UNIVERSITY OF NOTTINGHAM

Discussion Papers in Economics

Discussion Paper
No. 00/10

EMBEDDING AND SUBSTITUTION IN WILLINGNESS TO PAY

by Gwendolyn C. Morrison

July 2000

DP 00/10
ISSN 1360-2438

UNIVERSITY OF NOTTINGHAM

Discussion Papers in Economics

Discussion Paper
No. 00/10

EMBEDDING AND SUBSTITUTION IN WILLINGNESS TO PAY

by Gwendolyn C. Morrison

Gwendolyn Morrison is Lecturer, School of Economics, University of Nottingham.

July 2000

Embedding and Substitution in Willingness to Pay

Gwendolyn C. Morrison*
School of Economics
University of Nottingham

Abstract:
The embedding effect has frequently been observed in willingness to pay (WTP) studies. Since policy implications of valuation studies can be very sensitive to the composition of the package being valued, respondents' apparent insensitivity to such fundamental characteristics suggests that it is inappropriate to use the WTP method to evaluate existing or potential targets of public expenditure. This paper presents a model explaining that responses exhibiting embedding, while seemingly inconsistent, still conform to economic theory. The model is tested using data from a national sample survey in which WTP responses exhibit embedding. Examining responses within subject it is shown that, by isolating the different components of the model, the embedding effect can be removed.

Keywords: Willingness to Pay; Embedding Effect; Part-whole bias; Substitutes; Endowment Effect

JEL classification: D11, D61

---

* School of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
email: wendy.morrison@nottingham.ac.uk ; *telephone:* +44 (115) 951-4768 ; *fax:* +44 (115) 951-4159.
Acknowledgements: The data used in this paper is from a UK Department of Transport funded research project conducted under contract to the Transport and Road Research Laboratory (Contract Reference N05910; project reference S2/17B, Road Safety Division). An earlier version of this work appeared in chapter 4 of my DPhil thesis, University of York, UK. My thanks to Graham Loomes, Gavin Mooney, and John Cairns for their helpful comments and suggestions; any errors or omissions remain my own.

An inconsistency that has frequently been observed in studies employing the Willingness to Pay (WTP) method is that people are often WTP the same amount for a basket of goods as for separate components of that basket, or WTP the same amount for very different quantities of the same good (*e.g.,* Kahneman and Knetsch, 1992a; Boyle *et al* 1994; Jones-Lee *et al* 1993; Bateman *et al* 1997; Beattie *et al* 1998). Termed the embedding effect, or part-whole bias, this has very serious implications regarding the viability of the WTP method. Given that policy implications of WTP studies can be very sensitive to the composition and quantity of goods employed in the analysis, the observation that responses are very insensitive to such fundamental aspects of a good's (basket's) description suggests that the WTP method cannot be used as a means of eliciting values from the general public. Responses from studies using the contingent valuation method (CVM) are taken to represent the economic value that individuals and groups attach to goods. But it has been suggested that, given the nature of goods that CVM is typically employed to value, people may actually be stating their WTP for the satisfaction of having contributed to a 'good cause' rather than their value for the good *per se* (Andreoni, 1990; Kahneman *et al*, 1992a).

This paper presents a model that offers an alternative explanation which suggests that most of these seemingly inconsistent responses do conform to economic theory. The model states that WTP is comprised to two components: a generic component shared by all goods considered substitutes, and a specific component which encompasses characteristics particular to each good. A procedure is described for disaggregating WTP into its components—this provides a means of testing the model. The model is then tested using data from a UK national sample survey valuing non-fatal road injuries in which WTP responses exhibit embedding. Examining responses within subject it is shown that, by isolating the different components of the model the embedding effect can be removed.

I. Background

I.A. *The Embedding Effect*

One interpretation of embedding is that the "respondent values a broader or narrower policy package than the one intended by the researcher" (Mitchell and Carson, 1989). This form of part-whole bias is referred to as 'policy package' bias by Mitchell and Carson. This sort of misconception on the part of the respondents could, potentially, be alleviated by more detailed descriptions and by the inclusion of some questions designed to both elicit the respondent's perception of what they are valuing, and explain (or highlight) key attributes that are, and are not, included in the package. So, if this is the source of the anomaly, then there is the possibility of overcoming the problem and this need not suggest that CVM is not viable.

Another explanation for the embedding effect focuses on the fact that CVM is generally used to elicit values for public goods. The proposition is that people obtain a "warm-glow" or "moral satisfaction" from giving to a good cause, and that it is the value of this moral satisfaction that is reflected in the respondents' stated WTP, not the economic value of the good at hand (e.g., Andreoni, 1990; Kahneman and Knetsch, 1992a). This interpretation has much more serious implications for CVM. Whereas policy package bias involves respondents stating their WTP for a broader or narrower package than the one with which they are presented in the question, in the moral satisfaction argument respondents are valuing something other than the good for which a value is being elicited. The method is employed with the express purpose of estimating the economic value of goods for which market prices are not available. If the results obtained are not a reflection of the goods' own value, then the method is not serving its purpose. So, if it is the moral satisfaction that is being valued, then unless it is decided that the value attached to the moral satisfaction associated with contributing to the public good is an adequate appraisal of the good's worth, CVM is inappropriate and should be abandoned in this context.

CVM is frequently used in environmental economics to place a value on actual or proposed changes to the environment. Given the large amounts of money that are sometimes involved and the problems that are frequently reported in the application of CVM, the United States National Oceanic and Atmospheric Administration (NOAA) commissioned a report assessing the reliability of the method. Porney (1994) provides a summary of some of the main points of this report.[1] The panel, chaired by Kenneth Arrow and Robert Solow, provided guidelines intended to minimise the problems associated with the use of CVM as an evaluative tool (NOAA, 1993). The issues of embedding and moral satisfaction (or warm-glow effects) were addressed in this report. It states that if warm-glow effects are present then, stated WTP will incorporate an "element of self-approbation" (p.4607) and "CV responses should not be taken as reliable estimates of true willingness to pay, but rather as indicative of approval for the environmental program in question." (p.4605) So, if observed

[1] Hanemann (1994) discusses many of the empirical problems associated with CVM and offers suggestions as to how to minimise some of these problems.

inconsistencies in WTP responses do reflect moral satisfaction, then WTP is not a good way of eliciting values from the public.[2]

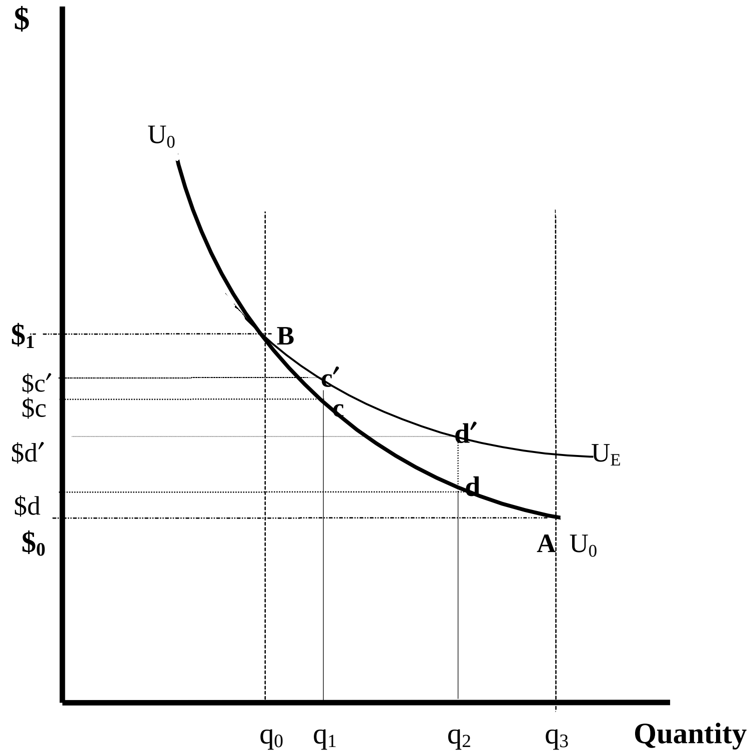
I.B. *Embedding & the Endowment Effect*

Embedding is not the only irregularity frequently reported in the contingent valuation literature. A natural extension of efficiency in exchange is that each individual's WTP for the n^{th} unit of the good must equal their willingness to accept (WTA) payment for that unit since an inequality would suggest that the individual has two different values for the same good—one for buying and one for selling. This is clearly inconsistent. However, economics experiments have frequently revealed WTA exceeding WTP (*e.g.*, Knetsch and Sinden, 1984 & 1987; Kahneman *et al*, 1990; Dubourg *et al*, 1994; Shogren *et al*, 1994; Morrison, 1997a & 1998). An explanation for this anomaly put forward by Thaler (1980), and later Knetsch (1989), is that people value things (whether they be money or a specific good) more once they hold them than before they own them. So, if they are endowed with a good (say a mug), then they are reluctant to part with the mug and will require financial compensation, for the loss of the mug, in excess of the amount that they would have been prepared to pay to acquire it in the first place. This has been termed the "endowment effect." Tversky and Kahneman (1991) argue that such anomalies arise because of loss aversion, where alternatives are viewed in terms of gains or losses relative to the current endowment (Kahneman & Tversky, 1979). As described by Morrison (1997b), the endowment effect can be illustrated as a pivot of the indifference map, in the relevant range, from a reference point (the point of endowment)—referring to figure 1, this would be a clockwise pivot from point 'A' for WTA values and a counter-clockwise pivot from point 'B' for WTP. This realignment of preferences in relation to initial endowment is shown to ultimately lead to WTA exceeding WTP. The endowment effect suggests that utility is a function of money, quantity of goods held, and the loss incurred: $U = f(\text{£}, X, \text{loss})$, where utility is a positive function of the first two arguments and a negative function of loss (Morrison, 1998).

[2] The evidence presented by Kahneman and Knetsch (1992a) (KK) to support their proposition that the embedding effect often observed in WTP responses stems from people valuing the moral satisfaction that they derive from contributing toward a public good has been subject to criticism. Smith (1992), Harrison (1992), and Nickerson (1995) highlight serious flaws in KK's study which serve to undermine the proposition that the

[Figure 1]

The Endowment Effect and Embedding



In figure 1, when the individual is not endowed at either point, the movement from 'B' to 'A' (or vice versa) along the pre-endowed utility curve, U_0, involves a loss of zero. But once endowed at point B and asked to indicate WTP to increase the quantity of X held from q_0 to q_3, the loss associated with changing from bundle B to bundle A is positive—they are exhibiting an aversion to losing some of their initial endowment of money, and consequently associate a lower level of utility with A than with B. A counter-clockwise pivot of U_0 from the point of endowment, B, places B on a higher utility curve, U_E, thus reflecting the alteration of preferences brought on by the endowment. It should be stressed that this realignment of preferences relates only to the quantity range q_0 to q_3 because, in stating their maximum WTP and minimum WTA, respondents are constrained to holding either q_0 or q_3.

What if WTP is elicited for different quantities of the same good between q_0 and q_3 — for example, eliciting WTP for an increase from q_0 to q_2 and then separately asking their

embedding effect is a result of respondents stating their WTP for the moral satisfaction of contributing to a public good.

WTP for an increase in quantity from q_0 to q_1? If WTP responses for the two alternative quantities are equal or roughly equal, then the respondents are exhibiting embedding. This could either be because the respondents are not distinguishing between the two different quantities, or because they have a very rapidly diminishing marginal value for the good. The endowment effect, although a separate concept from the embedding effect, may not be independent of it. Indeed the endowment effect would contribute to empirical findings of embedding by effectively increasing the rate at which marginal value diminishes. Consider WTP for two different quantities of the good in figure 1. Once a person is endowed at point 'B' (with q_0 and $\$_1$), the endowment effect alters the marginal rate of substitution between money and the good—they are willing to give up less money for additional units of the good than they were before they were endowed at 'B'.[3] So rather than being WTP ($\$_1$-$c) for the increase in quantity from q_0 to q_1, being endowed at point 'B' reduces their WTP to ($\$_1$-$c′). That endowment similarly reduces their WTP to increase the quantity of the good that they hold from q_0 to q_2: rather than their pre-endowed WTP of ($\$_1$-$d), endowment at point 'B' reduces their WTP to ($\$_1$-$d′). Consequently the endowment effect can be expected to reduce the difference between WTP for the two different quantities thus contributing to embedding in WTP responses.

II. Samples and Questionnaires

A national sample survey carried out in the UK for the Department of Transport (Jones-Lee *et al*, 1993; O'Reilly *et al*, 1994; Jones-Lee *et al*, 1995) estimated the value of preventing a serious non-fatal road injury using the WTP method with a random sample of 414 people, 48 of whom also answered a follow-up questionnaire which included repeat questions to check for temporal consistency. There is no statistical difference between the groups that answered the follow-up questionnaire and the main sample, and observed inconsistencies are broadly similar in these two groups. The study also obtained expected utility estimates, using von Neumann-Morgenstern standard gambles, for the same injury descriptions from another random sample with 409 completed questionnaires. Details regarding response rates can be found in the main study report (Jones-Lee et al, 1993). The

[3] Conversely, if they were endowed with more of the good and less money (as at point 'B'), then the endowment effect would have the opposite effect: they would be willing to give up less of the good for a given amount of money (i.e., they would require more money to compensate the loss of a quantity of the good).

WTP responses suggest that the indirect value[4] for preventing a serious non-fatal road injury (in 1990 prices) of between £202,288 and £231,874, while the SG method yielded a value of £58,494. This disparity is too large to be attributed to measurement error and, therefore, calls the validity of at least one of the methods into question.

Many studies using the contingent valuation method have identified inconsistencies, such as the WTP/WTA disparity, and embedding or part-whole bias, and some have proposed explanations for these inconsistencies (*e.g.*, respectively, Thaler 1980, Hanemann 1991 or Dubourg et al (1994) for the WTP/WTA disparity; Andreoni (1990) or Kahneman & Knetsch, 1992a, for embedding). This paper uses results from the Department of Transport WTP survey to investigate internal inconsistencies in the WTP method at the individual level, and a model explaining these results is presented and tested.

Two random samples were selected by the Office of Population Censuses and Statistics (OPCS). Interviewers from the Transport and Road Research Laboratory (now the Transport Research Laboratory) presented each sample with a different questionnaire. Each of the questionnaires asked questions regarding the respondent's driving and road accident experience, their demographic characteristics, and asked them to, first, rank a series of injury descriptions (listed in appendix A) and then place those injuries on a visual analogue scale assigning a value of zero to the worst (i.e., most serious) and of 100 to the best. This last part was intended to get the respondents thinking about the relative severity of the injuries.

The two different surveys then diverged. The WTP questionnaire asked respondents to rank a set of five risk reductions (one fatal and 4 non-fatal injuries) in order of priority and then to state their WTP for a reduction in their own risk of suffering each of those injuries. The other questionnaire elicited individual valuations for the same four non-fatal injuries using the SG method. In the WTP questionnaire, respondents were given a card describing an injury, told the annual statistical risk of their incurring that injury, and asked how much they would be willing to pay to reduce that risk by a set amount. Respondents were asked to state three levels of WTP for each of the risk reductions they were offered: the largest amount that they were sure they would pay, the smallest amount they were sure they would not pay, and the amount they felt was the best estimate of what they would pay. This paper analyses the 'best estimate' responses. The full set of risk reductions presented to respondents are listed in

[4] i.e. excluding avoided output losses and medical and emergency service costs which are estimated to be £11,000 in 1990 prices (see O'Reilly et al, 1992).

appendix B, and a sample question from the WTP questionnaire is in appendix C. The final section of the questionnaires contained questions asking the respondent whether he or she had been able, as requested, to ignore any direct financial costs of the accident. In addition, at the end of each questionnaire the interviewers were asked whether or not they felt that the respondent had understood the concept of risk as expressed in the survey (eg, 12 in 100,000), and to write any comments they wished to make concerning the interview.

III. Consistency of WTP Responses

When a random sample of the population is surveyed, it is expected that errors made by individuals will not bias the results of the entire sample—that is, it is assumed that any errors or irregular responses are random in nature. Systematic inconsistencies within an elicitation method indicates that any results obtained may be unreliable and, therefore, caution should be used when applying those results. If a method is internally inconsistent, then there is good reason to question the method's reliability—if not in general, at least in the particular context in which it is being used. Differences between methods may also suggest that one or more of the methods being compared is undependable, or it may simply be a symptom of the way the methods were applied (including, but not exclusively, already widely recognised framing effects).

Internal consistency checks were built into the WTP section of the questionnaire. These checks took the form of asking respondents to rank injury descriptions in order of severity, to prioritise a set of reductions in the risk of incurring those injuries, and then to state their WTP for those risk reductions. In order to check for consistency with respect to scope and consistency over time, WTP for two different sized reductions in the risk of incurring injury S was elicited and a sub-sample (48 of 414 original respondents) were later asked these same questions again during a follow-up interview. Thus, the questionnaire is constructed in a way that allows consistency checks to be performed at the individual level.

The internal consistency of respondents was discussed in the working paper prepared by the researchers who undertook the original Department of Transport study (Jones-Lee *et al*, 1993). These researchers considered the following inconsistencies for the WTP sample: WTP for S4 versus S12, ranking of injuries versus ranking implied by WTP, the ordering of risk reduction priorities versus that implied by WTP responses, and ranking of injuries versus priority ordering of risk reductions. They concluded that, although most respondents were responsible for answers implying at least one preference reversal, few were responsible for many and, therefore, these inconsistencies could reasonably be attributed to error. Rather than repeating their analysis, this section will address consistency checks as they are related to the problem of embedding in WTP.

III.A. *WTP for Road Safety*

The good being evaluated in this study is road safety, or road injuries avoided. The values were elicited to estimate the value of avoiding a statistical serious non-fatal injury. If questions were framed in terms of public safety on the roads, then it could be argued that the WTP responses incorporate a "warm glow" or "moral satisfaction" of contributing to a good cause (Andreoni, 1990; Kahneman *et al*, 1992a). However, the risk reductions employed in the survey are presented as a private good. Respondents were asked their WTP for a safety feature for their car that would reduce only their own risk of injury—it would not effect the safety of anyone else in the car. Therefore moral satisfaction cannot explain any evidence of embedding found in this study.

Mean and median WTP for each of the six risk reductions is presented in the second column of table 1. The aggregate results are consistent in that WTP increases with the common ranking of injury severity. The vast majority of respondents ranked the injuries in the same way, ranking from least to most serious J, W, X, S, R, K. The full injury descriptions are in appendix A, but in brief, J is no injury, W and X are temporary, S and R involve permanent disabilities, and K is fatal. WTP responses were constrained at the lower end with a minimum WTP of £0 and at the upper end with a maximum WTP of "more than £500." Responses of this latter type were assigned a value of £501. WTP for two different sized reductions in the risk of suffering injury S were obtained: one of 12 in 100000 (S12) and one of 4 in 100000 (S4). This provides an internal consistency check for the WTP method in this context. In general, if WTP for the larger risk reduction does not exceed that for the smaller, then this would seem to indicate an inconsistency in the method. Mean WTP for S12 does exceed that for S4 and a Wilcoxon match-pairs signed-ranks test rejects the hypothesis that the two are equal (p=.0000). However, on average people are only WTP 20% more for a risk reduction that is three times larger. It would not be expected that the relationship between WTP and quantity would be directly proportional, but these results do indicate a surprisingly rapidly diminishing marginal value.

Consistency checks at the individual level reveal that a large proportion of respondents did not distinguish between the two quantities in their WTP responses. Table 2 categorises individuals according to the logical consistency of their WTP responses to S4 and S12. A consistent respondent answered in one of the following ways: if S12>S4, the response is strictly consistent; if S12=S4=£0 (ie, the minimum permitted WTP) or S12=S4=£501 (ie, the maximum permitted WTP), then the response is weakly consistent. An inconsistent respondent either stated the same non-limiting WTP for both risk reductions (i.e., £0<S4=S12<£501) or were WTP more for the smaller quantity (S12<S4).

Table 1

Willingness to Pay for Safety: Effect of Inconsistent Respondents

Risk Reduction	**Complete Sample n=414**	**Excludes S4>S12 n=352**	**Excludes S4>S12 & £0<S4=S12<£Max n=212**
fatal injury:			
K4	164.94 *100.00*	165.31 *100.00*	168.29 *100.00*
	(163.49) [18]	(163.83) [2]	(168.77) [1]
permanent non-fatal:			
R4	138.27 *71.50*	138.42 *68.00*	135.05 *60.00*
	(152.83) [20]	(153.77) [1]	(161.18) [0]
S12	124.56 *70.00*	126.28 *72.50*	130.94 *60.00*
	(152.83) [18]	(140.79) [0]	(153.57) [0]
S4	103.00 *50.00*	98.39 *50.00*	84.64 *28.50*
	(131.21) [19]	(130.46) [0]	(136.01) [0]
temporary injuries:			
X12	103.08 *50.00*	102.55 *50.00*	99.81 *41.88*
	(125.56) [18]	(127.41) [0]	(139.01) [0]
W12	94.14 *40.00*	93.65 *40.00*	92.02 *33.76*
	(124.67) [17]	(125.92) [0]	(137.22) [0]

Notes: injuries and risk reductions are defined in appendices A & B.
Mean responses are listed on the left, median responses are on the right in italics, the standard deviation is below in round brackets and missing values in square brackets.

Table 2
Internal Consistency of WTP Responses

	Complete WTP Sample		WTP Follow-up Sub-sample	
	# of Respondents	**Percent**	**Original Responses**	**Follow-up Responses**
Consistent:				
S12>S4	180	43.5%	25	28
S12=S4=£0	17	4.1%	0	1
S12=S4=£max	15	3.6%	2	0
Inconsistent:				
£0<S4=S12<£max	140	33.8%	15	13
S4>S12	43	10.4%	5	2
Missing cases	19	4.6%	1	4

For the complete WTP sample, n=414; WTP follow-up sub-sample, n=48

Most respondents (51.2%) responded in a consistent manner and only 10.4% gave responses that are logically inconsistent. This is very encouraging given that the elicitation questions concerned small probabilities and members of the general public can have great difficulty dealing with probabilities. One third of individuals responded in a weakly inconsistent manner, stating the same positive WTP (<£max) for the two different sizes of risk reduction—embedding is complete in these cases.

The WTP follow-up questionnaire only asked two WTP questions, WTP for S4 and for S12, as opposed to six in the main questionnaire. If inconsistencies are driven by inexperience with the question format or by information over-load in trying to state WTP for six different risk reductions, then one would expect fewer inconsistencies in the follow-up WTP responses than in the original responses. Indeed this is the case. Comparing the original and follow-up responses of the sub-sample in table 2, we see that there were 3 more strictly consistent and 2 fewer logically inconsistent responses in the follow-up. However, the follow-up contained 3 more missing values than the original. One of the additional missing values is a respondent that gave a logically inconsistent response in the original survey. One individual gave strongly inconsistent responses in both the original and follow-up survey, and one respondent that gave a strictly consistent response in the original survey then gave a logically inconsistent response in the follow-up. On the other hand, two individuals giving a strongly inconsistent response in the first instance went on to give a strictly consistent response in the follow-up. So, although there is some improvement in consistency between the original and follow-up surveys, experience (from having already answered the same WTP questions) and a reduction in information over-load (in having only two WTP questions) do not have much of an impact on consistency. Therefore such inconsistencies are more deeply rooted and should be dealt with explicitly.

III. B. *Excluding Logically Inconsistent Respondents*

Economic theory predicts that rational individuals will behave in a way that maximises their utility. This, in turn, involves conforming to particular patterns in their decision making. Firstly, assuming the product in question is a 'good' (i.e., that a person derives positive utility from its consumption), then more of the good will always be preferred to less--this is the axiom of non-satiation. Furthermore, if a consumer values 'A' more than 'B', then the maximum amount they would be willing to pay to obtain 'A' will exceed their maximum WTP to obtain 'B'. The injury scalings derived from WTP responses depend on these assumptions.

Given the large proportion of internally inconsistent responses, the effect of excluding inconsistent respondents from the sample must be explored. Mean and Median WTP were recalculated removing inconsistent respondents from the sample; the results are shown in table 1. The first column of figures shows the results from the full sample, the second lists results excluding only those respondents that were strictly inconsistent (S4>S12), and the third column restricts the sample further excluding strictly and weakly inconsistent respondents.

Comparing the columns in table 1 it can be seen that consistent respondents were, on average, more sensitive to the difference in quantity between S4 and S12 than the inconsistent respondents. Whereas, for the full sample, it would be unpersuasive to argue that the small difference in WTP for S4 and S12 could be attributed to a diminishing marginal rate of substitution, this argument has credibility when considering the responses of only consistent respondents in last column.

III.C. *Relative WTP*

The WTP section of the survey asks respondents to state their maximum WTP to reduce their own risk of incurring an injury by a stated amount for one year. Each risk reduction was considered in isolation from the others on offer. Individuals valued a 4/100000 risk reduction for each of three injuries (K, R, and S) and a 12/100000 risk reduction for each of S, X, and W.[5] The relative WTP responses for the various risk reductions were used to obtain an injury scale ranging from zero to one hundred, where the latter is assigned to the state of 'normal health.' The scaling for each non-fatal injury is calculated with respect to each respondent's WTP to reduce their risk of death by 4/100000 (i.e., K4) under the assumption that death, as the 'absence of life', should receive a zero scaling.

Non-fatal injuries are scaled relative to death in the following way: (i) based on WTP for a 4/100,000 reduction in the risk of a non-fatal injury, such as injury S, the scaled value would be

$$S_{(via\_S4)} = 100 \cdot (K4 - S4)/K4 \tag{1}$$

and, (ii) based on WTP for a 12/100,000 reduction in the risk of a non-fatal injury, such as S, the scaled value would be

$$S_{(via\_S12)} = 100 \cdot \left(\frac{K4 - (S12/3)}{K4} \right). \tag{2}$$

So, if WTP for K4=£10 and WTP for S4=£6, then injury S has an injury scaling of 40, where death and normal health are assigned scaled values of zero and 100 respectively. It is assumed that if, for example, K4=S4=£0, then injuries K and S are equal to 'normal health' and should be given the same scaling as 'normal health', 100.[6] Furthermore, it is assumed that, if K4=£0 but S4>£0, then the scaling is set equal to zero—this prevents negative scalings. The same was applied when calculating scaled values from WTP responses to the other risk reductions.

Table 3 reports the mean relative WTP scalings obtained for each risk reduction. In column two, a large disparity can be seen between the mean scaling of injury S obtained using S4 and using S12. Given that we are dealing with a 100 unit scale, a difference of 30 is cause for concern. If applied to some policy making tool, such as the Quality Adjusted Life Year (QALY), the scalings implied by S4 and S12 could suggest very different resource allocations. For example, say that the injury effects only quality of life not life expectancy, that the average individual suffering this injury can expect 30 more years of life and that their pre-injury state was perfect health (a scale of 100). This is a conservative illustration since those most commonly injured in car accidents are young or middle aged. QALYs lost due to

[5] Injury descriptions and risk reductions are listed in appendices A and B.

[6] To test the results' sensitivity to the assumption that K4=S4=£0 implies a scaling of S via S4 equal to 100, the assumption was changed such that the same conditions lead to the miniumum scaling of zero. On average, the scalings decreased by 3 or 4 points, but the disparity between the S4 and S12 scaling is not reduced (Wilcoxon test result, p=.0000 using either assumption).

this injury would be calculated as 17.4 using the scaled value obtained from S4, or 8.1 using S12. Consequently, a cost-utility analysis of a safety measure to prevent a number of such injuries would look considerably better using the value from S4 than using S12.

Table 3
Relative WTP Scale

	Complete sample n=414	Excludes (S4>S12) n=352	Excludes S4>S12 & £0<S4=S12<£Max n=212
R4	25.076 (31.501) [21]	25.643 (31.944) [2]	33.272 (34.181) [1]
S12	72.916 (17.935) [20]	72.693 (18.253) [2]	73.701 (18.895) [1]
S4	42.014 (35.556) [21]	44.845 (35.470) [2]	59.058 (32.130) [1]
X12	75.989 (19.548) [20]	76.519 (19.083) [2]	79.012 (19.601) [1]
W12	78.168 (20.301) [20]	78.821 (19.625) [2]	81.244 (19.178) [1]
WTP scale S12=S4	z=-14.7869 (p=.0000)	z=-13.5275 (p=.0000)	z=-8.2322 (p=.0000)

Each cell reports the mean scaling together with the standard deviation in round brackets and the missing values in square brackets. Wilcoxon matched-pairs signed ranks test is used to test the hypothesis that the relative WTP scalings arising from S4 and from S12 are equal.

Excluding the respondents stating a WTP for S4 exceeding that for S12 necessarily increases the difference between the mean WTP for those risk reductions. Referring to column three in table 3 it can be seen that this exclusion also slightly reduces the disparity between the mean scalings calculated from S4 and S12. However, also excluding weakly inconsistent respondents (those stating £0<S4=S12<£Maximum), as in column four, greatly reduces that disparity. To determine whether or not this remaining difference is statistically significant the following null and alternative hypotheses must be tested:

H_0: $scaling_{S1}=scaling_{S2}$

H_1: $scaling_{S1}<scaling_{S2}$

The results of one-tailed Wilcoxon matched-pairs signed-ranks test[7] are that, at the 5% level of significance (indeed at α=.01), the null is rejected in favour of the alternative hypothesis for the complete sample, excluding strictly inconsistent respondents and even when weakly inconsistent respondents are excluded. This provides evidence that injury scaling via the WTP method is internally inconsistent, and that excluding logically inconsistent responses does not remove this inconsistency.

III.D. *On Excluding Inconsistent Respondents*

Excluding all inconsistent respondents greatly improves the internal consistency of the WTP method, as can be seen by comparing the second and fourth columns of tables 1 and 3. In particular, it appears to capture differences in WTP for different sizes of risk reduction more effectively and therefore the disparity between the scalings obtained from S4 and S12 is greatly reduced. This indicates that, despite the apparent insensitivity to scope in the complete sample, WTP can be a reliable method of eliciting preferences from the public when respondents understand the task at hand well enough to give responses that are logically consistent.

However, there is an inherent danger in removing responses that do not conform to economic theory—that denies the possibility that it is the method that is incorrect rather than the respondents. Removing almost half of a random sample simply because they do not answer questions in the way that the investigator expected, is unacceptable—the large proportion of inconsistent responses is likely to say more about the method or the way in which it was applied than about the offending respondents. Moreover the sample we would be left with would hardly be random. More to the point, such exclusion is not valuable on empirical grounds, since the disparity between the scaled values from S4 and S12 is still statistically significant. Therefore, it is better to try to understand and explain the presence of seemingly inconsistent responses, particularly when such a large proportion of individuals contradict expected behaviour.

[7] Scales very often have a non-normal distribution, moreover the WTP responses were elicited by presenting respondents with a list of amounts and asking them to indicate which they would and would not pay (see appendix C for an example). Consequently the relative scaled values are not likely to be normally distributed. Examination of the measures of skewness and kurtosis supports this conclusion, therefore a non-parametric equivalent of the correlated samples t-test, the Wilcoxon matched-pairs signed-ranks test is employed.

IV. The Model: Embedding & Substitution

The last section identified an internal consistency of the WTP results—an insensitivity to scope which is termed embedding (or part-whole bias). The embedding effect has frequently been noted in the literature regarding the contingent valuation method (Cummings *et al* 1986; Kahneman *et al* 1992a; NOAA 1993; Bateman *et al* 1997; Beattie *et al* 1998). This term is used to describe the situation where WTP for a single good is much the same as that for a much broader basket incorporating that good. For example, Kahneman & Knetsch (1992a) found evidence of a geographical and categorical embedding effect in median responses to questions such as the value of improving fish stocks in British Columbia versus improving them in Canada, and research into breast cancer versus research of all cancers. Kahneman *et al* concluded that their results were indicative of a WTP for the moral satisfaction of making a contribution toward a public good, rather than representing an estimate of that good's economic value. In the case of public goods, it is reasonable to propose that moral satisfaction is an important argument in the utility function and, therefore, Kahneman *et al*'s finding of a positive correlation between WTP and the level of satisfaction attained by contributing to a public good is not surprising.

However, the 'moral satisfaction' explanation assumes that respondents are not valuing the goods with which they are presented (in surveys or interviews)—this would seem to invalidate the use of CVM for eliciting preferences from the public. In this section, a model of WTP is introduced which provides an alternative explanation for the embedding effect. This model does not have such devastating implications for CVM as the 'moral satisfaction' argument since respondents *are* considered to be valuing the good in question. Furthermore, by specifying a functional form, the model permits the confounding element to be estimated and accommodated to reveal the independent value of the individual goods, whereas the embedding explanation leaves the values of the individual goods being measured hopelessly intertwined.

IV.A. *The Model*

The rationale behind the model, is that a respondent's stated WTP is comprised of two elements: (i) a generic component that applies to all goods that are of the same general type, and; (ii) a component based on the specific (distinguishing) characteristics of the good in question. That is,

$$WTP_{z1}=Generic_{Z}+Specific_{z1} \qquad (3)$$

where, 'z1' is the good for which a value is being elicited and 'Z' is the general category to which that good belongs. For example, 'z1' could be improved fish stocks in BC and 'Z' could be improvement in Canadian fish stocks in general. Alternatively, 'z1' could be research on breast cancer, and 'Z' could be research on all cancers. The generic component is the degree to which an individual or group considers improved fish stocks in one area of Canada to be a close substitute for improved fish stocks all over Canada; the specific component accounts for the fact that the two goods may not be regarded as perfect substitutes. So, the generic component represents the economic value that is common to each of the substitutes, while the specific component captures that value specific to the individual goods. At the limit, if the generic component is equal to the total WTP for two (or more) goods, then those goods are considered to be perfect substitutes—respondents stating the same WTP for S4 and S12 are an example of this limiting case. It follows that, if the generic component is much larger than the specific component, there may be very little absolute (or relative) difference between the WTP stated for each and the responses would display embedding. The 'close substitutes' argument may be more direct with respect to cancer research. A break-through in treating one kind of cancer is likely to be very useful, or even directly applicable to the treatment of other cancers. The observation that WTP for research into breast cancer is roughly equal to that for all cancers, may stem from the recognition of this spillover.

The explicit valuation of such spillovers could be viewed as a manifestation of ‘policy package bias’ (Mitchell *et al* 1989) whereby respondents may misinterpret the question and state their WTP for a larger package than the survey is intended to value. Harrison (1992) points out that even if respondents were clear as to the content of the package that they were valuing, they may regard many of the individual components as substitutes with only one (representative) component good entering their utility function. This is clearly a similar idea to the model presented in equation 3.

The presence of a generic component in stated WTP can be used to explain the apparent insensitivity of WTP responses to size of risk reduction. If equation 3 holds, then the following break-down of stated WTP is implied:

$$S4 = Generic + Specific_{s4} = £50;$$
$$S12 = Generic + Specific_{s12} = £70; \quad \text{and,}$$
$$W12 = Generic + Specific_{w12} = £40,$$

where, the generic component represents the WTP for a reduction in the risk of road injuries in general, the specific component is WTP for a reduction in the risk of incurring a specific injury (S4, S12, or W12), and the amounts listed are the median WTP values. While the risk reduction offered in S12 is three times that offered in S4, the ratio of total WTP for S12 to S4 is 1.4:1 thus indicating either an insensitivity of respondents to size of risk reduction, or a very rapidly diminishing marginal value for reductions in risk of injury S. If, however, a generic component is present (and greater than zero), then the responses are more sensitive to size of risk reduction than a comparison of *total* WTP would suggest. For example, if the generic component equals £35, then the relative *specific* WTP for S12 and S4 is 2.3:1 thus demonstrating much more sensitivity to size of risk reduction.

IV.B. *Embedding and Relative WTP*

The notion of an autonomous component of WTP can also be shown to explain why inconsistencies in the relative WTP arise. First there is the disparity between the scale value for S obtained from S4 and S12, but there is also the scale's insensitivity to injury severity. There is an enormous difference between injury S, which is a permanent disability, and injury W, which is minor and lasts a maximum of four months, yet this difference is not reflected in the injury scale. So, similarly, the mean relative WTP values calculated from S12 and W12 indicate either an insensitivity to injury severity or a rapidly diminishing marginal value of risk reductions with respect to injury severity. A generic component greater than zero could also account, at least in part, for this apparent anomaly.

The way in which the injury scale is calculated from the WTP responses is a convenient starting point for investigating these inconsistencies. The WTP scalings were obtained as in equations 1 and 2. Given those formulae, it is easy to see that the resulting injury scale is heavily dependent on the size of risk reduction involved. For example, K4 and S4 both concern risk reductions of 4/100000, but S12 deals with a reduction three times larger. The scaling method accommodates this difference by simply dividing the WTP value for S12 by 3. Thus, the scale calculation assumes that WTP is directly proportional to the size of risk reduction. There are two outstanding problems with this assumption. Firstly, the results discussed in the previous section showed respondents to be very insensitive to the size of risk reduction offered—44% of the random sample gave logically inconsistent responses of S4>S12 (10%) or £0<S4=S12<£maximum (34%). Secondly, economic theory predicts an eventually diminishing marginal utility of any good and, therefore, a diminishing marginal value.

In the questionnaire, WTP for each risk reduction was elicited in isolation from the others. That is, respondents were asked to state their WTP for each risk reduction as though it were the only risk reduction available. It is feasible, then, that those respondents stating £0<S4=S12<£maximum were stating their full 'safety expenditure budget' (or budget for reducing the risk of injury S) for both S4 and S12. Since they were not asked to allocate their WTP between the various risk reductions their responses to the different WTP questions may say as much about their WTP for road safety in general as about the particular risk reduction in question. Meanwhile, the relative WTP scale calculation is very sensitive to the size of risk reduction used and, therefore, in choosing this quantity the researcher predetermines the results to some extent. So, from both a theoretical and an empirical perspective, the assumption upon which the relative WTP scale calculation is based is seriously flawed.

The most straight forward solution to the relative WTP scale's sensitivity to risk reduction size, might be to offer the same sized risk reduction for each injury in question. After all, if we are interested in obtaining a health status scale for the injuries, then variables that strongly influence the calculated scaling, but that are unrelated to the 'good' we are trying to value, should be omitted from the calculation process. The size of risk reduction, however, cannot be completely excluded from WTP in this context—it is a necessary part of the good's description. How can a person state their WTP when they do not know the quantity of the good that they are purchasing? However, since in estimating an injury scale we are interested only in the relative value of reducing the risk of various injuries, offering the same sized risk reduction across the board might remove one form of scaling bias. This might also focus people's attention on the relative severity of the injury descriptions rather than on the size of risk reduction involved, although both the proportionate change in risk and any discrepancies between stated and perceived risk[8] would remain potential intervening variables.

But, adjustment for different sized risk reductions is not the only factor of the scaling calculation that warrants attention. In addition to the assumption that WTP is directly proportional to risk reduction size, the scaling calculation also assumes that WTP is directly

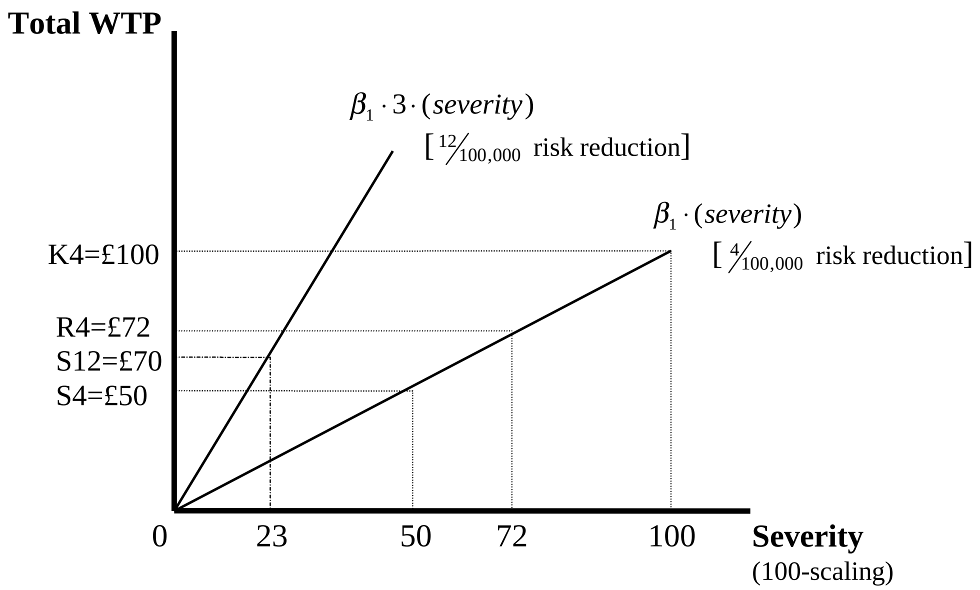
[8] When an individual is told the statistical risk of an injury, they may consider themselves to be at greater or lesser risk than the average individual. For example, if they rarely travel by motor vehicle, then they may consider themselves to have a smaller than average risk of incurring a road injury.

proportional to injury severity such as in the functional form expressed in equation 4. In equation 4, *severity* represents the true subjective severity that the respondent associates with a specific injury—this is the value that health status measures attempt to unveil—and β_1 is the marginal effect on WTP of a one unit increase in injury severity holding the size of risk reduction constant. However, just as theory predicts a diminishing marginal rate of substitution between money and a type of good, there might be a similar phenomenon with respect to injury severity. If this is the case, then the measured relative severity of the various injuries will be distorted as increases in WTP do not proportionately match increases in severity.

$$WTP = \beta_1 \cdot (severity) \quad (4)$$

The relationship between size of risk reduction, injury severity (severity=100-scaling), and WTP that the scaling procedures assume is depicted in figure 2. Equation 4 shows this relationship where the parameter, β_1, is simply multiplied by the ratio of the different sized risk reductions (12/4=3), to accommodate this difference. The median WTP values from the complete sample are used for this illustration. Given that K4=£100, S4=£50, and S12=£70, the severity of injury S is 50 when calculated using S4 and 23 when S12 is used—this corresponds to scalings of 50 and 77 respectively. Since R4=£72, the calculated scaling of injury R is 28.

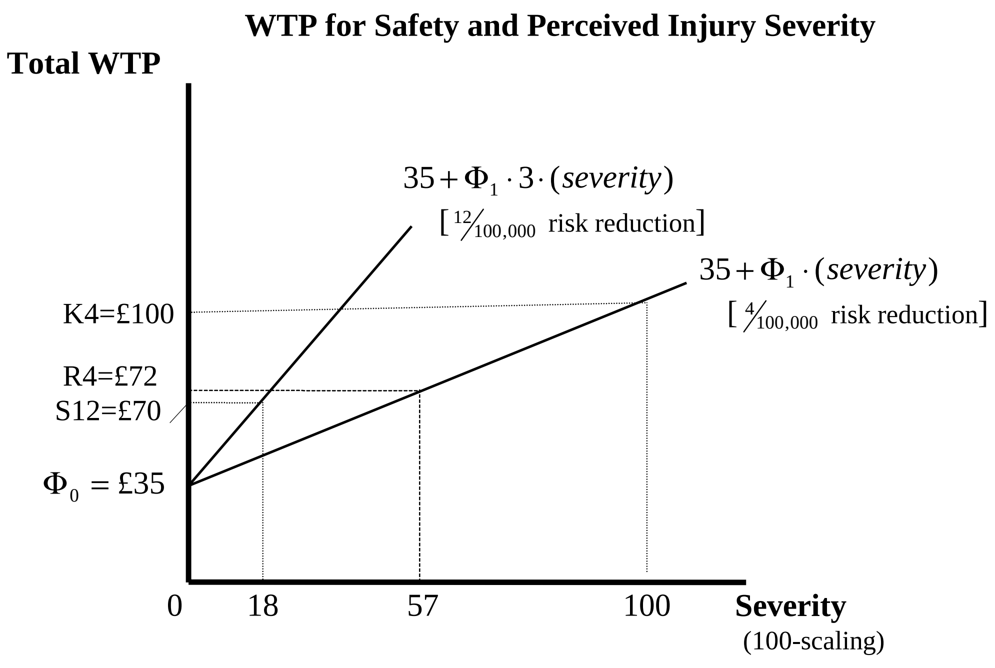
[Figure 2] **WTP and Calculated Severity**



But, even if there were a linear relationship between WTP and injury severity, does it necessarily hold that the calculated scalings are correct? Consider that the relationship takes the form expressed in equation 5 rather than that in equation 4. In equation 5, Φ_0 is WTP for reducing the risk of any road traffic related injury and Φ_1 is the marginal effect of injury severity on WTP for a risk reduction. Since relative WTP is taken to be a measure of perceived severity (scaling=100-severity), then unless Φ_0=0 all non-fatal injury scalings will be biased downwards. Say, for example, that Φ_0=35. This relationship is expressed graphically in figure 3. The level of perceived severity translates to injury scalings of S_p=82, R_p=57, and K_p=0. This is quite different from the situation depicted in figure 2. The reason for this difference is that, although the functional form in equation 5 is linear with respect to injury severity, the injury severity calculated as per equations 1 and 2, is not.[9] This results in the measured injury severities being biased upward and, therefore, the corresponding scalings are biased downward. The relationship between the calculated severity (as in figure 2) and the group's or individual's perception of severity (as in figure 3) is illustrated in figure 4.

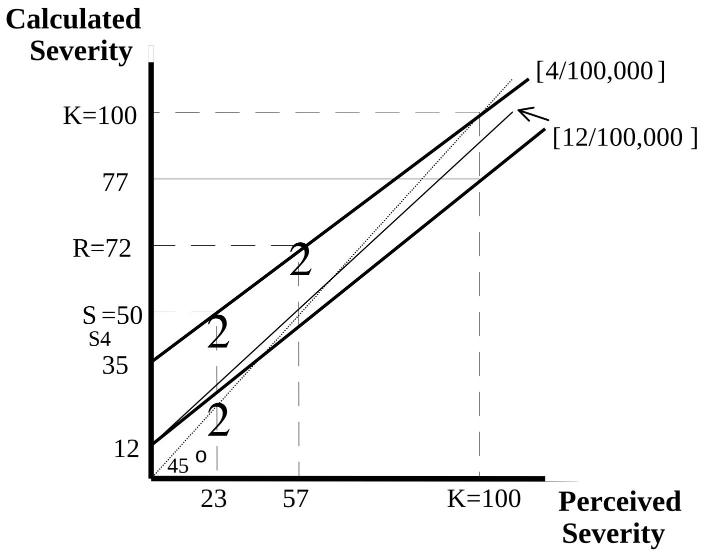
$$WTP = \Phi_0 + \Phi_1 \cdot (severity) \quad (5)$$

[Figure 3]



[9] Specifically, $\frac{WTP_{K4} - WTP_{S4}}{WTP_{K4}} = \frac{(\Phi_0 + \Phi_1 \cdot (severity_{K4})) - (\Phi_0 + \Phi_1 \cdot (severity_{S4}))}{\Phi_0 + \Phi_1 \cdot (severity_{K4})}$

[Figure 4] **Calculated Severity and Perceived Severity: WTP scaling**



where, severity is (100-scale); and, where $K4=\Phi_0+spec(K4)$:
$35=100\cdot\Phi_0/K4$; $12=100\cdot(\Phi_0/3)/K4$; and $77=100\cdot[\Phi_0+3*spec(K4)]/K4$.

The effect of Φ_0 on the calculated injury severity is illustrated in figure 4 with ‘calculated severity’ (where Φ_0 is present) sketched against ‘perceived severity’ (where Φ_0 is excluded). In our example, the only points at which calculated severity is equal to perceived severity are at 33 using 12/00000 risk reductions and at 100 using 4/100000 risk reductions (nb, the latter equality is by definition). It can be seen that the presence of a generic component in WTP results in a different predicted relationship between perceived severity and calculated severity when different sized risk reductions are used. The intercepts are calculated, for each size of risk reduction using the scaling calculations in equations 1 and 2 and the WTP model in equation 5 setting Φ_0=£35, *severity* equal to zero, and using the median value of K4 ($£100=\Phi_0+\Phi_1\cdot severity$=£35+65). For risk reductions of 4/100000 and 12/100000 the intercepts are $\left(\frac{£35+£0}{£35+£65}\cdot 100\right)\approx 35$ and $\left(\frac{(£35+£0)/3}{£35+£65}\cdot 100\right)\approx 12$, respectively. The calculated severity when the perceived severity is the maximum of 100 is, by definition, 100 when a risk reduction of 4/100000 is used

$\left(\frac{K4}{K4}\cdot 100=\frac{£35+£65}{£35+£65}\cdot 100=100\right)$; but, if a reduction of 12/100000 in the risk of death is scaled relative to K4, then this model predicts a calculated severity of death of $\left(\frac{(£35+3\cdot £65)/3}{£35+£65}\cdot 100\right)\approx 77$. So, the lines obtained using a 4/100000 and using a 12/100000 risk reduction are not even parallel—this has important implications in calculating the relative severity of injuries and, thus, the injury scale.

Taking the illustration a step further, imagine that severities are calculated for injuries S and R using these two different levels of risk reduction; let S_4 and R_4 indicate the severities calculated using a 4/100000 risk reduction while S_{12} and R_{12} are those calculated using a 12/100000 risk reduction. Figure 4 predicts three types of internal inconsistency where $\Phi_0>0$: (i) $S_4>S_{12}$ and $R_4>R_{12}$;[10] (ii) $(S_4/R_4)<(S_{12}/R_{12})$; and, (iii) $(R_4-S_4)/R_4<(R_{12}-S_{12})/R_{12}$. These are predictable inconsistencies because the line illustrating calculated severity estimated from a 12/100000 risk reduction is lower than and steeper than that for the 4/100000 risk reduction. So changing the size of risk reduction results in severity calculations (and scalings) that differ in absolute terms, in relative terms, and in terms of proportional difference between injuries. Thus, the model expressed in equation 5 and depicted in figure 4 predicts the observed internal inconsistencies in WTP scaling. The presence of diminishing marginal value would cause the line representing the 12/100000 risk reduction scalings to pivot from the origin toward the 4/100000 line. This reduces the difference between the calculated and perceived severity (type (i) inconsistency) for more severe injuries, but at the same time exacerbate the other two types of inconsistency. Figure 4 also conforms to, and might partly explain, the observed disparity between the scalings obtained from the WTP and standard gamble (SG) methods whereby the WTP scale suggests that injuries are much more severe than the SG scale (O'Reilly *et al* 1994; Jones-Lee *et al* 1995; Morrison 1996).

If stated WTP is comprised of both generic and specific components, and if we wish to construct a scale reflecting the relative values of various injuries, then the generic component (common to all injuries) should be removed to allow the comparison of the specific values associated with each injury. The injury (or health status) scale calculations shown in equations 1 and 2 would be altered to those in equation 6. But, in order to perform this calculation, it is first necessary to isolate the generic component of stated WTP. The next

[10] Indeed, $S_4>S_{12}>$(*perceived severity of S*) and $R_4>$(*perceived severity of R*)$>R_{12}$

section describes an approach to estimating the two components of the generic WTP model and applies it to the WTP data from the UK Department of Transport's study.

$$S_{(via\_S4)} = 100 \cdot \left(\frac{\left(K4 - \Phi_0^{S4}\right) - \left(S4 - \Phi_0^{S4}\right)}{K4 - \Phi_0^{S4}} \right)$$

$$S_{(via\_S12)} = 100 \cdot \left(\frac{\left(K4 - \Phi_0^{S12}\right) - \left((S12/3) - \Phi_0^{S12}\right)}{K4 - \Phi_0^{S12}} \right) \qquad (6)$$

V. Results

V.A. *Disaggregating WTP*

Since WTP responses are constrained in that the minimum response allowed was £0 and the maximum was 'more than £500' (set equal to £501 for analysis), ordinary least squares (OLS) may yield biased and inconsistent parameter estimates. Therefore, Tobit regression analysis was used in an effort to estimate the two components of the generic WTP model.

Respondents stating a WTP for S4 exceeding S12 are omitted from further analysis because they do not appear to have understood the task at hand. Those that did not complete all six WTP questions are also excluded for this reason. This leaves a sample of 350 from the 414 main WTP survey respondents. A subsample of respondents who had answered the SG survey also answered the full WTP questionnaire at a later date. These 53 respondents provide us with another sample on which to test the generic model. The same consistency tests that were applied to the main WTP responses were applied and the results are very similar though only about 6% of this group, compared to 10% in the main WTP sample, stated a WTP for S4 exceeding S12. As for the main WTP sample, these respondents and those that did not answer one or more WTP questions are excluded. This leaves a sample of 44 from the 53 SG follow-up respondents.

The crux of the generic WTP model is that part of an individual's WTP for a good is an amount that they would be WTP for any good of the same general type. In this case, part of WTP for each specific injury risk reduction, is an amount that the respondent would be WTP for any road injury risk reduction—this is the extent to which each risk reduction is a substitute for any other. Since it is expected that respondents are most likely to recall their WTP to the immediately preceding question (and may use this response as a reference point), the sequence in which the risk reductions were presented in the WTP questionnaire is relevant. The first WTP question elicited a value for K4, followed by R4, S12, S4, X12, and finally W12 (see appendix B).

To estimate the generic component, a series of regressions were performed using WTP for reductions in the risk of each non-fatal road injury as the dependent variable, and the WTP response to the preceding question as an explanatory variable (e.g., R4 was regressed on K4; S12 on R4; etc.). In each case, additional regressions were conducted including income and the visual analogue scaling (VAS) corresponding to the injury in the dependent variable. A significance level of 1% is used. For the main WTP sample, in every case with respect to income and all but one case with respect to VAS, these variables were insignificant (always insignificant at α=.14 and typically highly insignificant) while in each case the WTP response to the preceding question remained highly significant (α=.0000 even when VAS and income were included). So, if WTP is related to injury severity, it is generally not related to injury severity as measured by VAS. The insignificance of income of course violates economic theory. However, it should be noted that the survey broke income into 12 ranges and asked individuals to indicate the income range into which they fit. Perhaps, there was insufficient variation in that variable because of that grouping. Nonetheless, income is insignificant so it is excluded from the final estimations. Severity of injury R as measured by VAS (R_{VAS}) is significant at α=.01 in the censored regression of R4 on K4 and R_{VAS}, so the parameter estimate for R1 from this regression is used to disaggregate WTP for R4, however it should be noted that the results are not sensitive to the exclusion of this variable.

Similarly, among the SG follow-up respondents VAS was always highly insignificant and income was always insignificant at α=.01, so Φ_0 is estimated from the regression results which do not include these variables. In one case income was significant at α=.05 (when included in a regression of S12 on R4) and one could argue that, given this is a small sample of 44, perhaps a 1% significance level is too strict. However, the results are not significantly altered when the coefficient on R4 from the regression including income is used instead (as will be seen in table 4).

Estimates of generic WTP, Φ_0, are calculated by multiplying the WTP explanatory variable by its parameter estimate, β. For example, in the main WTP sample $\Phi_0^{R4} = K4 \cdot \beta$ where β is from the regression of R4 on K4 and R_{VAS}, and $\Phi_0^{S12} = R4 \cdot \beta$ where β is from the regression of S12 on R4. The assumption here follows the generic model—that is, responses to each WTP question will share the same generic component, that component being WTP for a reduction in the risk of any road injury. Any variation in stated WTP that is not explained by variations in the WTP response to the previous question is attributed to the differing characteristics of each risk reduction (such as injury severity). Table 4 presents generic WTP estimates for each non-fatal risk reduction and, consequently, that part of WTP that is specific to each of those risk reductions (*i.e.*, $WTP^{i}-\Phi_0^{i}$).

Table 4

Disaggregated WTP: Generic Model

	Main WTP		**SG follow-up**	
	generic	**specific**	**generic**	**specific**
$\mathbf{R_4}$	137.464 (136.233)	24.967 (51.959)	130.243 (135.437)	15.864 (29.744)
$\mathbf{S_{12}}$	120.597 (134.090)	20.915 (33.986)	109.425* (136.322)	18.294* (37.335)
$\mathbf{S_4}$	111.196 (124.610)	8.329 (15.664)	109.015 (131.939)	3.819 (7.856)
$\mathbf{X_{12}}$	88.264 (117.725)	22.423 (47.153)	83.643 (129.228)	13.364 (25.164)
$\mathbf{W_{12}}$	95.533 (119.510)	12.856 (40.854)	95.662 (134.742)	6.291 (21.093)
	Generic WTP: ANOVA			
severe	F=3.572 (sig F=.028)		F=0.358* (sig F=.700)	
all	F=8.443 (sig F=.000)		F=0.749* (sig F=.560)	

Omitting strictly inconsistent respondents, main WTP survey n=352 and SG follow-up survey n=44. The table reports mean estimates with standard deviation in brackets. ‘Severe’ tests H_0: $\Phi_0^{R4}=\Phi_0^{S12}=\Phi_0^{S4}$ and ‘all’ tests H_0: $\left(\Phi_0^{R4}=\Phi_0^{S12}=\Phi_0^{S4}=\Phi_0^{X2}=\Phi_0^{W2}\right)$ against H_1: at least one is not equal.

VAS is significant at α=.01 in the regression of R1 on K1 & R_{VAS}, so the parameter estimate for K1 from this equation was used to estimate Φ_0^{R4}. The results are not sensitive to the exclusion of this variable: ANOVA(severe) F=4.036, sig F=.018; ANOVA(all) F=8.890, sig F=.000.

* Income is an insignificant explanatory variable at α=.01 and therefore is not included in the regression used to estimate Φ_0^{S12}, but it is significant at α=.05. The results change very little when income is included: Φ_0^{S12} =120.376 (==149.964); ANOVA(severe) F=0.256, sig F=.775; ANOVA(all) F=0.824, sig F=.511.

V.B. *Relative WTP – Generic Model*

If the premise of WTP responses being composed of a generic and a specific component is correct, then the scaling calculation must be changed accordingly. Taking the scaling formula as given, the problem lies in the autonomous component of WTP (represented by Φ_0 in equation 5). The generic and specific WTP estimates from table 4 can be applied to equation 6 to calculate an injury scale removing the component of WTP that is assumed to be common to all risk reductions—that is, subtracting the generic component of WTP.

The injury scales obtained from relative WTP responses are shown in table 5 using the original calculation and generic model. Table 5 shows that the removal of the estimated generic component greatly reduces the internal inconsistency of the scaling for injury S arising from WTP for S12 and for S4. A reduction in this disparity is not an achievement in itself—indeed, it is a mathematical certainty. Subtracting a constant from both the numerator and the denominator of a fraction will (providing the constant is less than the numerator which in turn is less than the denominator, and that all three are positive) necessarily decrease the fraction; and the smaller the fraction was to begin with, the greater the decrease. This translates to the scaling calculated from WTP for S12 increasing by less than that calculated from WTP for S4 and, thus, the gap between the two narrows. However, the degree of the reduction in this disparity is encouraging. Wilcoxon matched-pairs signed-ranks tests (also presented in table 5) show that, for both samples, the hypothesis that the scaling of injury S calculated from S4 and from S12 are equal must be rejected for the original WTP scale at α=.01, but cannot be rejected for the generic model scale even at α=.05. So although the original WTP injury scale is internally inconsistent, the generic model scale is not. This finding supports the hypothesis that the embedding effect arises, at least in part, because stated WTP is comprised of a generic and a specific component.

Table 5

Relative WTP Scale: Original and Generic Model

	Original WTP Scale		**Generic Model WTP Scale**	
	main WTP **n=350**	**SG follow-up** **n=44**	**main wtp** **n=350**	**SG follow-up** **n=44**
R_4	25.643 (31.944)	31.627 (35.227)	47.336 (49.908)	54.545 (50.369)
S_{12}	72.693 (18.253)	77.023 (12.775)	75.592 (30.956)	81.358* (29.824)
S_4	44.845 (35.470)	53.707 (34.820)	71.083 (43.036)	84.343 (33.896)
X_{12}	76.519 (19.083)	82.398 (13.406)	80.355 (27.010)	88.029 (19.294)
W_{12}	78.821 (19.625)	81.593 (18.424)	82.305 (29.699)	90.127 (25.363)
relative WTP $S_{12}=S_4$	z=-13.5275 (p=.0000)	z=-4.1532 (p=.0000)	z=-1.8575 (p=.0632)	z=-0.0427* (p=.9659)

Equality of the scaling of S via S4 and S12 is tested using Wilcoxon matched pairs signed-ranks tests.
The results are much the same if VAS is excluded: scale=47.380 (see corresponding notes in table 4).
* The results change very little when income is excluded: scale=88.697 (==26.910), z=-0.3484, p=.7276 (see table 4 notes).

It must be noted that these results are not haphazard estimates—they demonstrate a high level of consistency and robustness. For example, rank order is preserved between original and generic scaling and is the same across the two samples.[11] This is a type of validity check given that this is the same ranking given by the vast majority of respondents in both main surveys of the study in the injury ranking exercise and VAS at the beginning of each questionnaire, as well as the ranking inferred from WTP responses (and from SG reponses in the SG survey). As can be seen in table 5, the test results are not changed when other variables are included in the regressions used to estimate generic WTP.

Although the scalings obtained from the SG follow-up sample are higher for every injury, this is a much smaller sample (n=44 versus n=350) so a difference is not very surprising. Moreover, there is consistency in that both the original and generic model scalings are higher for that sample, and the relative severity between injuries are very similar between the two. That is, comparing the ratio of mean injury scales—R_4:S_{12} (0.62:1), S_{12}:S_4 (1.06:1), and X_{12}:W_{12} (0.98)—all but S_4:X_{12} are roughly equal between the two samples.

[11] The estimates of X and W (both very minor and temporary injuries) in the SG follow-up sample are not statistically different in either the original scaling or the generic model scale calculation.

Given that one is a randomised national survey sample of 350 and the other a small follow-up sample of 44, this consistency yields further confidence in the model.

A further test of the model concerns the generic WTP estimates. If the generic model of WTP is correct, then the estimated generic component of WTP should be equal for all goods in the same category. Hence, it is necessary to test whether the severity and size of risk reduction influences the size of the estimated generic components listed in table 4. To this end, two one-way analyses of variance (ANOVA) are employed.[12] The first tests the hypothesis that all risk reductions share the same generic component of WTP regardless of size of risk reduction or injury severity. Specifically, the null hypothesis that generic WTP is equal for every risk reduction (ie, $\Phi_0^{R4} = \Phi_0^{S12} = \Phi_0^{S4} = \Phi_0^{X12} = \Phi_0^{W12}$, where Φ_0^{i} is the estimate of generic WTP for injury risk reduction *i*) is tested against the alternative hypothesis that they are not all equal.

However, given that two of the non-fatal injuries described in the risk reductions are minor and temporary (X & W) while the other two are serious and permanent, it is possible that respondents treat these as two separate categories of goods. Indeed the generic component of WTP is smaller for the temporary injury risk reductions than for the permanent in both samples—this is intuitively appealing since it says that reducing the risk of a temporary injury is not a good substitute for reducing the risk of a permanent injury. With this in mind, a second ANOVA is run to test the hypothesis that the severe injury risk reductions share the same generic component (ie, H_0: $\Phi_0^{R4} = \Phi_0^{S12} = \Phi_0^{S4}$).

The results of these tests, which are reported in the bottom row of table 4, reveal that for the SG follow-up sample the null could not be rejected for either test even at very high levels of significance. For the main WTP sample, however, the null must be rejected at α =.01 when all risk reductions are compared, but cannot be rejected for the test including only the severe injury risk reductions. Given that the main WTP sample is so much larger, the emphasis is placed on these results although the SG follow-up sample results are even more favourable. Hence, we conclude that the generic component of WTP does not differ significantly between the risk reductions for permanent injuries despite differences both in size of risk reduction and severity of injury. This result lends further support to the generic model of WTP.

[12] This is because, by using the previous question to obtain the generic component estimates, the estimation technique assumes the presence of sequencing (or carry-over) effects in the WTP responses and, so, nonparametric test statistics are inappropriate.

VI. Discussion

This paper describes a model that provides a theoretically consistent interpretation of seemingly inconsistent WTP responses. We have also described a procedure for disaggregating those responses into their individual components in order to facilitate the formation of policy recommendations stemming from WTP surveys. A more simple alternative to this disaggregation procedure lies in the design of contingent valuation surveys. In this survey, respondents were asked to state their WTP for each risk reduction independent of the other risk reductions. They were asked how much they would be willing to pay for K4 without knowing which, if any, other risk reductions they would be offered. Asking individuals their WTP for one risk reduction in isolation may lead them to state an amount close to their full risk reduction budget. Even if they do not go on to state that same amount for each consecutive risk reduction that they are offered, the relative (and nominal) values they attach may be quite different from those they would have given had they been presented with all of the risk reductions simultaneously. If faced with all of the relevant risk reductions at the same time and asked how much of their money they would allocate to each, then the respondent would be forced to consider the relative importance of each of the risk reductions they are asked to value (as recommended by Morrison *et al* 1992). As noted by Bateman *et al* (1997), it is not surprising that values elicited in this way exhibit embedding where the different goods could be viewed as substitutes. An alternative approach to overcoming embedding in WTP studies may be to ask respondents to value all goods together rather than in isolation.

VII. Conclusion

A common result in WTP studies is that the sum of the parts exceeds the whole. A simple explanation for this is found in basic consumer theory—people see different goods from the same general category to be imperfect substitutes. However, this substitutability has overwhelmingly been interpreted in the literature as an internal inconsistency and has thus been used as an argument against the use of WTP in project appraisal (Diamond *et al*, 1994; Jones-Lee *et al*, 1995). The model presented in this paper contests this interpretation and the prescribed procedure for disaggregating WTP provides a means of testing our model.

The model states that WTP is comprised to two components: a generic component shared by all goods considered substitutes, and a specific component which encompasses characteristics particular to each good. The results of the disaggregation of WTP responses and recalculation of relative WTP in accordance with the generic model support the hypothesis that the embedding effect arises, at least in part, because stated WTP is comprised of a generic and a specific component. Firstly, the hypothesis that the same generic WTP estimate is shared by other similar goods cannot be rejected. Secondly, although the original injury scale calculated from relative WTP for risk reductions is internally inconsistent the generic model scale, which is recalculated removing generic WTP, is not.

It is clear that any apparent inconsistencies must be understood before the WTP method can be safely applied in a policy setting context. These results indicate that our approach to disaggregating WTP allows the removal of generic WTP and thus removes one frequently observed internal inconsistency in the WTP method—the embedding effect. Given the obvious need for shadow pricing in project evaluation, and that inconsistencies such as the embedding effect have been used as an argument against the use of WTP in this role, these results have important implications for the future of public project appraisal.

References

Andreoni, J., 1990, Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. Economic Journal 100, 464-77.

Bateman, I.; Munroe, A.; Rhodes, B.; Starmer, C.; and Sugden, R., 1997, Does Part-whole Bias Exist? An Experimental Investigation, Economic Journal 107(441), 322-32.

Beattie, J.; Covey, J.; Dolan, P.; Hopkins, L.; Jones-Lee, M.; Loomes, G.; Pidgeon, N.; Robinson, A.; 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, no. 1, 5-25.

Boyle, K.J.; Desvousges, W.H.; Johnson, F.R.; Dunford, R.W.; and Hudson, S.P., 1994, An Investigation of Part-Whole Biases in Contingent Valuation Studies, Journal of Environmental Economics and Management 27, 64-83.

Broome, J., 1993, QALYs, Journal of Public Economics 50, 149-67.

Cummings, Brookshire, and Schulze (eds.), 1986, Valuing Environmental Goods: An assessment of the contingent valuation method (Rowman and Allanheld, Totawa, N.J.).

Diamond, P.A. and Hausman, J.A., 1994, Contingent Valuation: Is Some Number Better than No Number? Journal of Economic Perspectives 8(4), 45-64.

Dubourg, W.R.; Jones-Lee, M.W.; Loomes, G., 1994, Imprecise Preferences and the WTP-WTA Disparity, Journal of Risk and Uncertainty 9, 115-33.

Hanemann, W.M., 1991, WTP and WTA: How Much Do They Differ? American Economic Review 81, 635-47.

Hanemann, W.M., 1994, Valuing the Environment through Contingent Valuation, Journal of Economic Perspectives 8(4), 19-43.

Harrison, G.W., 1992, Valuing Public Goods with the Contingent Valuation Method: A critique of Kahneman and Knetsch, Journal of Environmental Economics and Management 23, 248-57.

Jones-Lee, M.; Loomes, G.; O'Reilly, D.; and Philips, P., 1993, The Value of Preventing Non-fatal Road Injuries: Findings of a willingess-to-pay national sample survey (Working Paper WP/SRC/2, Transport Research Laboratory, Crowthorne, Berks.)

Jones-Lee, M.; Loomes, G.; and Philips, P.R., 1995, Valuing the Prevention of Non-Fatal Road Injuries: Contingent Valuation vs Standard Gambles, Oxford Economic Papers 47, 676-95

Kahneman, D. and Knetsch, J.L., 1992a, Valuing Public Goods: the purchase of moral satisfaction, Journal of Environmental Economics and Management 22, 57-70.

Kahneman, D. and Knetsch, J.L., 1992b, Contingent Valuation and the Value of Public Goods: Reply, Journal of Environmental Economics and Management 22, 90-2.

Kahneman, D.; Knetsch, J.L.; and Thaler, R., 1990, Experimental tests of the Endowment Effect and the Coase Theorem, Journal of Political Economy 98(6), 1325-48.

Kahneman, D. and Tversky, A., 1979, Prospect Theory: An Analysis of Decision under Risk, Econometrica 47(2), 263-91.

Knetsch, J.L., 1989, The Endowment Effect and Evidence of Nonreversible Indifference Curves, American Economic Review 79(5), 1277-84.

Knetsch, J.L. and Sinden, J.A., 1984, Willingness-to-Pay and Compensation Demanded: Experimental Evidence of Unexpected Disparity in Measures of Value, Quarterly Journal of Economics 99, 507-21.

Knetsch, J.L. and Sinden, J.A., 1987, The Persistence of Evaluation Disparities, Quarterly Journal of Economics 102, 691-5.

Mitchell, R.C. and Carson, R.T., 1989, Using Surveys to Value Public Goods: the Contingent Valuation Method (Resources for the Future, Washington, D.C.).

Morrison, G.C. and Gyldmark, M., 1992, Appraising the Use of Contingent Valuation, Health Economics 1(4), 233-43.

Morrison, G.C., 1996, An Assessment and Comparison of Selected Methods for Eliciting Quality of Life Valuations from the Public (D.Phil. Thesis, University of York, UK).

Morrison, G.C., 1997a, Willingness to Pay and Willingness to Accept: Some evidence of an endowment effect, Applied Economics 29, 411-7.

Morrison, G.C., 1997b, Resolving Differences in Willingness to Pay and Willingness to Accept: Comment, American Economic Review 87(1), 236-40.

Morrison, G.C., 1998, Understanding the Disparity between WTP and WTA: endowment effect, substitutability, or imprecise preferences? Economics Letters 59(2), 189-94.
National Oceanic and Atmospheric Administration, 1993, Report of the U.S. NOAA Panel on Contingent Valuation.', K. Arrow & R. Solow [co-chairs]), Federal Register 58(10), 4601-14.
Nickerson, C.A.E., 1995, Does Willingness to Pay Reflect the Purchase of Moral Satisfaction? A Reconsideration of Kahneman and Knetsch, Journal of Environmental Economics and Management 28, 126-33.
O'Reilly, D., 1992, Costing Road Traffic Accidents: The value of lost output (Transport Research Laboratory Working Paper, Crowthorne).
O'Reilly, D.; Hopkin, J.; Loomes, G.; Jones-Lee, M.; Philips, P.; McMahon, K.; Ives, D.; Soby, B.; Ball, D. and Kemp, R., 1994, The Value of Road Safety: UK research on the valuation of preventing non-fatal injuries, Journal of Transport Economics 28(1), 45-59.
Portney, P.R., 1994, The Contingent Valuation Debate: Why economists should care, Journal of Economic Perspectives 8(4), 3-17.
Shogren, J.F.; Shin, S.Y.; Hayes, D.J. and Kliebenstein, J.B., 1994, Resolving Differences in Willingness to Pay and Willingness to Accept, American Economic Review 84(1), 255-70.
Smith, V.K., 1992, Arbitrary Values, Good Causes and Premature Verdicts, Journal of Environmental Economics and Management 22, 71-89.
Thaler, R., 1980, Toward a Positive Theory of Consumer Choice, Journal of Economic Behavior and Organization 1, 39-60.
Tversky, A. and Kahneman, D., 1991, Loss Aversion in Riskless Choice: A Reference Dependent Model, Quarterly Journal of Economics 106, 1039-61.

APPENDIX A: INJURY DESCRIPTIONS

K	X
-immediate unconsciousness, followed shortly by death **R** **In Hospital** -several weeks, possibly several months -moderate to severe pain **After Hospital** -continuing pain/discomfort for the rest of your life, possibly requiring frequent medication -substantial and permanent restrictions to your work and leisure activities--possibly some prominent scarring **S** **In Hospital** -1 to 4 weeks -moderate to severe pain **After Hospital** -some pain gradually reducing, but may recur when you take part in some activities -some restrictions to leisure and possibly some work activities for the rest of your life	**In Hospital** -1 to 4 weeks -slight to moderate pain **After Hospital** -some pain/discomfort, gradually reducing -some restrictions to work and leisure activities, steadily improving -after 1-3 years, return to normal health with no permanent disability **W** **In Hospital** -2 to 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 **J** -your normal state of health

APPENDIX B: RISK REDUCTIONS USED IN THE CONTINGENT VALUATION SURVEY

1) **K4** INJURY K: **Current annual risk** 8 in 100,000 **Reduced by** 4 in 100,000 (8 in 100,000 DOWN TO 4 in 100,000)	4) **S4** INJURY S: **Current annual risk** 24 in 100,000 **Reduction by** 4 in 100,000 (24 in 100,000 DOWN TO 20 in 100,000)
2) **R4** INJURY R: **Current annual risk** 16 in 100,000 **Reduced by** 4 in 100,000 (16 in 100,000 DOWN TO 12 in 100,000)	5) **X12** INJURY X: **Current annual risk** 30 in 100,000 **Reduced by** 12 in 100,000 (30 in 100,000 DOWN TO 18 in 100,000)
3) **S12** INJURY S: **Current annual risk** 24 in 100,000 **Reduction by** 12 in 100,000 (24 in 100,000 DOWN TO 12 in 100,000)	6) **W12** INJURY W: **Current annual risk** 16 in 100,000 **Reduced by** 12 in 100,000 (16 in 100,000 DOWN TO 4 in 100,000)

APPENDIX C: SAMPLE WILLINGNESS TO PAY QUESTION

K1

Immediate unconsciousness, followed shortly by death.

CURRENT ANNUAL RISK 8 in 100,000
REDUCED BY 4 in 100,000
(8 in 100,000 DOWN TO 4 in 100,000)

Put a next to each amount you are sure you would pay
Put a × next to each amount you are sure you would not pay
Put a * next to the amount which you think is your best estimate of what you would be willing to pay for that annual risk reduction

0
£1
£2
£3
£5
£8
£10
£15
£20
£25
£30
£40
£50
£60
£75
£100
£150
£200
£250
£300
£400
£500
More than £500