic disparities can also be produced within particular elicitation procedures by changing the framing from willingness to pay to willingness to accept. In the twenty years since Knetsch and Sinden (1984) drew attention to it, a mass of evidence of this phenomenon has accumulated in the field of health, safety and the environment. To check for such effects in the road safety pilot studies cited earlier, two WTA questions were also included. These involved risk increases for two of the injuries of the same magnitude as the risk reductions for which interviewees had just given WTP responses. As in the WTP exercises, the WTA procedure elicited both *min* and *max* figures well as *best* estimates.

For a substantial proportion of respondents there was *no overlap whatsoever* between the min-max intervals produced by the WTA questions and those generated by the WTP questions: more than half of the 152 individuals answering all the relevant questions gave *min* WTA responses which strictly exceeded their *max* WTP for the same marginal change in risk⁷.

⁶ For details of the main study, see Jones-Lee *et al.* (1995a); for a further discussion of why the two methods should have produced similar results, but did not, see Jones-Lee *et al.* (1995b).

⁷ For further details, see Dubourg *et al.* (1994).

Altogether, then, there is now a considerable body of evidence suggesting that for unfamiliar and complex goods such as those in the field of health and safety, people's stated preferences are susceptible to a number of influences that are theoretically irrelevant but that are very difficult to eliminate – or even attenuate – in practice. At the same time, there is also an uncomfortably large body of evidence suggesting that respondents may not always give sufficient weight to other considerations which are regarded by theorists and practitioners as being of crucial importance – in particular, as discussed in the next subsection, the magnitude of the benefits being evaluated.

Insufficient Sensitivity to Theoretically Relevant Factors

Some of the most pervasive and problematic patterns of response have been labelled *scope* or *embedding* effects. These occur when respondents state that they are willing to pay just the same, or only a little more, for a benefit that is much larger and that might therefore be expected to receive a substantially higher valuation. Such effects have been reported in many studies valuing environmental goods (for examples, see Kahneman and Knetsch, 1992). And it has also proved extremely difficult to conduct surveys in the field of health and safety where respondents' values show anything like the degree of sensitivity to scope that theory and practical policy would require.

The problems that arise are particularly sharp in relation to the elicitation of values for reducing physical risks. Recall that in the example given earlier, each member of a representative sample was asked how much he or she would be willing to pay for a safety measure which would reduce their risk of premature death on the roads by 1 in 100,000. By taking the average response (£12 in that example) and multiplying it by 100,000, we arrive at a VPF of £1.2m.

But what if those respondents had been asked what they would be willing to pay for a reduction of a different size – say, 3 in 100,000? Under conventional assumptions, so long as the sum of money is a relatively small fraction of respondents' incomes, they would be expected to pay almost three times as much for this three-times-bigger benefit. So, for example, if the average WTP for this larger risk reduction turned out to be £30, that would mean that 100,000 people would, collectively, pay £3m to prevent 3 deaths, giving a VPF of £1m. And although the latter figure is a bit smaller than the estimate derived from the 1 in 100,000 question, both are in roughly the same ballpark, and

policymakers might feel reasonably confident that a VPF of that order could be used in road safety cost-benefit analysis; and that the same figure could be used as well for smaller innovations expected to prevent one or two deaths over some period as for larger innovations which might in due course prevent 50 or 100 deaths.

However, suppose that people's responses display the kind of insensitivity to scope that has been so frequently observed. That is, something like 30%-40% of respondents give *exactly the same value* for both sizes of risk reduction, even if they are asked about them one immediately after the other, while a similar proportion give a value for the bigger benefit which is more than they stated for the smaller benefit, but falls far short of proportionality (see Jones-Lee *et al.*, 1995, and Beattie *et al.*, 1998, for examples). Typically, then, the average WTP for the 3 in 100,000 reduction would be no more than 50% higher than the average WTP for the 1 in 100,000 reduction. In terms of the example, people give an average WTP of no more than £18 for the 3 in 100,000 risk reduction, from which we would derive a VPF of £0.6m. In other words, respondents' insensitivity to the size of the benefit means that the estimate of a VPF may be halved (or doubled, or possibly affected to an even greater extent) depending on the size of benefit *the researchers choose to ask about*. If such very different estimates of a VPF can be generated by the same people answering adjacent questions which differ only in the magnitude of the risk reduction, which number – if any – should be chosen to be used for policy?

It might be thought that the difficulties arise from asking respondents about probabilities – especially small changes in already small probabilities, which are difficult for even quite numerate people to imagine and manipulate. However, there is evidence that the problem is more deeply-rooted than that. For example, in one study respondents were asked to consider different road safety programmes where the benefits were expressed not in terms of risk reductions but in terms of the numbers of deaths prevented each year (alternatively, over a 5-year period) in their region. They were told that one programme would reduce the number of deaths by 5 per year, while a more extensive programme would prevent 15 deaths per year (alternatively, 25 or 75 deaths would be prevented during the next five years).

Initially, respondents were asked for a provisional estimate of what each programme would be worth to their household, expressed in terms of their WTP for each year of the programme's life. They were then presented with the following prompt:

“In the past, we've found that some people say that preventing 15 (75) deaths on the roads is worth three times as much to them as preventing 5 (25) deaths on the roads; but other people don't give this answer. Can you say a bit about why you gave the answers you gave?”

Despite this rather pointed prompt and despite the fact that the question involved numbers of deaths prevented rather than small probability changes, 22 of the 56 respondents placed the same non-zero value on both programmes, while only 11 gave a value for the larger programme that was more than double the value they placed on the smaller programme. Overall, then, the mean (median) value placed on the 15-per-year reduction was just 33% (41%) higher than the corresponding value placed on the 5-per-year reduction. Moreover, the one-year form of question produced significantly different distributions of values from those generated by the five-year form, with the result that the implied VPFs varied by as much as a factor of 4 (between £2.56m and £11.07m) depending on the size of programme and the form of question (for further details, see Beattie *et al.*, 1998).

In the recent DEFRA study, respondents were asked to value increases in life expectancy in normal health for themselves and all members of their immediate household. Participants in the survey were allocated at random to one of three subsamples, with the length of extra life expectancy being varied between the subsamples: 1 month, 3 months and 6 months. In all cases, this benefit was presented as a sure thing enjoyed by all household members (whose names were recorded at the start of the interview and specifically repeated at the time of the value elicitation). Despite this, it was apparent from regression analysis that even after controlling for *per capita* income and other variables, responses were insufficiently sensitive to the two key factors, namely the number of members of the household who would benefit and the number of extra months each household member stood to gain. The mean WTP for an extra 6 months was only just over 30% higher than the 1-month figure, so that

computing the ‘value of an extra year in normal health’ on the basis of responses to the 1-month question gave a figure more than four times bigger than the value computed on the basis of responses to the 6-month question. (See Chilton *et al.*, 2004, for further details.)

It might be thought that the insensitivity of people’s WTP to differences in the size of benefit is due to them running up against a budget constraint: they may value the larger benefit much more, but simply can’t afford to pay the required multiple. From a conventional perspective, this explanation may have *some* force: in the DEFRA study above, for example, paying six times the 1-month amount would, on average, have involved paying just under 1.5% of *per capita* income for the benefit – a proportion which, although modest, is not trivial.

But budget constraints are by no means the full story. As noted in footnote 1 earlier, an alternative to asking what people are willing to *pay* for small increases in benefit is to ask them what compensation they would be willing to *accept* to offset small losses of benefit. In such WTA questions, the respondent’s answer is not constrained by their budget. However, as reported in Dubourg *et al.* (1994) and Baron and Greene (1996), WTA responses show no greater sensitivity to magnitude than WTP responses.

Overall, then, the history of trying to elicit people’s values for health and safety (and for other goods such as environmental protection) is that the typical respondent does not give sufficient (as judged from the perspective of economic theory and public policy) weight to the relevant features of the scenario, while at the same time paying too much attention to those features of the elicitation procedure that are supposed to be irrelevant. What explains this? Can anything be done which will bring patterns of response into line with the precepts of conventional economic rationality? And if not, is there any alternative basis for providing policy makers with valid, reliable and usable measures of the values to be attached to health, safety and environmental goods?

3. A Psychological Perspective

From a conventional economist’s standpoint, one common reaction to the kind of evidence outlined above is to attribute many of the ‘anomalies’ in the data to diverse shortcomings in the study design: perhaps the scenarios weren’t well enough described; perhaps the question wasn’t sufficiently ‘incentive compatible’; perhaps respondents

needed more time to consider and form a judgment; and so on. Underlying this type of reaction is the belief that people do have ‘true’ values and preferences in there somewhere, if only the researchers were smart enough to devise instruments that will draw them out without causing bias and distortion.

For the economic model to stand a chance of working in practice, two components would seem to be essential. First, for any good or benefit, the utility an individual will experience would have to be accurately anticipated (save for inaccuracies which have the characteristics of random errors with no systematic biases involved). Second, the individual would have to be able to translate such unbiased estimates into expressions of preference or value which would give the same comparisons between goods irrespective of the particular elicitation procedure employed to elicit them. If these conditions hold, what Kahneman (1994) has referred to as *experienced utility* would tend to coincide with *decision utility* and both would map onto money or any other appropriate magnitude scale.

However, there is a body of psychological research which suggests that neither of those conditions is likely to hold to the extent required for the standard economic model to work. Rather, people often fail to recollect and/or predict their experienced utility accurately; and in the absence of a firm foundation of this kind, responses to survey questions may have less to do with well-behaved economic preferences and much more to do with *attitudes* and *affective responses* that may be subject to a variety of procedural and framing effects. The next subsections expand on this point.

Recalling and Predicting Experienced Utility

One graphical way of thinking about experienced utility is to imagine that the *level* or *intensity* of pleasure/pain associated with some activity is represented on the vertical axis, with *time* on the horizontal axis, so that the total utility/disutility experienced during some period is the area under the line tracing the level of intensity at each moment during that period. In the course of our daily lives, many factors may contribute to the way in which this level rises or falls from moment to moment. In reality, it may be difficult even to identify all of these factors, let alone tease out their individual contributions to the overall level. But conceptually at least, one could imagine the total utility (or disutility) attributable to the consumption of a particular good (or bad) being

thought of as the increment (or decrement) in that area over the period during which the consumption of the good is having an effect⁸.

In cases of goods where there is already some past consumption experience, it would seem reasonable to suppose that an accurate prediction of the utility to be experienced in the course of any future consumption of that good would depend, in part at least, on a reasonably accurate recollection of the utility of the previous experience(s).

However, there is evidence that '*remembered utility*' may not be a reliable indicator of experienced utility. Kahneman (2000a) reviews a body of evidence which shows clear tendencies for people's remembered utility – as measured by retrospective ratings of (generally aversive) experiences – to neglect the duration of the experience while being overly sensitive to the *trend* in the profile of the intensity of the experience. The latter effect means that the same total area under the curve is liable to be coded as less unpleasant if the intensity of discomfort falls in later moments as compared with cases where it starts lower but ends higher. When this effect is combined with duration neglect, it turns out to be possible to take an aversive experience lasting, say, 10 minutes, and add an extra minute of milder discomfort, and generate a lower rating of unpleasantness among those exposed to the 11-minute experience than among a comparable sample who endured the smaller total of discomfort entailed by the 10-minute experience. Much of the evidence reviewed⁹ is consistent with a 'peak-end' rule (see Fredrickson and Kahneman, 1993) whereby a past experience is evaluated as a simple average of the most extreme intensity (the peak) and the most recent intensity (the end). This 'representative moment' comes to constitute the retrospective affective evaluation as recorded on some magnitude scale.

In its most pure form, the peak-end rule gives no weight at all to duration. As Kahneman readily acknowledges, this is too extreme: *all other things being equal*, longer

⁸ Although this way of thinking/talking about utility is considered by many orthodox economists to be outdated – even antediluvian – there is a strong similarity between this approach and health economists' depiction of 'Quality Adjusted Life Years' (QALYs), where the vertical axis represents an index of health status (with 'full health' set at 1, 'dead' at 0, and negative values denoting states worse than death). The impact of an intervention on the total area under the curve represents the QALY gain afforded by that intervention, and cost-per-QALY analysis may be used to help inform judgments about which interventions to promote or discourage.

⁹ Kahneman points out that the experimental evidence largely relates to a fairly narrow range of situations, and that other patterns might emerge from other types of episode e.g. those involving a mix of positive and negative experience.

episodes tend to be rated as worse than shorter episodes. But the effect of duration seems to be relatively weak, and *additive rather than multiplicative*: that is to say, unpleasant episodes that last two or three times as long may be scored, respectively, as just 10% or 20% worse rather than as two or three times as bad, which is what would be required in order to accurately reflect the ‘area under the curve’. What is striking about this result is its resemblance to the insensitivity to scope in the CV studies reported in the latter part of Section 2 of this paper. This is a resemblance to which Kahneman and colleagues attach some significance – about which, more later.

The fact that remembered utility may not accurately reflect the actual experience does not bode well for accurate prediction of the utility of future experiences. And indeed, there is plenty of evidence of failures to predict accurately, whether in relation to life-changing events or to things as ordinary as consuming ice cream or listening to recorded music (see Kahneman and Snell, 1990, and other sources cited in Kahneman, 2000a).

It has been suggested that the reasons for failures to predict accurately are somewhat analogous to the reasons for failures to remember accurately: that is, a combination of (a) neglect of the full profile of intensity over the whole duration of the experience with (b) relatively too much weight being placed on the *front* end of that profile.

Taking (a) first, people often appear to underestimate the extent to which they will *adapt* to some new experience. Kahneman cites work by others on paraplegics and lottery winners which indicates that both groups adapt to their changed circumstances. For many (though perhaps not all) of those affected, the initial loss and distress involved in becoming paraplegic reduces in intensity over time as they substitute activities they can do for those that have become precluded. Indeed, for some people the adaptation is such that after a period of time they report themselves as being as happy as they ever were. Viewed in terms of the image of the graph tracing the level of utility from moment to moment throughout the day, the substitution of other sources of utility becomes progressively more important and the time spent encountering and/or thinking about the restrictions due to paraplegia come to weigh less. For lottery winners too there is adaptation. Initially, the ability to avail themselves of all kinds of goods and services is

exhilarating; but it is hard to maintain this ‘high’ as people become accustomed to their new level of wealth. Such adaptation seems to be widespread: numerous other examples are cited in Kahneman and Sugden (2005) which support the idea of a *hedonic treadmill* whereby people strive to increase their real income, but after adaptation are not much happier than they were before; so they strive some more, perhaps achieving some short-term gain – but this too tends to fade; and so on. Yet somehow most of us fail to learn this lesson, so that for many experiences we are liable to underestimate the extent of adaptation and thus fail to estimate accurately the ‘true’ area under the curve.

Turning to (b), the suggestion is that as part of neglecting the intensity of the experience after adaptation, people are prone to attend too much to the front end of the profile: that is, they give too much weight to their perception of the initial experience, and in particular, the contrast with their current position. Not only may the difference between their current state and a new state before adaptation be more extreme than after adaptation, but also the comparison may cause people to focus on particularly contrasting and/or salient attributes, coded as gains and losses relative to the *status quo*.

Thus when asked to predict the (dis)utility of some prospect, two kinds of distorting *focusing effects* may come into play. First, when someone is asked to think about it, a prospect may be given more weight than it will receive in normal daily life when it becomes one of many possible sources of utility and is not receiving such special attention. Second, if the prospect in question involves some change, or *transition*, this may draw particular attention to some features of the prospect that are perceived as gains and losses which not only loom relatively large but where rather greater weight is attached to losses than to gains.

Such considerations have at least two possible implications for CV studies. First, the reality of commissioning research means that most issues are evaluated one-at-a-time in separate studies: one sample of people is asked in one study to think about the value of benefit A; another sample is asked in another study about B; and so on. Even if the sampling procedure is such that the samples are comparable, they are different individuals and at the time of responding, will be focusing either on A or else on B, etc., but will not be considering A as one of many possible benefits which include B, and vice-versa. The pure economic model may *assume* that it is as if people engage in

sophisticated judgments of this kind, but simply including an invocation in the interview script along the lines of “When answering this question, please consider all of the other things you might be spending your money on” is unlikely to do the trick in practice. Hence there is a danger that each separate study will result in each benefit being given greater weight than it would receive in the context of all of the other possible benefits and the many other sources of welfare in normal daily life¹⁰.

Second, the focus on changes and their coding as gains and losses may contribute substantially to the disparity between WTP and WTA. Even if the value of a benefit, when considered in isolation, may be overestimated by WTP responses, there are still constraints operating on WTP: the respondent has to consider losing some money/current consumption in order to effect the acquisition. When the problem is framed the other way round, in terms of WTA, it is the loss of the benefit (or the imposition of some disbenefit) which receives greater weight, while the monetary response is not constrained by the respondent’s wealth.

This is not to say that the disparity is entirely illegitimate: it may well reflect the difference between the way the respondent feels about one direction of change compared with its opposite. But because it is concerned with the evaluation of the *transition* and not with the evaluation of the whole profile associated with the resulting *state*, it may not be an accurate basis for predicting the total utility the respondent would actually experience.

Altogether, then, there appears to be a substantial body of evidence which suggests that both remembered utility and predicted utility are liable to deviate from experienced utility in significant and systematic ways. To the extent that decisions are made on the basis of such impressionistic memories and/or imperfect anticipations, convergence between decision utility and experienced utility is liable to be the exception rather than the rule. This being the case, it would not be surprising to find that evaluations of prospects fail to exhibit the kind of consistency and coherence required by conventional economic theory. But why do such failures manifest themselves in the particular ways reported in section 2 above, where stated money values for health and

¹⁰ Kemp and Maxwell (1993) provide a dramatic illustration of this: respondents who were asked about their WTP to protect against oil spills off the coast of Alaska gave an average response of \$85 when asked about this issue in isolation; when it was presented as a component of a wider programme of public goods, the average WTP was 29 cents.

safety goods were seen to be systematically influenced by theoretically irrelevant features of the elicitation procedure, while at the same time being insufficiently sensitive to certain theoretically key factors? Psychological research offers a number of suggestions, as outlined in the next subsection.

Translating Affective Responses into Money Values

Respondents' generation of money values for the kinds of goods that are the subject of CV studies may be usefully broken into two parts: first, the process of evaluating the good(s) in question, irrespective of how such evaluations are expressed; and second, the mapping of such evaluations to a monetary scale.

Consider first the process of evaluation. If it were the case that people could predict reasonably accurately the experienced utility associated with each prospect, there would be no great problem. But if instead we are dealing with affective responses based on a few 'snapshots' of particular moments of an experience, evaluations may be much more volatile and susceptible to being influenced by the particular features highlighted and the particular angles from which they are viewed.

For example, Hsee (2000) shows that when items are evaluated separately, attributes which are 'important but hard to evaluate' in isolation may be less influential than 'unimportant but easier to evaluate' features. The point can be illustrated by the results of a study cited by Hsee where respondents are asked to say what they would pay (in the range \$10-\$50) for a music dictionary. One subsample is asked to consider a single dictionary (A) which has 10,000 entries and is in mint condition. Another subsample is given the same task for dictionary B, which has 20,000 entries, but also has a torn cover. A third group is told that both dictionaries are available. When the two dictionaries were evaluated separately, A was valued on average at \$24, as opposed to \$20 for B; but when both were considered alongside one another, the ordering was reversed: \$19 for A, but \$27 for B.

What appears to be happening here is that in separate evaluation, the affective response is evoked by the appearance attribute: mint = good, torn = bad. The valuation responses reflect this. Whether some number of entries is good or bad is hard for the non-expert respondents to judge: seen in isolation, 20,000 is 'just a number', and carries much the same affective weight as 10,000 when that number is seen in isolation. Even when the

two dictionaries are seen alongside each other, it still may be hard to judge whether 20,000 is 'good' relative to some absolute yardstick such as the maximum number of entries in the biggest dictionary in existence. But it is easy to see that B has twice as many entries as A; and since the main point of a dictionary is to provide data, B appears to have much more of the key ingredient – against which, a torn cover might be thought to be a rather inconsequential defect. Hence in joint evaluation, B is assigned a greater value.

The operation of affect in separate evaluation can even produce a 'more is less' result. A 24-piece dinnerware set A which is complete and in good condition was valued more when considered in isolation than a 40-piece set B consisting of the same intact 24 pieces as A plus another 16 pieces, 7 of which were intact but 9 of which were damaged. This result appears to have structural similarities with the case cited earlier where adding an extra period of mild discomfort to a painful 10-minute procedure increases the total amount of discomfort 'under the curve' but is evaluated as less aversive.

Hsee notes that the kinds of anomalies illustrated above do not depend on using one method for eliciting separate evaluations and a different method for eliciting joint evaluations: in the examples above, WTP values were used throughout. However, when different methods *are* used, even more discrepancies are liable to occur – see Goldstein and Einhorn (1987) and Tversky *et al.* (1988) for a variety of examples and a more detailed discussion of the possible factors at work.

In addition to all this, asking respondents to express their evaluations in the form of sums of money adds another layer of complexity and allows further room for departures from the conventional economic model.

Kahneman, Ritov and Schkade (1999) point to the body of work in the field of psychophysical measurement which suggests that although there may be a fair degree of agreement among respondents about the *relative* values attached to different stimuli such as different intensities of light or different amplitudes of sound, there may be large differences in the *absolute* values assigned by different people unless some guidance, such as a common modulus, is provided. They argue that asking people to give monetary values for the kinds of goods that are typical of CV studies is, in effect, a special case of magnitude scaling without a modulus.

Such a proposition is most directly applicable to studies that use open-ended questions. Apart from a zero lower bound, there is no other indication of a modulus. From the perspective of the standard economic model, there is no need: people simply (!) estimate the utility of the target good, identify some bundle of currently consumed goods that give the same utility, assess how much money would be released by foregoing those goods to acquire the target, and then report that sum as their maximum WTP. But from the psychological perspective, respondents are trying to express their attitudes to the good on a scale where they have to select their own modulus. Debriefing may give insights into that selection process. Some candidates emerge more often than others: what the good might reasonably *cost to provide* (rather than its benefit to the respondent, which is what the researcher really wants to know); or what the respondent can afford (which often means whatever bit of ‘spare’ income she has that she could spend without making an appreciable dent on current consumption). Both of those considerations are likely to exert some restraining influence on OE responses, while the reference to disposable income may be an important reason why one of the (few) things that regularly shows up as a significant explanatory variable is the respondent’s income¹¹. However, there are many other possible candidates for modulus, and what has been observed in psychophysical scaling *without* a common modulus is also observed in CV studies: namely, a distribution with a long upper tail and substantial unexplained variance.

Of course, many elicitation procedures are not open-ended, and thus may be tacitly suggesting some modulus. Payment cards offer a range, while iterative bidding and dichotomous choice present initial amounts that may act as ‘anchors’. The general proposition from the anchoring literature is that, in the absence of a firm idea about the precise right answer, presenting respondents with some number – even one that seems to have no bearing on the question in hand, such as the digits from one’s social security number – may invite them to think of factors consistent with that number, and hence allow that anchor to exert some pull on the final response. With many CV procedures, the ‘hint’ – however unintended as far as the researchers are concerned – may be seen as

¹¹ This is often claimed as evidence of the ‘expectation-based validity’ of the results (see Bateman et al., 2002, Chapter 8): i.e. as evidence that the results conform with the underlying theoretical model where WTP is expected to be a positive function of income. However, if the same correlation can be explained in terms of the psychological account outlined here, it reduces the force of that claim as evidence of the validity of the conventional economic model.

even stronger: if a particular payment card range is presented, the respondent could reasonably suppose that it is a range which is likely to encompass most people's values and hence, if one is not sure about one's own answer, going for something towards the middle of the range may seem safe and sensible. The first value presented in an iterative bidding exercise, or the *only* value presented in a pure dichotomous choice, may carry even greater weight.

So if what is happening is that many respondents are generating a money response on the rather impressionistic, affective, 'snapshot' basis suggested above, then the range or number presented by the researcher is liable to influence which snapshots come to mind, and hence shape to some extent not only the *best* estimate but also the individual's *min* and *max* judgments. And picking the starting number at random, either via a computer as in the New Zealand study mentioned earlier, or even after an explicit show of shuffling a pack of cards as in the DEFRA study, does not prevent this happening: after all, if a modulus is needed, then whatever number is available, irrespective of how it is generated, may be adopted to meet that need.

At the same time as theoretically irrelevant features of the elicitation procedure are having an undue influence on stated money values, *extension neglect* – giving insufficient attention to the total size of the (dis)benefit of the experience (of which duration neglect is a particular manifestation) – results in lack of sufficient sensitivity to scope. Kahneman (2000a) presents graphs portraying data from other studies which show how stated WTP to reverse species decline displays much the same patterns as ratings of aversiveness to different durations of loudness and of pain: that is, to the extent that people give a higher WTP to reverse a decline of 75% compared with 50% or 25%, the differences are additive rather than multiplicative. This same broad pattern is evident in almost all of the health and safety studies cited in section 2 above, and in many more CV studies of environmental goods. If scope insensitivity is understood in this way, Kahneman *et al.* (1999) argue that it is not simply a reflection of poor study design which could be eliminated by improving the design, but is an *inevitable* consequence of the psychological factors influencing people when they try to generate money value responses. As indicated in section 2, if the aim of a study is to generate some unit value,

such as the ‘value of preventing a fatality’ or the ‘value of a life year’, such insensitivity to scope seriously undermines the whole enterprise.

Overall, then, the message coming from this body of psychological research appears to be discouraging for those wishing to elicit valid and robust measures of economic values and preferences for health, safety and environmental goods. How might/should theorists, applied researchers and policymakers respond to such a message?

4. Implications/Concluding Remarks

At the heart of the problem is the distinction between decision utility and experienced utility, or between what Loewenstein and O’Donoghue (2004) refer to as the *affective system* and the *deliberative system*¹² – the general proposition being that instead of a single internally-consistent all-purpose judgment and decision apparatus, the relative weight of the two systems may vary from one context or decision task to another. With respect to the particular issues discussed in this paper, the deliberative system may be seen as processing decisions much more in the way that the economic model supposes – balancing one possible outcome against another, weighting them by the risks and uncertainties involved and the periods of time during which they will be experienced – while the affective system delivers initial, emotive, more impressionistic and partial reactions that are much more susceptible to ‘effects’ such as framing, focusing and response mode effects of the sorts discussed earlier. While the affective system may have a kind of primacy, having evolved first and usually being activated first in any situation, the deliberative system may be seen as being called into play, monitoring and modifying (and sometimes overriding) affective responses if the broader or longer-term goals of the individual call for it and if sufficient cognitive capacity is available and merited.

With respect to the kinds of tasks involved in stated preference surveys about the value of health, safety or environmental goods, very substantial cognitive demands are entailed by the kind of deliberation needed to satisfy the requirements of the standard economic model: respondents would have to pay sufficient attention to balancing complex and unfamiliar outcomes against each other and give due weight to the small

¹² Loewenstein and O’Donoghue briefly review the history of dual-system models, including several that they consider similar to, though distinct from, the one they are proposing. Kahneman (2003) refers to another such ‘architecture of cognition’, labelled Systems 1 and 2 by Stanovich and West (2000),.

probabilities and temporal considerations involved. On the other hand, the practicalities of survey research – a short interview or questionnaire, often with little opportunity for reflection, with little incentive to do other than be helpful to the interviewer/researcher – are likely to activate the affective system, with only minimal monitoring from the deliberative system. Viewed in this way, it does not seem surprising that the data from such surveys, however carefully the questions are constructed and however sincerely the respondents try to answer them, fail to conform with the economic model.

However, even those who are in broad agreement about the source of the problem do not necessarily share the same view about the implications for practical policy.

One possible way of reacting has been proposed by Sugden (2004, 2005). His argument is that even though consumers may not exhibit the kind of stable and coherent all-purpose preferences assumed by conventional theory, this does not prevent markets from working reasonably well, in the sense of making a wide variety of goods and services available to people and giving them *opportunities* to trade. In his approach, the ‘traditional’ notion of consumer sovereignty – the proposition that individuals should, so far as possible, be free to choose because they are the best judges of their own interests and welfare – is replaced by the notion of value/welfare being derived from consumers being free to take (or decline) opportunities to trade according to whatever they want and are willing to pay for at the time. To the extent that there will still be failures of the competitive market to deal with problems such as damaging externalities and underprovision of public goods, cost-benefit analysis would still have a role in guiding policy makers towards the outcomes that would result from the market if only it could operate in ideal competitive circumstances. Sugden’s argument is therefore that methods of collecting, analysing and reporting the information that is fed into cost-benefit analysis should aim to simulate as closely as possible the information generated by the market.

One way of interpreting this approach¹³ is as follows. It is not only health, safety and environmental goods whose stated values have a large affective component and which may not meet the standard coherence criteria. Many market decisions about goods and services which account for a significant proportion of many people’s wealth, and a

¹³ This is my interpretation/elaboration, no doubt with my ‘spin’ on it: it does not necessarily coincide with Bob Sugden’s own views.

number of people's lifestyle choices which may impact substantially upon their health and happiness, may also be subject to large affective influences. Houses are often purchased after only two or three 'viewings', with first impressions apparently playing a substantial role. The information available when deciding on holiday packages generally gives little basis for a rounded prediction of the utility that will be experienced, and enormous resources are devoted to advertising/promoting many consumer goods – cars, furniture, fashionable clothes, cosmetics, electronic goods – in ways which are aimed at getting attention, amusing and/or arousing, and focusing on 'image', but which are light on the kind of information that an estimate of experienced utility would require.

More generally, consumer durable goods are prime candidates for focusing effects: because they may involve relatively large lump sums, people may give them (or some features of them) more weight when making the purchasing decision than they will subsequently receive in normal daily life, with particular initial emphasis on novel features whose ability to deliver a 'hit' will in fact diminish with familiarity and adaptation. The vulnerability of choices between consumer durable goods to various manipulations has been a fertile area of research (see, for example, Shafir et al., 1993).

Nor are such goods the only major items of wealth or welfare that are apparently susceptible to affective influences. Akerlof and Dickens (1982) made an early contribution to the literature by drawing attention to the ways in which aversion to contemplating the unpleasantness of prospects such as injury at work or old age (and one could extend this to unemployment, ill-health, and so on) might lead to people underproviding for protection and insurance. More recently, Barberis and Thaler (2003) have reviewed the ways in which investment behaviour at both the individual and the aggregate level may depart from what would be expected in a world of rational traders, while Madrian and Shea (2001) and Benartzi and Thaler (2004) have shown how seemingly modest manipulations may greatly affect savings and pension plan behaviour.

It could be argued that if such effects are rife in the market, perhaps we should be less concerned if similar kinds of inconsistencies are being picked up in surveys to elicit values for health, safety and environmental goods. After all, if the purpose of cost-benefit analysis is to guide the allocation of resources not only within the public sector but also between public and private consumption claims, is there anything unreasonable about

using market-mimicking procedures to try to put cost-benefit data on the same footing as the prices generated in the markets that operate to distribute the kinds of goods listed above?

One response may be to argue that this *is* unreasonable, literally: that is, it surrenders to the primitive primacy of emotion over reason, of affect over deliberation, whereas the success of the human species may largely be attributable to the development of faculties which allow us to temper, and if necessary override, the affective system if it is in our broader or longer-term interests to do so. And although this may be too difficult, demanding or costly for people always to achieve at an individual level, that does not mean they would not do it if they could. Indeed, many individuals may look to social institutions to harness a range of skills, expertise and judgment that no single individual could be expected to exercise. They may want public policy to protect and promote their individual interests and welfare when it is difficult for them to do so themselves, and to make the choices on their behalf that they would make if they had the time and resources to deliberate carefully about them. So the fact that individual cognitive capacities are limited and subject to errors and biases would not be grounds for public policy mimicking those limitations.

The problem is, of course, to identify those things that are errors and biases, as opposed to genuine preferences that do not accord with the particular postulates underpinning conventional economic theory, and to get measures of what really would make people better off. Some cases seem uncontroversial. For example, (almost) everyone subscribes to the principle of choosing a dominant option over a dominated one, and (nearly) all respondents make such a choice when the dominance relationship is easy to see¹⁴. The fact that the majority of respondents can be induced to choose the dominated alternative when the dominance relation is cleverly disguised – see Tversky and Kahneman (1986) – does not constitute an argument for policy makers to do other than select the dominant alternative whenever more careful analysis reveals such an alternative to exist. Unfortunately, however, dominance is a rather rare situation. The

¹⁴ If one option is at least as good as another in all respects and strictly better in at least one valued respect, it dominates. The reason for the qualifying brackets is that even when such dominance is transparent to the overwhelming majority, there may be the occasional individual who does not see it, or who selects the dominated alternative, perhaps through inattention or error.

nature of most problems is that competing options are better on different dimensions, so that trade-offs are required: the essence of decision-making is to identify the terms on which those trade-offs are – or should be – made.

If it is accepted that what policy makers should aim to do is enhance as far as possible the actual wellbeing of the population they represent, how should we proceed? One possibility is to try to obtain measures of experienced utility to feed into the public policy arena. Kahneman (2000b) discusses this possibility. The essence of the idea is to chart “a distribution of moment utilities that adequately represents the intended population of individuals, times and occasions” (p.681). He cites as an example Experience Sampling Methodology – see Csikszentmihalyi (1990) and Stone, Shiffman and DeVries (1999) – whereby respondents carrying palmtop computers are prompted at various times during each day to report their current affective state. Other approaches may also be considered: for example, asking people to reconstruct stretches of time or experiences from the day they have just had, with subsequent analysis attempting to find associations between features of the experience and the level of utility reported on some scale. The general idea is that the best estimate of the loss of wellbeing entailed by becoming diabetic, or suffering non-fatal injuries in a road accident, or having a new housing development built on nearby farmland can be obtained by measuring what life is actually like for people who have those experiences compared with people who have not.

However, from a policy perspective there are also some potentially important practical difficulties. Perhaps the most daunting of these is the difficulty of aggregating, across people and across time, the ‘scores’ on whatever scale is used to measure utility. One appealing feature of the conventional economic model – if only decision/remembered/predicted and experienced utility were the same – is that benefits and costs can be measured in a common currency: money. The beauty of this is that if the overall monetary value of any benefits really do exceed the monetary value of the costs, it is possible (in principle) to transfer money from beneficiaries to those who bear the costs in such a way that no-one is worse off and at least some are better off. Moreover, even if there is some spread in the timing of costs and benefits (e.g. if costs are incurred now but benefits are not realized until some time in the future), the monetary measure has the

property that it is amenable to transfers across time: money can be borrowed now to cover the costs and reclaimed from beneficiaries at a later date.

But it is far from clear how a similar analysis could or should be applied when utilities are measured on some other magnitude scale. Consider the QALY – referred to in footnote 8 above – which is perhaps the most widely used measure that is in some respects (although by no means fully) comparable with the notion of experienced utility¹⁵. In order to calculate QALYs, respondents are asked to locate particular health states on a scale which assigns a score of 1 to ‘full health’ and 0 to ‘dead’. In order to estimate population mean indices, it is necessary to assume something about interpersonal comparability – usually, that full health is the same for everyone, as is dead. For experienced utility, this is by no means uncontroversial: the experience of spending a year in full health at the age of 60 is not obviously equally as good as a year in full health at the age of 30. Moreover, even if we accept the calculation of QALYs on this basis and then evaluate different health care interventions on a cost-per-QALY basis, this only allows those alternative interventions to be *ranked* against one another within the domain of health care and subject to the public expenditure budget imposed (by whatever process): in the absence of a monetary value for a QALY, little can be said about the extent to which benefits exceed costs or the point at which benefits equal costs. Thus while the measures may provide some guidance *within* a domain such as health care (although even then, only if one buys the rather strong assumptions involved), they are of limited use when deciding how to allocate resources across domains¹⁶.

Is there scope for closing the gap between experienced utility and decision utility and delivering measures of value which meet some basic requirements of coherent deliberative judgment and which, although unlikely to ever be demonstrably optimal, can be defended as boundedly rational and as likely to advance the wellbeing of the population? There is clearly no consensus about the answer to this question, so what

¹⁵ Current QALY measures are generally based on respondents’ answers to questions involving health states that they have not experienced, and are thus likely to reflect predicted utilities (with all the affective influences involved) rather than experienced utilities.

¹⁶ At the time of writing this, there is a research project just getting underway which will examine the feasibility of assigning a monetary value to a QALY, and which will also explore the extent to which QALYs may be given different weights according to various characteristics of the recipients. Information about the progress of this research will in due course be available on <http://www.publichealth.bham.ac.uk/nccrm>.

follows is intended as a contribution to the debate rather than as a claim to providing a definitive solution.

It would appear to be in keeping with the notion of deliberation as the monitor and organizer of affect to propose a two-part process. The first component would be to establish, by discussion moderated by decision researchers and involving representatives of the population, a list of objectives and reasonable, broadly agreed requirements.

For example, when asked to think about the question carefully, do people subscribe to the view that a three-times bigger reduction in their risk of premature death from cause X is, in terms of personal wellbeing, roughly three times as big a benefit? Does an increase in their life expectancy of three months constitute a three-times bigger gain of wellbeing compared to an increase of one month? Are there some classes of goods to which the proportionality of wellbeing to size of good or length of experience, when deliberated about, *does* apply, and other classes where it does *not* – and if so, how far can reasons/explanations be adduced? For instance, do individuals who like drinking instant coffee consider that the contents of a 300gm jar will give them roughly three times as much coffee drinking pleasure as the contents of a 100gm jar? Do individuals who like to donate to a particular charity feel that a £30 donation makes them feel roughly three times as good as a £10 donation? If the answer to (some) such questions is No, then why? And in such cases, what can we learn about the relationship between wellbeing and the dimensions of the good?

There are plenty of other examples that could be given, but the point is to illustrate the idea that one part of the process is to establish whether there is broad agreement about certain principles to which measures of value should conform if those measures are to guide policies intended to promote wellbeing efficiently.

The other – arguably (even) more difficult – part of the process is to elicit measures of what I will call *considered* utility. What I mean by this is not only that the measures should conform with the principles derived from the first part of the process but also that when their implications for the respondents' worlds are spelled out, they cannot identify any obvious objections to those implications, nor any other obvious way of

organizing their world so as to improve their wellbeing, subject to the resource constraints that obtain.

One aspect of this would involve trying to find ways of getting respondents to contemplate the full profile of the experienced utility from any particular source (making allowance for possible adaptation) *as judged in the context of the other sources of utility that would constitute overall wellbeing*. This would be a demanding task, often not amenable to the kinds of survey practices that are currently the norm, and is likely only to be imperfectly achieved. But it might be possible to learn from the ‘experience sampling’ and/or the ‘typical day reconstruction’ methods mentioned earlier to assist the process. It may also be possible to compare the rankings and ratings of those people who have particular experience – e.g. of spending time in certain health states – with those who do not have such experience, and modify the estimates of value accordingly¹⁷.

In order to obtain measures suitable for the purpose of allocating scarce resources between many competing candidates, it would be necessary for people to give some systematic consideration to the opportunity costs involved. As suggested earlier (see footnotes 10 and 11 and the associated paragraphs), asking people to consider one good in isolation and simply exhorting them to bear in mind all the other things they are, or could be, spending their money on, is not enough. If the purpose is to get considered assessments of the contributions to wellbeing of different public goods relative to one another and to private goods, then it may be necessary to present arrays of such goods and engage people in such tasks as ranking them as sources of wellbeing and then trying to move towards some estimates of the rates of trade-off between different items – always checking for conformity with the general principles emerging from other parts of the process. To the extent that money amounts are a convenient way of scaling and generalizing the trade-offs and feeding them in to cost-benefit analyses, such values might be sought as the culmination of the process.

¹⁷ This may be a rather controversial suggestion that may look dangerously like advocating that we ask people to give us their best shots, then (partially) override them because ‘we know better’. Unfortunately, however, the disparities between predicted and experienced utilities documented earlier do indicate that people’s best shots may not be on target; and what is being advocated is the exercise of judgment to modify those shots, not in the direction of the researchers’ or policymakers’ prejudices but in the direction suggested by those with the relevant ongoing experience. As a safeguard, any such modifications and the reasons for them should be explicit and open to scrutiny and challenge. In the end, though, policymakers will have to make judgments and take responsibility for them.

Such procedures may, of course, be susceptible to ‘effects’ of their own¹⁸. So an important part of the process would involve taking the values obtained from respondents using one set of stimuli and seeing how far they produce predictions about a somewhat different set of stimuli that respondents endorse – or reject.

Of course, conformity with principles reached by deliberation and consistency across option sets is not enough to guarantee the ‘truth’ of the value measures obtained; but, imperfect though they would undoubtedly still be, they may be more persuasive as a basis for allocating resources in line with long term wellbeing than most estimates produced by existing methods which have proved so demonstrably vulnerable to such a wide variety of ‘effects’.

There are, however, two practical considerations that may limit the feasibility of implementing the kind of approach outlined above.

The first is that they are likely to make heavier demands on respondents and researchers, and will therefore be much more costly to implement than the sorts of surveys widely used at present. Interviewers would require a good deal more training and respondents might be required to commit to multiple sessions and/or extended periods of engagement – inputs more at the levels of those invested in randomized controlled trials for a medical interventions, with costs to match.

However, social science has come to be regarded as an inexpensive area of research compared with natural science. Paradoxically, those social scientists who have sought to present themselves as closer to the traditions of natural science, making extensive use of mathematical and statistical tools, involving (often spuriously) high degrees of precision and excessively simplified and mechanistic models of human behaviour, have in the process reinforced the notion of social science research being inexpensive: the complex and difficult stuff is done inside people’s heads and all we social scientists have to do is to tap into the well-behaved, comprehensive and highly-articulated preferences that rational agents are supposed to have evolved – as Fischhoff (1991) put it, ironically, “if we’ve got questions, then they’ve got answers”.

¹⁸ For example, an exercise involving ranking sets of money lotteries and sure sums of money found that one previously robust violation of conventional decision theory was virtually eliminated; but a theoretically irrelevant manipulation produced a different, and no less subversive, anomaly. Details can be found in Bateman et al. (2004), available on request.

What stated preference experience and psychological studies during the past quarter of a century have shown is that this simply isn't the case: human decision processes are as complex and elusive as anything in biology, physics or chemistry, and the research efforts needed to study them effectively are considerable. The stakes are high – the allocation of thousands of millions of pounds/dollars/euros/etc., and the human welfare implications of that allocation, may be significantly influenced by the results of this research – and current attitudes to funding are inappropriate. However, traditions are slow to change, and in the meantime it will be difficult to obtain sufficient resources to explore thoroughly the kind of approach suggested above.

Moreover, even if we were able to develop better methods for estimating the longer-term experienced utility associated with different goods, services or policies, another practical consideration would remain. Many decisions about public policy are subject to political processes – which, for societies with elections every 3-5 years, may be heavily influenced by shorter-term, affective factors. If people neglect duration/extension and focus on their reactions to transitional gains and losses and if their attitudes to politicians and parties have a large emotive, impressionistic component formed by an accumulation of affective responses to particular decisions or policies, then basing policy on what would actually make people better off in the longer term will not necessarily be a recipe for electoral success. Such considerations are no reason for researchers to abandon the attempt to obtain better estimates of wellbeing and develop a fuller and more sophisticated understanding of how people evaluate experiences; but they may be a reason for not supposing that such research, however successful, will find an easy passage through to impacting upon policy and increasing the wellbeing it seeks to measure.

References

- Barberis, N. and Thaler, R. (2003). A survey of behavioural finance. In *Handbook of the Economics of Finance*. Constantinides, G., Harris, M. and Stulz, R. (eds.). Elsevier.
- Baron, J. and Greene, J. (1996). Determinants of insensitivity to quantity in valuation of public goods: contribution, warm glow, budget constraints, availability and prominence. *Journal of Experimental Psychology: Applied*, **2**, 107-25.
- Bateman, I., Carson, R., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Ozdemiroglu, E., Pearce, D., Sugden, R. and Swanson, J. (2002). *Economic Valuation with Stated Preference Techniques*. Edward Elgar.
- Bateman, I., Day, B., Loomes, G., Orr, S. and Sugden, R. (2004). Does a ranking procedure eliminate the usual violations of expected utility theory? Mimeo, presented at Foundations of Utility and Risk conference, Paris, June 2004.
- Beattie, J., Covey, J., Dolan, P., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N., Robinson, A. and Spencer, A. (1998). On the contingent valuation of safety and the safety of contingent valuation: Part 1 – *Caveat Investigator*. *Journal of Risk and Uncertainty*, **17**, 5-25.
- Benartzi, S. and Thaler, R. (2001). Naïve diversification strategies in defined contribution savings plans. *American Economic Review*, **91**, 79-98.
- Benartzi, S. and Thaler, R. (2004). Save more tomorrow: using behavioural economics to increase employee saving. *Journal of Political Economy*, **112**, S164-87.
- Chilton, S., Covey, J., Jones-Lee, M., Loomes, G. and Metcalf, H. (2004). *Valuation of Health Benefits Associated with Reductions in Air Pollution*. DEFRA, UK.

- Csikszentmihalyi, M. (1990). *Flow: the psychology of optimal experience*. New York: Harper and Row.
- Dubourg, W.R., Jones-Lee, M. and Loomes, G. (1994). Imprecise preferences and the WTP-WTA disparity. *Journal of Risk and Uncertainty*, **9**, 115-33.
- Dubourg, W.R., Jones-Lee, M. and Loomes, G. (1997). Imprecise preferences and survey design in contingent valuation. *Economica*, **64**, 681-702.
- Fischhoff, B. (1991): Value elicitation: is there anything in there? *American Psychologist*, **46**: 835-47.
- Fredrickson, B. and Kahneman, D. (1993). Duration neglect in retrospective evaluation of affective episodes. *Journal of Personality and Social Psychology*, **65**, 45-55.
- Goldstein, W. and Einhorn, H. (1987). Expression theory and the preference reversal phenomena. *Psychological Review*, **94**, 236-54.
- Guria, J., Jones-Lee, M., Leung, J. and Loomes, G. (2004). The willingness to accept value of statistical life relative to the willingness to pay value: evidence and policy implications. Forthcoming in *Environmental and Resource Economics*.
- Hsee, C. (2000). Attribute evaluability: its implications for joint-separate evaluation reversals and beyond. In *Choices, Values and Frames*. Kahneman, D. and Tversky, A. (eds). Cambridge: Cambridge University Press.
- Jones-Lee, M., Loomes, G. and Philips, P. (1995a). Valuing the prevention of non-fatal road injuries: contingent valuation vs. standard gambles. *Oxford Economic Papers*, **47**, 676-95.

Jones-Lee, M., Loomes, G. and Robinson, A. (1995b). Why did two theoretically equivalent methods produce two very different values? In *Contingent Valuation, Transport Safety and Value of Life*. Schwab, N. and Soguel, N. (eds). Boston: Kluwer.

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, **150**, 18-36.

Kahneman, D. (2000a). Evaluation by moments: past and future. In *Choices, Values and Frames*. Kahneman, D. and Tversky, A. (eds). Cambridge: Cambridge University Press.

Kahneman, D. (2000b). Experienced utility and objective happiness: a moment-based approach. In *Choices, Values and Frames*. Kahneman, D. and Tversky, A. (eds). Cambridge: Cambridge University Press.

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *American Economic Review*, **93**, 1449-75.

Kahneman, D. and Knetsch, J. (1992). Valuing public goods: the purchase of moral satisfaction. *Journal of Environmental Economics and Management*, **22**, 57-70.

Kahneman, D., Ritov, I. and Schkade, D. (1999). Economic preferences or attitude expressions? An analysis of dollar responses to public issues. *Journal of Risk and Uncertainty*, **19**, 203-35.

Kahneman, D. and Snell, J. (1990). Predicting utility. In *Insights in Decision Making*. Hogarth, R. (ed). Chicago: University of Chicago press.

Kahneman, D. and Sugden, R. (2005). Experienced utility as a standard of policy evaluation. Forthcoming in *Environmental and Resource Economics*.

Kemp, M. and Maxwell, C. (1993). Exploring a budget context for contingent valuation. In *Contingent valuation: a critical assessment*. Hausman, J. (ed.) Amsterdam: North-Holland.

Knetsch, J. and Sinden, J. (1984). Willingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of value. *Quarterly Journal of Economics*, **99**, 507-21.

Loewenstein, G. and O'Donoghue, T. (2004). Animal spirits: affective and deliberative processes in economic behavior. *Mimeo*.

Madrian, B. and Shea, D. (2001). The power of suggestion: inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, **xx**, 1149-87.

Stanovich, K. and West, R. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences*, **23**, 645-65.

Stone, A., Shiffman, S. and DeVries, M. (1999). Rethinking self-report assessment methodologies. In *Well-being: the foundations of hedonic psychology*. Kahneman, D., Diener, E. and Schwarz, N. (eds). Cambridge University Press.

Sugden, R. (1999). Public goods and contingent valuation; alternatives to the neo-classical theory of choice. Chapters 5 and 6 in *Valuing Environmental Preferences*, Bateman, I. and Willis, K. (eds). Oxford University Press.

Sugden, R. (2004). The opportunity criterion: consumer sovereignty without the assumption of coherent preferences. *American Economic Review*, **94**, xxx-xx.

Sugden, R. (2005). Coping with preference anomalies in cost-benefit analysis: a market-simulation approach. Forthcoming in *Environmental and Resource Economics*.

Tversky, A. and Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, **59**, S251-78.

Tversky, A., Sattath, S. and Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, **95**, 371-84.