

Eliciting Measures of Value for Health and Safety

Michael Jones-Lee and Graham Loomes

University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK.

University of East Anglia, Norwich, NR4 7TJ, UK.

Corresponding Author: g.loomes@uea.ac.uk

Abstract

Many transport policies and innovations are liable to have implications for human health and safety. How should such implications be weighed against the other costs and benefits? In particular, when policy makers are undertaking social cost-benefit analysis, what monetary value should be attached to any health or safety components? Standard welfare economics suggests that the answer should reflect people's collective willingness-to-pay (WTP) for any such improvement (or their willingness-to-accept (WTA) compensation for any deterioration). But eliciting such values from the general population presents a number of practical challenges. This talk will describe some of the problems encountered in the course of several large studies conducted during the past 10 years, and consider what we have learned about the nature of people's preferences and the possible implications for public policy in this area. Much of that work has had a transport focus, but the issues raised are relevant to a number of other areas, including environmental health, crime, safety at work and in the home, and the allocation of health care resources.

Keywords: Value of life, health, safety.

1 Introduction

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: that is, those who will bear the costs and/or reap the benefits of those decisions. 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 want – or be required – to consider how the costs that will fall on society compare with the benefits expected to result. In effect, that requires a monetary value to be attached to each component of the costs and benefits, including any changes in expected quality and/or length of life. But how should society ascribe a monetary value to such changes in the quality or length of human life? In this paper, we outline the theoretical principles underpinning the conventional welfare economics answer to this question, and then discuss some of the issues

involved in trying to turn theory into practice, with particular reference to transport safety, but with indications of the broader relevance of the issues to a number of other areas, including environmental health, crime, safety at work and in the home, and the allocation of health care resources.

2 Modelling Individual Decisions

Since the welfare of the individual member of society is the primary building block of standard welfare economics analysis, we begin by considering a basic model of how the typical individual is assumed to handle decisions involving risk and uncertainty, with particular reference to his/her health and safety.

In most standard economic models, a typical 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 or she is well informed about the prices and qualities of the large array of goods and services available to them, it is supposed that each individual will choose a pattern of present and planned future consumption which will bring him or her the greatest overall level of satisfaction/wellbeing (or, in economists' terminology, utility) that their wealth will allow. This is, of course, a highly stylised portrait of a rational economic agent, and should not be taken too literally as an accurate description of any randomly-chosen man or woman in the street. However, the model appeals to the assumption that on average it is as if the population at large operates in this way. 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 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 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 reduction of that magnitude. Suppose that the average answer is £X. If the sample is representative, we can infer that every 100,000 members of the population would, between them, be prepared to pay a total of £X 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 £X x 100,000. Thus if, for example, X turned out to be 12, the figure to be used in road safety cost-benefit analysis to represent the value people place on each prevented fatality would be £1.2m. And 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 question then is to find some robust and reliable way of eliciting the appropriate figures. Unfortunately, this turns out to be a great deal more difficult to do than standard economic theory might lead us to suppose.

The main reason is this. The assumptions of standard theory are probably not too bad an approximation of the way people behave when making routine and frequently repeated purchase-and-consume decisions, such as shopping for food, or choosing travel mode, or going to the hairdresser: in such cases, there is scope for reflecting on the experience of consumption, and some opportunity to try out and compare alternatives, thereby allowing tastes and preferences to be refined and more satisfying choices to be made.

However, when it comes to the valuation of health and safety benefits, the task is much more demanding. Instead of

¹ It may be that some innovations have the potential for an adverse effect on the health/safety of at least some members of a population. For such cases, the standard prescription is to measure 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, which can be regarded as an element on the cost side of any social cost-benefit analysis. 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; but this is not an issue that will be developed in detail in the current paper – the interested reader is referred to Sugden, 1999, and Guria *et al.*, 2004.)

asking about the *certain* consumption *in the near future* of a good with which the decision maker is *familiar*, questions about health and safety are liable to ask people about *small changes in already small risks* of consequences which are generally *unfamiliar and difficult to imagine* accurately and whose *timing is uncertain and possibly quite remote*.

To illustrate the point, imagine you are told that for someone of your age, gender and driving pattern, the risk that you will at some point in your remaining life be involved in a road accident serious enough to leave you paralysed from the waist down is 10 in 100,000. Next, suppose you are asked to consider the most you would be willing to pay to halve that risk. How precise an answer can you give?

Experience suggests that many people find it very hard to home in on a precise figure. On the one hand, they might feel pretty sure that if this risk reduction were going to cost them a one-off payment of £1, they would quite happily pay that. On the other hand, if it were going to cost them £10,000, they might be fairly confident that this was too much and that they would rather tolerate the existing level of risk rather than forego the goods and services that £10,000 could buy them. Logically, therefore, it would seem that such an individual's value for this particular risk reduction must lie somewhere between £1 and £10,000: but to what extent can a respondent narrow the range down, and how confident does he or she feel about their answer?

The essence of the problem appears to be that although the typical inhabitant of an economics textbook has a complete set of highly articulated values which she can access and process quite readily, the typical inhabitant of the real world does not have such well-honed and easily accessible preferences for the kinds of complex and unfamiliar goods which are the subject of the kinds of questions we would like them to answer. Rather, they – or perhaps I should say 'we' – are more likely to have only rather imprecise and poorly-articulated values for many such goods, so that when confronted with questions of the kind indicated above, we have to *construct* our responses, rather than simply take them, already fully-formed, off the shelf. And under the sorts of conditions that apply in surveys – where respondents are asked to give answers in a rather limited period of time and with little opportunity for careful reflection – they may be liable to use whatever simplifying strategies may 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. Unfortunately, all of this may produce patterns of responses which do not conform with what would be expected under standard assumptions, and which pose real problems of interpretation for policy. An illustrated selection of such phenomena is given in the next section, together with some discussion of the challenges they present for policy makers (and for economic theorists, too).

3 Persistent Practical Problems in Direct Willingness To Pay Studies

3.1 Oversensitivity to Supposedly Irrelevant Factors

If respondents had a reasonably firm and precise idea of their preferences (and were trying to answer truthfully), we should expect that different methods of eliciting those preferences should all be expected to yield much the same estimates of whichever values are being sought. However, in circumstances where people's preferences are rather imprecise and only partly formed, features of the elicitation method may exert undesirable systematic effects on responses.

Given this possibility, it is important to try to allow for it and check for it. In particular, rather than try to force respondents to go more or less directly to stating a single precise amount as the maximum they are willing to pay for some benefit, there are various ways of enabling them to express their uncertainty or imprecision about their values. Essentially, this involves asking them to identify amounts they are quite sure they *would* be prepared to pay, other amounts they are quite sure they would *not* be prepared to pay, and the 'band of imprecision' in between – that is, the range of amounts for which they are less than sure whether or not they would pay. The aim is to start by identifying the limits of this band – from the largest amount they are sure they would pay (the lower bound, or *min*, of the band) to the smallest amount they are sure they would not pay (the upper bound, or *max*) – and then obtain some estimate of the point inside the band where the respondent finds it hardest to say whether or not she would pay (their *best* estimate).

Our initial hopes were that, for any particular item being evaluated, the min and max would prove to be reasonably stable and robust, even if the position of the best estimate within that range might be influenced somewhat by the particular elicitation method. However, early tests suggested that this was an overly optimistic expectation. In 1991 we undertook a major study for the UK Department of Transport (DoT) to elicit values for preventing non-fatal road injuries. This study was preceded by several phases of piloting, during which we tested for two possible undesirable influences: starting point effects and range effects.

In both cases, respondents were asked to think about their willingness to pay for various specified reductions in the risks of injuries of different severities. Two such injuries, labelled as W and X, were described as follows.

W

In hospital

- 2-7 days
- Slight to moderate pain

After hospital

- Some pain/discomfort for several weeks

- Some restrictions to work and/or leisure activities for several weeks/months
- After 3-4 months, return to normal health with no permanent disability

X

In hospital

- 1-4 weeks
- Slight to moderate pain

After hospital

- Some pain/discomfort, gradually reducing
- Some restrictions to work and leisure activities, steadily reducing
- After 1-3 years, return to normal health with no permanent disability

In the case of injury W, respondents were asked to think about some safety measure that would reduce their risk of experiencing such an injury during the forthcoming year by 10 in 100,000; for injury X, a separate question asked them to think about a risk reduction of 18 in 100,000. Other reductions in the risks of more severe injuries, labelled R and S but not described here, were also used presented, as well as reductions in the risk of instant death².

To test for starting point effects, interviewers used a procedure involved presenting respondents with a plain circular disc which had a window cut out at the '12 o'clock' position. This window displayed a sum of money, and after describing the particular risk reduction in question, interviewers asked respondents whether they were sure they would be prepared to pay the amount displayed, or whether they were sure they would not pay that amount, or whether they were unsure. Depending on their answer, the amount displayed was adjusted until the respondent's min, max and best estimates were identified.

Prior to interview, respondents were allocated at random to one of two subsamples: in one, the first amount displayed in the window was always £25; in the other, the starting point was set at £75. As mentioned earlier, we had hoped that (at worst) any starting point effect might be limited to influencing just where the best estimate was positioned within the band of imprecision the *best* estimate fell, but would not have any substantial effect on the *mins* and *maxs* – which were, after all, amounts that respondents had asserted they were quite definite about. However, such hopes were comprehensively dashed. The data allowed ten separate pairwise comparisons to be made, and the null hypothesis of no significant starting point effect was rejected at the 1% level in six cases, at

² Further details of the procedures used, the visual aids provided to help envisage the risk reductions, the descriptions of the other injuries in the study, etc., can be found in Dubourg *et al.* (1997).

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. Moreover, this was not just a matter of the position of the best estimate being influenced within otherwise reasonably stable bands of imprecision. In fact, in *every one* of the ten comparisons the mean *max* generated by the £25 starting point was substantially lower than the mean *min* for those initially presented with £75, with the latter being between 35% and 114% higher than the former. In other words, there was not a single case where the comparable bands of imprecision overlapped to any extent whatsoever.

Given the power of the starting point in this procedure, in a subsequent round of piloting we replaced the disc with a ‘payment card’ on which respondents were 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. And although the effect here was somewhat less dramatic, it still represented a striking departure from what might be expected from respondents with robust, well-formed preferences. In the ten possible comparisons between subsample means, the £0-1500 range produced higher mean best estimates in nine cases, while the mean min values for the £0 - £1500 subsamples exceeded mean max values for the £0 - £500 subsamples in five of the ten possible comparisons.

Since that time, we have tried various ways of trying to overcome – or at least, minimise the impact of – such effects, with 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 – effectively, a variant on the ‘window in the disc’ 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. Respondents were assigned at random to one of three starting values – NZ\$20, NZ\$100 or NZ\$400. While there was some evidence that the starting point effect became weaker as respondents progressed through the interview and became more accustomed to the procedure, there was nevertheless evidence that a number of significant influences persisted, even though the extent and the direction of the bias involved was unclear (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 both the starting point and range effects by using a *random card sorting* procedure. Here the idea was that 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, before inviting the respondent to take each one in turn and place it into one of three piles: certainly *would* pay; certainly *would not* pay; and *unsure*. The hope 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, knowing that it was randomly decided which card was turned over first, she would have no reason to attach any particular significance to that initial value. But once again, an effect showed through: although not so marked in the nonlinear regression analysis, the linear correlation between the amount on the first card the respondent saw and their willingness to pay for the package of benefits was positive and highly significant (see Chilton *et al.*, 2004).

Thus there appears to be a considerable body of evidence consistent with the idea that when people’s preferences for unfamiliar and complex goods are only part-formed and imprecise, respondents may be susceptible to a number of ‘cues’ which would be regarded as theoretically irrelevant, but which are nevertheless very difficult to eradicate 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, the magnitude of the benefits being evaluated. It is to some examples of this issue that we now turn.

3.2 Undersensitivity to 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 appears to be much larger and which might therefore be expected to receive a proportionally higher valuation. Such effects have been reported in many studies valuing environmental goods (for examples, see Kahneman and Knetsch, 1992). And although defenders of WTP methods have argued that a number of the instances of these effects can be attributed to poor survey design, and that in many other cases, valuations have been responsive to the size of the benefit (see Carson, 1997), the fact remains that it has 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.

To see what is involved, consider a study which seeks to provide an estimate for preventing road accident fatalities. In the example given in Section 1 above, one way of setting about this is to ask each member of a representative sample 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, then take the average response (£12 in the earlier example) and multiply that by 100,000 to give a VPF of £1.2m. But an alternative design could just as well ask what respondents 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' income, they might be expected to pay approximately 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 we have so frequently observed: that means, roughly speaking, 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, but less than twice as much, as they stated for the smaller benefit (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 our example, people would give an average WTP of no more than £18 for the 3 in 100,000, 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 designers of the survey choose to ask about. If such very different estimates of a VPF can be generated by the same people answering adjacent questions where only the magnitude of risk reduction is changed, which figure should be used for policy? And how much confidence can policy makers have in that number rather than one that is half, or twice, as big?

It might be thought that the problem lies in 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 deep-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 per se, 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 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 prevented during the next five years).

Initially, they 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 we were using 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. Thus even our somewhat blatant attempt to encourage greater sensitivity to the magnitude of the benefit appears to have failed. Moreover, the annual form of question produced a significantly different distribution 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).

More recently, we asked respondents 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 from 1 month to 3 months to 6 months. Thus the benefit was presented as a sure thing, and there was no doubt that respondents were being asked to think of it as being enjoyed by all household members (whose names were noted 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 gives a figure more than four times bigger than the value of an extra year 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 above, 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.

Thus it seems that when presented with unusual and demanding tasks – which is what questions about WTP or WTA for health/safety gains/losses most certainly are – it is difficult for respondents to give appropriate (as judged from the perspective of economic theory and public policy) attention to all of the relevant features of the scenario while ignoring those features of the elicitation procedure that are supposed to be neutral. One possibility would appear to be to simplify the tasks and/or reduce the cognitive burdens in some way – perhaps by using more indirect methods, or breaking procedures down into more digestible elements, or by hanging new values on an existing ‘peg’ that has stood the test of time. However, as discussed in the next subsection, the various alternative approaches used by researchers bring further issues and difficulties to the surface.

4. Alternatives to Direct Willingness To Pay

4.1 Inferring Values from Choices

When respondents are presented with a standard WTP question, what they are in effect being asked to do is this: they are being asked to evaluate the extra satisfaction/utility some innovation will bring them, then judge what other items of their current consumption they could forego up to the point where this would counterbalance the extra utility, and then estimate how much money this foregone consumption would release to be spent on the innovation – this latter amount being the WTP figure we seek to elicit. So not only are they being asked to evaluate some unfamiliar and quite complex scenario to which they have only just been introduced, but they are also being asked to do something that many of us do only occasionally, namely generate a monetary equivalence for a good or service.

It has been suggested that one way of making the task less unusual is to set it up not as a value-generation exercise but as something much more like the consumer choice decision that many of us make dozens if not hundreds of times every week. The argument is that for most goods, what we are presented with is the good itself (or some picture/description of it) and a price, and the decision we have to make is a simple dichotomous choice: do we buy it – yes or no?

To implement this approach in its purest form to elicit the values of the public, it is necessary to recruit a large sample and then randomly allocate respondents to different subsamples and present each respondent with the innovation in question and a single monetary amount, with this amount being varied from one subsample to the next. Each respondent is then asked whether or not they would pay the particular amount assigned to their subsample. By bringing together the responses from all of the subsamples, it is then possible to build up a picture of the relationship between the different sums of money and the proportion of people who say ‘yes’ to each one (the expectation being that as the sum gets lower, more and more will be willing to pay that

amount). Analysis of the relationship allows us to infer an average WTP for whatever is being valued.

Of course, the greater simplicity of the task presented to each respondent means that less information is collected from each person and very much larger samples are required, together with a number of statistical assumptions, in order to infer an overall WTP figure. But that may not be the only issue. There is now a body of evidence to suggest that dichotomous choice questions systematically elicit higher estimates than more open-ended direct WTP questions. It may be that the reason for this is a ‘yea-saying’ effect, whereby respondents with rather imprecise preferences for such goods may suppose (albeit subconsciously) that the designers of the survey consider whatever sum is presented to be a reasonable amount, so that respondents may feel more inclined to say ‘yes’ (especially if, as is often the case, the health/safety/environmental innovation in question is perceived to be a ‘good thing’ and the respondent does not want to appear to be negative or mean-spirited about it). The net result may be to give WTP estimates that overstate people's values – although, of course, this is difficult to know since, as we have already seen, any underlying values are themselves very hard to pin down with much confidence.

4.2 Breaking Down the Task, then Putting the Pieces Together Again

An alternative way of trying to simplify the task is to break it down into elements that respondents may find more manageable and which the researchers can then put back together to produce the estimates required. This strategy was explored in a study described in more detail in Carthy *et al.* (1999), the key features of which were as follows:

- (i) Respondents were first presented with questions designed to elicit both their WTP and their WTA values for the *certainty* of a more easily imaginable injury scenario such as that described as Injury W in subsection 2.1 above. The idea here was that that most respondents could more readily evaluate such a prospect on the basis of their past experience of injury and illness, and that such an evaluation would not be complicated by considerations of manipulating probabilities. Although the prospect of such an injury being experienced with certainty cannot be regarded as sufficiently minor that WTP and WTA should be quite similar, some estimate could still be made of the monetary equivalent of that injury as a weighted average of the two responses.
- (ii) Next, respondents were presented with questions sought to establish how they balanced risks of less serious injuries against risks of more serious consequences, including death³. At this

³ The type of question used here was a form of standard gamble (SG) of the kind widely used to elicit relative utility

stage, respondents were no longer being asked to think about comparing money with health, but were being asked questions framed entirely within the domain of health status and thus focusing on comparisons that might be said to be much more “like with like”.

- (iii) Finally, the various responses were “chained” together to infer the respondent’s WTP to reduce the risk of death. To illustrate with an example, if the certainty of Injury W worked out to be equivalent to a 1 in 300 risk of death, and if the certainty of that injury were assigned a money equivalent of £4,000, then the combination of the two generated a VPF of $300 \times £4,000 = £1.2\text{m}$.

A drawback with this approach is that imprecision, error or bias at any one stage of the procedure can become exacerbated when combined with other links in the chain. We observed this not only in relation to the study reported in Carthy *et al.* (1999) but also in the New Zealand study reported in Guria *et al.* (1999). As a result, the usual pattern in WTP studies, whereby the distribution of values tends to be right-skewed, was exaggerated, with some of the products of the chaining process being so *very* much larger than the rest that there were serious doubts about their reliability. This required us to consider various levels of trimming of the upper and lower tails. This is not uncontroversial, on at least two grounds. First, if the process has produced distortions in some responses serious enough to lead to their exclusion, might it not also have distorted the remaining responses, albeit to a lesser extent? And second, if policy makers are to exclude those responses they judge too large (and/or too small), does this not undermine the principle of acting on the preferences of a representative sample of the population? It has to be acknowledged that both questions raise important concerns for which there is no easy palliative answer. What can be said is that it is important to present all the data, making explicit the degrees of trimming and the assumptions involved; and that in an area where imprecision, error and bias are all in evidence, no pure ‘gold standard’ values are ever likely to be available, and some degree of judgment is always likely to be required.

4.3 Relating New Values to an Existing ‘Peg’

The strategy outlined in the previous subsection involved starting with the monetary value of a certain but less serious injury and then ‘chaining’ up to more serious injuries and death. An alternative strategy is to start with some

measures, including health state utility indices. The essential idea is to compare the certainty or near-certainty of one health state prognosis with an alternative prospect involving some chance of a better outcome but also some complementary chance of a worse outcome. The balance of chances that makes the certainty of the intermediate health (or in this case, injury) state seem equally as good (or bad) as the mix of better and worse outcomes gives a measure of how the utilities of the different states stand relative to each other.

reasonably well-established (or at least, broadly accepted) VPF and try to establish how other values relate to it.

In the study reported in Jones-Lee *et al.* (1995a), the sample was randomised between two ‘treatments’: one was a conventional WTP format, directly eliciting money values for various reductions in the risks of different severities of injury, including death; the other involved SG questions similar to those used in step (ii) of the chaining procedure described in the previous subsection. In theory, both procedures should have generated broadly similar values for the different injuries relative to death; but in practice, the kinds of insensitivities to scope mentioned earlier led to much less differentiation in the WTP data than in the SG data. On the basis of the WTP data, Injury W was estimated to have about one-fifth of the value of preventing a death, whereas on the basis of SG responses, its mean value was one-fiftieth of the value of preventing a death. A follow-up study gave reasons to believe that the SG responses gave a much better estimate of the kinds of trade-offs between injury and death on the roads that most people would subscribe to (see Jones-Lee *et al.* 1995b for details), and it was these responses which provided the basis for the policy decision to value the prevention of a ‘serious’ injury at about 10% of the VPF when conducting cost-benefit analysis⁴.

The study cited above involved taking the VPF that was already in use for road safety policy and assigning other money values on the basis of how the severity of road accident injuries were judged relative to death. That is, the comparisons were made *within* the road safety context. However, given the fact that the roads VPF has now been in use for more than 15 years⁵ and seems broadly acceptable, an obvious question to ask is whether values in other areas of health and safety can be set with reference to it.

The simple default position would be to attach the same value to premature death across all contexts, so that the roads VPF would also be applied to, say, railways, fires, murder, drowning, occupational cancers, food poisoning, air pollution related deaths, and so on. But it is not obvious that this would necessarily accord with people’s values. There is evidence that people regard some types of premature death as worse than others – what Sunstein (1997) refers to as ‘bad deaths’. It may be that differences in certain characteristics of various risks – for example, whether they are associated with lifestyle choices people have made rather than imposed on them by others, or whether they affect younger people in otherwise good health as opposed to older people in already poor health – may cause people to put different values on

⁴ Injuries W and X were at the less severe end of the spectrum, and the overall value of preventing a serious injury was based on a weighted average of these and more severe injuries which were also evaluated in the study.

⁵ In 1988 the UK Department of Transport adopted a figure of £500,000 in 1987 prices; that figure was indexed to inflation and subject to review, and now stands at approximately £1.25m, a figure that it is broadly in line with values in other countries where preference-based values are used, and which has not produced decisions that have caused people to challenge it.

preventing different kinds of premature death. The question is: how to establish the appropriate relativity between the roads VPF and its counterpart in any other context?

For this purpose, the SG format mentioned above is not directly applicable, although in principle a related method – *risk-risk trade-offs* of the kind used by Viscusi *et al.* (1991) – could be used in some instances. The essential idea here is to elicit the change in the risk of death from some cause that respondents consider to be equivalent to some specified change in the risk of death from a ‘reference’ cause: for example, what increase in their risk of death from a train crash would the respondent consider equivalent to, say, a 2 in 100,000 increase in the risk of death from a road accident? But (as the reader will by now not be surprised to discover) there are a number of problems with such questions, including the following: if the respondent does not consider herself to be at risk from certain causes (e.g. quite a few people rarely/never travel by train), the question may lack credibility; and even if applicable to the respondent, the risks involved may be very low and/or difficult to quantify, and the manipulation of such small probabilities may be highly susceptible to biases of various kinds⁶.

An alternative – which has also been used widely to elicit relativities between health state utilities – is the *person trade-off* (PTO) method (see, for example, Nord 1995). A version of this approach was used in a study reported in Chilton *et al.* (2002) to try to establish how the value of preventing deaths from three other causes – train crashes, domestic fires and fires in public places – stood relative to preventing deaths on the roads. Essentially, respondents were asked to consider pairs of safety programmes that would prevent deaths from different causes, prioritise them when they both cost the same and prevented the same number of deaths (10 each), and then say how many deaths the less-favoured programme would have to prevent to be as good as preventing 10 deaths via the more-favoured programme. So, for example, if a respondent would give priority to a programme that prevented 10 deaths on the roads over a programme that would prevent 10 deaths in domestic fires, she would be asked how many domestic fire deaths the latter programme would need to prevent (for the same cost) to be given equal priority with the road safety programme. For a respondent whose answer was ‘15’, the inference drawn was that she would effectively put a VPF on domestic fire deaths that was two-thirds of her VPF for the roads.

While this is a more straightforward task than the risk-risk trade-off, it has drawbacks of its own. First, the status of the response to a PTO question in welfare economic theory is controversial: it is not really asking about respondents’ valuations of their own risks so much as asking them for an opinion about how society should balance different kinds of hazard; and although this may be of interest to policy

⁶ Indeed, even in cases where the probabilities were not *that* small and where the respondent *was* at risk, patterns of response to risk-risk questions departed systematically from those elicited by an SG method that would conventionally be expected to produce much the same results – see Dolan *et al.* (1995).

makers, it does not have the same theoretical and normative foundations as several of the other methods discussed above. In addition, it appears to be susceptible to ‘effects’ of its own. In particular, there was evidence that many respondents were reluctant to move too far from the initial ‘equal numbers of deaths prevented’ position, so that inconsistencies appeared in their responses.

To illustrate, suppose that in one question a respondent had said that preventing 10 road deaths was equivalent to preventing 15 rail deaths, and in a second question that preventing 10 rail deaths was equivalent to preventing 18 domestic fire deaths. Chaining the second answer to the first would suggest that 10 road deaths would equate to 27 fire deaths ($10 \times 1.5 \times 1.8$), implying a fire VPF 37% the size of the roads VPF; but in a direct comparison between roads and domestic fires, such a respondent would typically be reluctant to go that far, stopping perhaps at a ratio of 10:20, implying a fire VPF half that of the roads VPF. It was as if the 10:10 starting point, together with some discomfort about being inequitable, inhibited respondents from differentiating too strongly, and thereby gave estimates which differed, depending on the ‘route’ by which they were calculated.

Moreover, if the question were re-framed (as was done during developmental work) so that the starting point was not in terms of *deaths prevented* but rather in terms of *life years saved*, a similar ‘starting point plus inequality aversion’ effect was observed. For example, if the typical road accident fatality entails a loss of 40 years of life compared with average life expectancy, whereas the typical domestic fire fatality may entail a loss of, say, 20 years, then the starting point may be presented in terms of two programmes each saving the same number of life years – 1,000 – with that involving preventing 25 road deaths or 50 fire deaths. Even a respondent who favours the road programme but is unwilling to go beyond a ratio of 1,000:2,000 will nevertheless effectively be equating 25 road deaths with 100 fire deaths, implying a fire VPF just one quarter of that for the roads. Thus even though PTO appeared to be a less demanding task than either direct WTP for small risk changes or the matching of pairs of risk changes across contexts, it still seemed vulnerable to influences from what should, in theory, be irrelevant manipulations.

Having said that, such questions may provide some additional information and/or checks on the broad acceptability of values elicited via other methods. For example, when the DEFRA study discussed earlier gave the result that the estimated ‘value of an extra year of life expectancy in normal health’ could vary by a factor of more than four, depending on whether it was derived from the 1-month subsample or the 6-month subsample, a qualitative follow-up study used a PTO question to ask respondents to choose between giving 2 people an extra 40 years of life by reducing road accidents or giving 1,000 people an extra month by reducing air pollution⁷. Opinion was fairly evenly divided, and although it would be unwise to place too much

⁷ It was made explicit that both programmes involved about the same total of extra months.

weight on such a small and non-representative sample, there was at least some indication that using a value that would be comparable to the roads VPF was not without support.

5 How Far Have We Got? How Much Further Can We Go?

One possible reaction to the various difficulties and shortcomings listed in the previous two Sections would be to conclude that we are about as likely to establish a robust set of preference-based values of health and safety as we are to discover Shangri-La; and that we should cut our losses and put time and public money into other, more achievable quests. On the other hand, perhaps there is a more positive, constructive interpretation of where we are now and what has been achieved to date.

Certainly, we cannot proceed on the assumption that people respond in aggregate as if they have the kind of highly-articulated, readily-accessible values for health and safety which conventional theory entails. On the other hand, we do not (yet) have a viable alternative model whose descriptive and prescriptive implications have been worked through. But we are accumulating a body of knowledge and experience which enables us (a) to avoid the false estimates we would derive if we mistakenly took the standard model to be true, (b) to identify some sources of bias and try to avoid/control for them and (c) to begin to develop a richer model of the way people's preferences evolve and operate and start to consider the implications for policy of that richer model.

So although there remains a considerable gap between the quality of the data we are currently able to gather and the information we would ideally like to have, and although it continues to be necessary to apply a substantial measure of judgment (ideally, transparent and open to challenge) to those data, it may be argued that this is nevertheless a more coherent approach to the formulation of public policy than to leave it to be determined on an *ad hoc* basis by the hunches of policy makers and/or the interests of particular pressure groups. Worse still, perhaps, is to have policy driven by short-term politically expedient reactions to events and to media pressure, without regard to the wider and longer-term costs involved. Thus although we may not know with as much precision as we would like how the VPF for rail safety stands relative to the VPF for the roads, all of the evidence to date suggests that the multiple of 2.8 used in recent railway safety policy is seriously out of line even with the values of regular rail commuters in the aftermath of several train crashes (see Chilton *et al.*, 2002), while the figure that would be necessary to justify the introduction of Automatic Train Protection on the UK rail network would, in effect, be saying that it is better to prevent 10 deaths in train crashes than more than 100 deaths on the roads – a trade-off to which almost no member of the UK public would subscribe.

Finally, although the imprecision of people's preferences and their vulnerability of their responses to all kinds of unwanted influences is a source of frustration – and occasionally, despair – to those researchers trying to elicit robust values, the other side of this coin is that if a response which implies a VPF of £0.5m seems as plausible to the person giving it as a response which

implies a VPF of £2m, it may be that any value within that range would constitute an acceptable basis for policy. In which case, looking for more precise figures may not be as crucial as the conventional model of preferences might suggest.

As mentioned earlier, UK transport safety policy has for a number of years worked with a VPF of about £1.25m in current prices, supplemented by values for preventing injuries geared to that VPF, and no major problems have been generated by such values. There has also been some progress in estimating values for other fields of transport and for the health effects of air pollution that would seem to be reasonable starting points for cost-benefit analyses in those areas (with the option to modify those values if the decisions they imply can be shown to run contrary to public preferences). Meanwhile, we and our colleagues are currently engaged in exploring further whether there are indeed 'bad deaths' – and if so, how the relevant VPFs might be set to take account of such considerations. And at the time of writing, we are about to embark on studies to see whether money values can be ascribed to measures of health gain in the form of Quality Adjusted Life Years, and whether there are weights that can be assigned to QALYs to represent any desire on the part of the public to favour some recipients or some forms of health care interventions over others. From what has been said in the course of this paper, it would be unrealistic to expect that neat numerical solutions will readily emerge in respect of any of this ongoing research. But it will, we hope, give us further insights into the scope – and limitations – of such measures, add to our understanding of human judgment of values, and take another step or two towards the goal of coherent public sector resource allocation in the field of health and safety which reflects, as best we can, the interests of the population who bear the costs and enjoy the benefits of the decisions made in their name.

6 Acknowledgments

In this paper we draw on work we have undertaken with a number of collaborators, including Jane Beattie, Trevor Carthy, Sue Chilton, Judith Covey, Paul Dolan, Richard Dubourg, Jagadish Guria, Lorraine Hopkins, Hugh Metcalf, Peter Philips, Nick Pidgeon, Angela Robinson and Anne Spencer. Funding has come from the Department of Transport, the Economic and Social Research Council, London Underground Limited, the Ministry of Agriculture, Fisheries and Food, the New Zealand Land Transport Safety Authority, and a consortium comprising the Health & Safety Executive, the Department of Environment, Food and Rural Affairs, the Home Office and Her Majesty's Treasury. However, the views expressed in this paper do not necessarily represent those of any of the individuals or organisations listed above.

7 References

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.
- 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.
- Carson, R. (1997): Contingent valuation surveys and tests of insensitivity to scope. In *Determining the Value of Non-Marketed Goods: Economic, Psychological and Policy Relevant Aspects of Contingent Valuation Methods*. KOPP, R., POMMEREHNE, W. AND SCHWARZ, N. (eds). Boston: Kluwer.
- Carthy, T., Chilton, S., Covey, J., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N. and Spencer, A. (1999): On the contingent valuation of safety and the safety of contingent valuation: Part 2 – The CV/SG “chained approach”. *Journal of Risk and Uncertainty* **17**: 187-213.
- 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.
- Chilton, S., Covey, J., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N. and Spencer, A. (2002): Public perceptions of risk and preference-based values of safety. *Journal of Risk and Uncertainty* **25**: 211-32.
- Dolan, P., Jones-Lee, M. and Loomes, G. (1995): Risk-risk vs. standard gamble procedures for measuring health state utilities. *Applied Economics* **27**: 1103-11.
- 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.
- Guria, J., Jones, W., Jones-Lee, M., Leung, J, Loomes, G. and Keall, M. (1999): *The Values of Statistical Life and Prevention of Injuries*. Draft Report, New Zealand Land Transport Safety Authority.
- Guria, J., Jones-Lee, M., Leung, J, 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*.
- 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. and Knetsch, J. (1992): Valuing public goods: the purchase of moral satisfaction. *Journal of Environmental Economics and Management* **22**: 57-70.
- Nord, E. (1995): The person trade-off approach to valuing health care programs. *Medical Decision Making* **15**: 201-8.
- 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.
- Sunstein, C. (1997): Bad deaths. *Journal of Risk and Uncertainty* **14**: 259-82.
- Viscusi, W.K., Magat, W.A. and Huber, J. (1991): Pricing environmental health risks: survey assessments of risk-risk and risk-dollar trade-offs for chronic bronchitis. *Journal of Environmental Economics and Management* **21**: 35-51.