Comparing the results of face-to-face and web-based PTO exercises.

Angela Robinson[1], Judith Covey[2], Mike Jones-Lee[3], Graham Loomes[1]

[1] **University of East Anglia**

[2] **University of Durham**

[3] **University of Newcastle**

Address for correspondence:

Angela Robinson

School of Medicine, Health Policy & Practice,

University of East Anglia,

Norwich

10603 593620

e mail: angela.robinson@uea.ac.uk

Introduction

For some years now the Rail Industry – along with the Department for Transport (DfT), the Health and Safety Executive (HSE) and various other UK public sector agencies – has been using *willingness to pay* (WTP)-based values of safety in the cost-benefit analysis (CBA) of proposed rail safety projects. In considering the level at which to set the Value of Preventing a Statistical Fatality (VPSF) for rail accidents, the Rail Safety & Standards Board (RSSB) started with the "baseline case" of a single-fatality rail accident involving an adult passenger behaving responsibly. Following careful consideration of the available empirical evidence, it was decided to set this VPSF at a level equal to the willingness-to-pay-based value used by the DfT in road project appraisal (currently about £1.4 million in 2004 prices). However, this still left open the question of the appropriate level at which the rail VPSF should be set for a number of other cases, involving, for example, the death of a responsible adult in a *multiple*-fatality rail accident; or a child who has wandered in error into the path of an oncoming train; or a track worker; or, indeed, an adult trespasser or suicide.

In view of this, in 2005 the RSSB commissioned a research project aimed at establishing how members of the public would prioritise the prevention of these other types of rail fatality relative to the baseline case. More specifically, the aim of the project was to estimate "multipliers" or "discount factors" which could be applied to the baseline VPSF and non-fatal injury values in order to derive corresponding values for the other cases, such as multiple fatality accidents, trespassers, suicides and so on. It was decided to base the estimation of the public's relative valuation of the prevention of different types of statistical rail fatality on so-called Person Trade Off (PTO) or as they are sometimes referred to, "matching" questions. Essentially these questions aim to establish the number of rail fatalities of a given type that would need to be prevented by a safety improvement in order for the respondent to regard that safety improvement as being "equally as socially desirable" as the prevention of a given number of baseline case fatalities.

The question that then arises is how such prioritization data might best be gathered from a representative sample of the public. Four broad avenues of approach are; 1) by conducting face-to-face interviews, 2) by assembling focus groups, 3) via postal questionnaires and 4) via the internet using computers/PCs. Given the relative complexity and difficulty of the type of questions that need to be put to respondents in a sample survey concerning preferences and

attitudes to health and safety, there is clearly an *a priori* case in favour of face-to-face interviews or focus groups given that both of these approaches allow respondents to seek direct clarification of questions that they may not fully understand. On the other hand, postal or internet surveys are for obvious reasons far less costly to carry out.

In view of all this, RSSB decided to commission *two separate and independent* sample surveys, the first – involving direct face-to-face computer assisted personal interviews (CAPI), and the second using the internet. The essential purpose of commissioning these studies was to examine the workability of a nationally representative large scale web-based study in the safety field and to provide a second independently generated data set.

PTO values have been elicited previously via the internet with some reported success (Baron and Ubel, 2002, Schwappach, 2002, Damschroder et al, 2007) but without an alternative elicitation method to act as a comparator. One US study has set out to compare PTO responses in a) a 'face-to-face' group, in which an interviewer used a web-based survey instrument in a structured interview and b) a 'computer' group, in which the subjects used the same web-based survey directly without an interviewer (Damschroder et al 2004). They found the quality of responses did not differ significantly between the two treatments. The study was, however, conducted as a randomised controlled trial with all subjects required to attend the University of Pennsylvania to take part in the experiment. The resulting sample was small (N=95) and unrepresentative of the general population, making it difficult to generalise to 'real world' situations. In contrast, we set out here to conduct a more natural field experiment in order to compare the feasibility of conducting large scale CAPI and internet studies in the UK using survey organisations' standard recruitment practices and procedures.

Clearly in order to render the whole comparative exercise worthwhile it was essential to ensure that the questions put to respondents in the two studies were, to all intents and purposes, identical. In the event, the CAPI survey (which produced a useable sample of 1033 respondents) was carried out in June 2006 and the internet survey (which yielded 1957 useable responses) took place in August 2006.

Methods

CAPI design

Four different versions of the CAPI were produced. Having introduced to the aim of the study, respondents were asked some questions about how far they lived from their nearest railway station and how often they travelled on both local and long distance trains. Respondents were then presented with a showcard which showed descriptions of 15 different types of single-fatality deaths that can occur on the railways (see appendix one). These included:

- **Rail passengers killed on trains (L,S)**
- **Rail passengers falling from station platforms (G,M)**
- **Trespassers killed on the railways (A,B,N,R).**
- **Suicide victims (E,Q)**
- **Drivers killed on level crossings (D,K)**
- And finally, **track workers (C,F)**

In versions 1 and 2 the descriptions were presented in the following order - L, S, H, R, A, N, B, M, G, Q, E, K, D, F, C. In versions 3 and 4 the order was reversed. Having read the descriptions respondents were asked to rate whether they thought that preventing each type of death was HIGH, MEDIUM or LOW PRIORITY. In giving their answer they were asked to set aside their views about how easy, difficult or costly it might be for the railways to prevent these types of deaths.

Respondents were then given a set of cards showing 7 of the 15 descriptions. The set of cards used varied between the 4 versions. However, as shown below, some cards were common across versions. L and R were included in all four versions, A and N were included in both versions 1 and 2, and F and C were included in both versions 3 and 4.

- Version 1: L, R, A, N, B, Q, E
- Version 2: L, R, A, N, D, K, M
- Version 3: L, R, F, C, S, G, K
- Version 4: L, R, F, C, M, G, H

When respondents had read through the cards they were asked to rank them according to the priority they thought should be given to their prevention. Equal ranks were allowed. The ranking task was followed by six PTO questions which were used to estimate the implied values of

preventing the different types of rail fatalities. The six matching questions for each version of the questionnaire are shown below.

Version 1	**Version 2**	**Version 3**	**Version 4**
L vs R	L vs R	L vs R	L vs R
R vs A	R vs N	R vs R	R vs G
A vs L	N vs L	F vs L	G vs L
A vs B	N vs M	F vs C	G vs M
L vs E	L vs D	L vs S	L vs H
E vs Q	D vs K	S vs K	H vs C

Before the first PTO question was presented respondents were asked to suppose that the railway industry has some extra money to spend on safety and wants to know how the public would prefer to see it spent (see figure 1).

Figure 1: *Example choice*

In this question the choice is between a safety programme which is expected to prevent deaths of the type described on card L, and another which is expected to prevent deaths of the type described on card R. The computer has started by entering values of 10 for both programmes. Would you prefer the money to be spent on preventing 10 deaths of type L OR on preventing 10 deaths of type R OR do you think both programmes are equally good?

In giving your answer please bear in mind:

- **Things can definitely be done to prevent these types of deaths**
- **The programmes will cost the same amount of money**
- **The numbers shown are accurate estimates of the deaths that could be prevented over the next 10 years or so**.

PROGRAMME L	**PROGRAMME R**
• Adult rail passenger killed on train • Collision accident caused by signal failure • No other people hurt or killed at the same time	• Child trespasser killed on railway line • Taking shortcut across the track - inadequate fencing or climbed through hole • No other people hurt or killed at the same time
Prevent 10 deaths of this type	**Prevent 10 deaths of this type**

RECORD RESPONSE – PREFER L, PREFER R, EQUALLY GOOD

If the respondent thought the programmes were equally good they were directed to the next question. However if they preferred L or R the numbers of deaths prevented by each programme were changed in accordance with the flowchart shown in Appendix B. For example, if they preferred L the next choice they were presented with a choice between preventing 5 type L fatalities compared to 10 type R fatalities.

Then, depending on the choice the respondent made they were presented with subsequent choices in which the numbers of deaths prevented were changed again, following the scheme shown on the flowchart in Appendix B. As shown on the flowchart if the respondent indicated that they would prefer to prevent one death of one type rather than 10 of the other they were directed to a follow-up question in which they were asked to given an open-ended response for the number of deaths which would make the programmes equally good. At this point respondents were given the option of a "don't know" or "no upper limit/ infinite" response.

When the respondents had answered all six PTO questions they were presented with questions designed to test whether the multiple fatalities were the same as killed in 10 separate incidents. Space does not permit a detailed discussion of these questions here. The questionnaire concluded with demographic questions including age, gender, household size, number of children under 16 in the household, socio-economic status, and income.

Internet design

The internet-based survey was designed to be as close as possible to the CAPI survey with the questions posed – and the order of the questions- identical to above. The main difference was that all information displayed previously on a series of 'showcards' was now transferred to the screen for respondents to read online. In summary, the key steps used in both surveys were as follows:

- Introduction to aim of the study
- Presentation of list of 15 different types of single-fatality deaths that can occur on the railways (see Appendix A).
- Rating of each description in terms of whether they thought that preventing each type of death was HIGH PRIORITY, MEDIUM PRIORITY or LOW PRIORITY.
- Presentation of list of 7 of the 15 descriptions (varied according to version).
- Ranking of the 7 descriptions according to the priority given to their prevention.

- Six PTO questions involving single-fatality incidents.
- Two PTO questions comparing multiple vs single-fatality incidents (not discussed here).

Aggregating results

Whilst it may seem natural to consider the arithmetic mean of the individual PTO ratios as an appropriate measure of central tendency, we demonstrate here why this is problematic. Suppose the following represent data from 5 respondents who were indifferent between programmes R and L when programme R prevented X deaths of type R and programme L prevented Y deaths of type L (with either R or L = 10).

	Programme R	Programe L
	X	Y
Resp 1	10	5
Resp 2	10	9
Resp 3	10	5
Resp 4	5	10
Resp 5	9	10

This pattern indicates that respondents 1 to 3 preferred programme L at the first PTO iteration - i.e. when both programmes prevented 10 deaths- and the number of deaths prevented by programme L was subsequently set below 10 until their point of indifference was reached. In contrast, respondents 4 and 5 preferred programme R at the first PTO iteration and the number of deaths prevented by programme R was subsequently set below 10. Clearly, respondents 3 and 4 have identical- but opposite- preferences. Yet, the ratio X/Y is 2:1 for respondent 3 and 0.5 for respondent 4 (and vice versa for Y/X) resulting in a very different influence on the arithmetic mean of individual ratios. As taking the ratio X/Y or Y/X is arbitrary, such asymmetry is undesirable.

It is important to note here that the alternative approach of setting X at 10 and allowing Y to vary – either above or below 10- was rejected on the grounds that would introduce bias into the individual PTO responses themselves. Were X to be fixed at 10 and the respondent considered programme L to be twice as good as R, Y would be set at 5 (i.e. 5 fewer than X). If on the other hand, the respondent considered R to be twice as good as L, they would need to set Y at 20 (10 more than X). There is, however, evidence to suggest that respondents are susceptible to anchoring effects and ratio bias phenomenon- in which subjects focus on the *absolute* numbers in the numerator and ignore the denominator (Pinto- Prades et al, 2006). Such effects would tend to 'suppress' the value

of Y/X when R was considered to be twice as good as L above (i.e. Y would be set < 20) introducing an asymmetry to the individual responses.

There are a number of different ways of aggregating data of the type collected here, the first of which is similar to a method we have argued in favour of previously (see Chilton et al, 2002).

1. The ratio of means (medians) method

In this method the more favoured programme attracts a value of 1 and the less favoured a value equal to the number of deaths prevented by the more favoured programme divided by the number of deaths prevented by the less favoured programme. For example, programme L was respondent one's 'more favoured' programme and he set Y at 5. Programme L then attracts a value of 1 and programme R a value equal to Y/X= 0.5. Following this principle for all 5 respondents yields the following values;

	Programme R	Programme L
Resp 1	0.5	1
Resp 2	0.9	1
Resp 3	0.5	1
Resp 4	1	0.5
Resp 5	1	0.9

Means (medians) of each column may then be taken and the ratio of the means (medians) calculated. In this case, the mean value for programme R is 0.78 whilst the mean value for programme L is 0.88. The implied weight of programme L over R using ratio of means is then 0.88/0.78 = 1.13. Likewise, the ratio of medians is 1/0.9= 1.11.

2 *Medians (geometric means) of individual ratios.*

Alternatively, the ratio X/Y may be computed for each individual in the sample, but we have already argued against using arithmetic means of such ratios.

	X/Y
Resp 1	2
Resp 2	1.11
Resp 3	2
Resp 4	0.5
Resp 5	0.9

Taking medians of the individual ratios, however, results in an implied weight of programme L over programme R of approximately 1.11 (n.b. this method will always give the same result as taking the

ratio of medians above). The geometric mean of the individual ratios is 1.15. Whilst we took the view here that other aggregation methods were more appropriate than geometric means, we present the data below for illustrative purposes.

As above, those respondents who still preferred programme L over programme R when L prevented only 1 death (compared to 10 deaths in programme R) were asked to identify how many deaths programme R would have to prevent (some number greater than 10) to be just as good as preventing the 1 death in programme L.

Results

The samples

As above, there were a total of 1033 and 1957 usable responses in the CAPI and internet samples respectively. The mean (median) ages were 45.4 (43.0) and 41.8 (39.0) for CAPI and internet samples respectively. Five hundred and four (48.8%) of the CAPI sample were male, compared with 907 (46.4%) of the internet sample. Tables 1a and 1b show the breakdown of the samples according to age and gender, the main difference appearing to be the lower proportion of respondents in the 65 years and over category in the internet sample compared to CAPI (7.06% versus 19.02%).

Table 1a: Age/gender breakdown, CAPI sample (N= 1025)*

	Age range						Total
	18 – 24	25 - 34	35 – 44	45 - 54	55 - 64	65 +	
Male	88	80	86	73	69	107	503
Female	78	102	105	78	71	88	522
Total	166 (16.20%)	182 (17.76%)	191 (18.63%)	151 (14.73%)	140 (13.66%)	195 (19.02%)	1025

* Eight respondents did not give their age

Table 1b: Age/gender breakdown, internet sample (1955)*

	Age range						Total
	18 -24	25 -34	35 -44	45 -54	55 -64	65+	
Male	115	224	183	147	169	69	907
Female	132	305	184	150	208	69	1048
Total	247 (12.63%)	529 (27.06%)	367 (18.77%)	297 (15.19%)	377 (19.28%)	138 (7.06%)	1955

*Two respondents did not give their age

Tables 2a and 2b details the total annual household income in the CAPI and internet samples respectively. Unfortunately, differences in the way the data were collected by the two survey organisations and the relatively large number of 'refuse to answer' or 'don't know' responses makes direct comparisons difficult. Of the 628 respondents giving details of their income in the CAPI sample, 135 (21.5%) earned £40k or over. Of the 1588 respondents giving detail of their income in the internet sample, 414 (26.07%) earned £39k or over.

Table 2a: Total annual household income, CAPI sample N=(1033)

		Frequency	Percent	Cumulative Percent
Valid	Up to £10,000	130	12.6	12.6
	£10,000 to £14,999	80	7.7	20.3
	£15,000 to £19,999	78	7.6	27.9
	£20,000 to £24,999	66	6.4	34.3
	£25,000 to £29,999	56	5.4	39.7
	£30,000 to £34,999	52	5.0	44.8
	£35,000 to £39,999	31	3.0	47.8
	£40,000 to £44,999	36	3.5	51.3
	£45,000 to £49,999	27	2.6	53.9
	£50,000 or more	72	7.0	60.9
	Refused	258	25.0	85.9
	Don't know	146	14.1	100.0
	Total	1032	99.9	
Missing	System	1	.1	
Total		1033	100.0	

Table 2b: Total annual household income, internet sample N=(1957)

		Frequency	Percent	Cumulative Percent
Valid	Up to £5,000	50	2.6	2.6
	£5,000 to £7,999	89	4.5	7.1
	£8,000 to £12,999	154	7.9	15.0
	£13,000 to £17,999	184	9.4	24.4
	£18,000 to £23,999	234	12.0	36.3
	£24,000 to £31,999	265	13.5	49.9
	£32,000 to £38,999	198	10.1	60.0
	£39,000 per year or more	414	21.2	81.1
	Don't know	94	4.8	85.9
	Refused	275	14.1	100.0
	Total	1957	100.0	

In terms of social class grade, 567 of the 1033 (54.9%) in the CAPI sample and 969 of the 1957 (49.6%) in the internet sample were in either class AB or C1. As above, respondents were also asked about how often they used both local and long distance trains. This data shows that 18.8% of the internet sample never use long-distance trains compared with 33.4% of the CAPI sample.

Ranking data

Recall that respondents were asked to rank a set of 7 of the 15 descriptions of fatalities. Table 3 shows the mean and median rank of the fatalities in terms of the priority given to their prevention.

Table 3: Ranking data from CAPI and internet samples (rank = 1 is most preferred)

	CAPI		Internet	
	mean rank	median rank	mean rank	median rank
Version 1				
L	1.78	1.00	1.59	1.00
R	2.38	2.00	2.52	2.00
A	3.29	3.00	3.92	4.00
N	3.95	4.00	4.25	4.00
B	4.60	5.00	5.50	6.00
Q	5.19	6.00	5.56	6.00
E	4.18	4.00	4.19	4.00
Version 2				
L	2.04	1.00	1.74	1.00
R	2.77	3.00	2.88	3.00
A	3.70	4.00	4.31	4.00
N	4.09	4.00	4.55	5.00
D	2.47	2.00	2.19	2.00
K	4.72	5.00	5.33	6.00
E	4.96	5.00	5.39	6.00
Version 3				
L	2.22	1.00	1.95	1.00
R	3.41	4.00	3.95	4.00
F	2.98	3.00	2.98	3.00
C	4.62	5.00	5.38	6.00
S	3.19	3.00	2.77	2.00
G	2.98	3.00	3.03	3.00
K	5.01	6.00	5.69	6.00
Version 4				
L	1.98	1.00	1.71	1.00
R	3.02	3.00	3.45	4.00
F	2.60	2.00	2.74	3.00
C	4.33	5.00	5.43	6.00
M	4.99	5.00	5.59	6.00
G	2.79	2.00	2.62	2.00
H	4.75	5.00	4.86	5.00

The rankings are clearly very similar in the CAPI and internet samples, indicating that both groups had broadly similar views on the priority that ought to be given to the prevention of the various fatalities. Further, it is clear that those where the victim was themselves responsible are deemed to be of lowest priority. For example, fatality type K- ranked lowest in version 3-is the adult driver who zig-zagged round the barriers whilst Q- ranked lowest in version 1- is the adult suicide who jumped from the platform.

PTO data

The ranking task was followed by six PTO questions which were used to estimate the implied values of preventing the different types of rail fatalities. Table 4 presents the results of the matching questions for each version and aggregation method. A value greater than one indicates that more weight is given to preventing the first fatality in each pairing than the second. For example, the values in the first row are all greater than 1 indicating that more weight is given to preventing fatality type L- adult passenger behaving responsibly- than R- child trespasser. In contrast, the values of less than one for the N/L comparison in version two indicate that less weight is given to preventing fatality type than N- adult trespasser- than L- adult passenger behaving responsibly.

Clearly the broad pattern is similar for the CAPI and internet samples and in no case does the direction of the priority 'switch' from one fatality type to the other between the two samples at the aggregate level. Clearly the weights based on geometric means of the individual responses are significantly greater than the other methods used to aggregate the PTO data.

Correspondence between the priorities inferred by the ranking and PTO data is good. For example, in version 1, the PTO data indicates: L would be ranked above R, A and E. A would be ranked above B, and E would be ranked above Q. Taking either mean or median rankings these predictions can be seen to hold for both mean and median rankings in both samples. Hence, at the aggregate level there is a strong convergent validity across methods. (although not shown here, we also analysed the 'priority' data -high, medium and low priority- and found good correspondence between those responses and the implied weights from PTO).

Table 4: Aggregate PTO data

	CAPI			Internet		
	Ratio of means	Medians	Geometric mean	Ratio of means	Medians	Geometric mean
Whole sample						
L/R	1.33	1.00	1.96	1.44	1.00	1.91
Version 1						
R/A	1.75	2.00	4.24	1.77	2.00	3.30
A/L	0.53	0.45	0.22	0.49	0.50	0.20
A/B	1.80	2.00	4.36	1.49	1.00	2.64
L/E	2.26	3.33	6.01	1.97	2.00	4.43
E/Q	1.74	2.00	3.82	1.57	1.82	2.79
Version 2						
R/N	1.76	2.00	4.24	1.44	1.00	2.36
N/L	0.43	0.30	0.15	0.40	0.30	0.14
N/M	1.50	1.43	2.77	1.25	1.00	1.60
L/D	1.05	1.00	1.67	1.02	1.00	1.07
D/K	2.66	6.67	8.59	3.22	10.00	10.98
Version3						
R/F	0.95	1.00	0.89	0.83	1.00	0.64
F/L	0.86	1.00	0.78	0.91	1.00	0.82
F/C	2.46	3.33	7.35	2.81	4.00	8.70
L/S	1.23	1.00	1.66	1.10	1.00	1.26
S/K	2.24	3.33	6.55	2.59	4.00	8.32
Version 4						
R/G	0.99	1.00	0.54	0.76	1.00	0.54
G/L	0.83	1.00	0.65	0.88	1.00	0.78
G/M	3.14	10.00	11.13	2.87	3.33	7.83
L/H	2.60	3.33	7.35	2.41	3.33	5.99
H/C	1.02	1.00	1.09	1.21	1.00	1.45

Individual responses

Whilst there appears to be a good deal of convergent validity in the aggregate data, we turn here to examine individual responses more closely. As space is limited, we restrict this to the L-R pairing on which we have data from the whole sample. Tables 5a and 5b show the correspondence between ranking and PTO responses at the level of the individual respondent. Observations lying on the top- left to bottom- right diagonal (shaded) may be thought of as 'strictly consistent' in that L was ranked above/equal/below R and the respondent went on to attach more/the same/less weight to L in the PTO exercise. Observations in the top right and bottom left cells may be thought of as 'strictly inconsistent' in that L was ranked above/below R but the respondent went on to attach

less/more weight to L in the PTO exercise. The remaining observations may be considered to be 'weakly consistent'. The figures in brackets represent the percentage of respondents in that cell.

Table 5a: Consistency of individual responses: CAPI sample

		Ranking			
		Prefer R	Equal	Prefer L	
PTO	Prefer R	89 (8.62)	20 (1.94)	87 (8.42)	196 (18.98)
	Equal	84 (8.13)	68 (6.58)	166 (16.07)	318 (30.78)
	Prefer L	76 (7.38)	44 (4.26)	399 (38.63)	519 (50.24)
		249 (24.10)	132 (12.78)	652 (63.12)	1033

- all percentages calculated out of 1033

Table 5b: Consistency of individual responses: internet sample

		Ranking			
		Prefer R	Equal	Prefer L	
PTO	Prefer R	50 (2.59)	8 (0.41)	47 (2.43)	105 (5.43)
	Equal	189 (9.78)	163 (8.43)	518 (26.80)	870 (45.01)
	Prefer L	80 (4.14)	37 (1.91)	841 (43.51)	958 (49.56)
		319 (16.50)	208 (10.76)	1406 (72.74)	1933

- all percentages calculated out of 1933- 24 respondents did not respond to the L vs R PTO question.

A total of 556 (53.8%) and 1054 (54.5%) of responses to the L vs R comparison were 'strictly consistent' in the CAPI and internet samples respectively. A total of 163 (15.78%) and 127 (6.57%) of responses to that same question were 'strictly inconsistent' in the CAPI and internet samples respectively. Whilst it may appear at first glance that the internet sample is doing rather better in having fewer strictly inconsistent responses, this pattern can be explained by the associated larger numbers of 'weakly consistent' responses. In particular, internet respondents were more likely that their CAPI counterparts to rank L above R, but then go onto rate them equally in the PTO (26.8% versus 16.07%).

Although the data for the remaining 20 PTO questions is not shown here, we can report that 43% of all PTO responses in the internet sample and 30% of all PTO responses in the CAPI sample were- 'equals' at 10/10 i.e. respondents did not differentiate between the two programmes when each prevented 10 fatalities.

Discussion

To the best of our knowledge, no such parallel CAPI and internet PTO studies on this scale have been carried out, although a smaller more 'controlled' experiment has taken place in the US (see Damschroder et al 2004). As with the Damschroder study, our results are broadly encouraging for the feasibility of using an internet–based approach to elicit responses to the type of questions posed here. Results were broadly similar to those of the more intensive (and expensive) face-to-face study and at first glance the 'quality' of the responses did not appear to vary markedly between the two. We believe, however, there are a number of important caveats to be made before the use of internet surveys of this nature can be recommended.

First, although the tasks respondents were asked to do here were ones with which they would not be familiar, the fatality descriptions used were fairly 'stark' and not particularly wordy or detailed. As such, respondents would be able to absorb the gist of the information fairly quickly and differences between the scenarios would be relatively easy to spot. In contrast, many studies including those conducted in health economics, require respondents to consider a set of more detailed descriptions with the differences between each more subtle than was the case here. It seems plausible that internet-based studies may be more problematic when the respondent is required to digest a lot of detailed information.

Second, it was not possible here to check for the *internal* consistency of PTO responses as no fatality-type could be considered to 'dominate' another in the way that health states often do (in that they are better on at least one dimension and just as good on others). Hence, we have no means of identifying from the PTO responses alone whether certain respondents got the task 'wrong' and ended up giving more weight to their least preferred programme.

Third, it is worrying that there were a greater number of respondents giving 'equivalence' responses in the internet compared to CAPI samples (43% of the internet sample and 30% of CAPI). In contrast, Damschroder et al found that the number of equivalence responses -or respondents 'refusing to trade' as they refer to them- was roughly the same between the interview and internet only groups (approx 21% in each). That those respondents motivated sufficiently to travel to campus to take part in an experiment behave in a broadly similar manner whether or not there is an interviewer present is perhaps not surprising. It has to be acknowledged, however, that clicking on 'equivalence' responses is the quickest way through the PTO questions and it is a concern that this played some role in the internet responses in particular here. It seems plausible that respondents clicking on a survey in their own homes- or places of work- may be more inclined to take the shortest route through than those participating in a structured interview. We believe this is something which requires further investigation.

Whilst the policy implications of our work is not the focus of this paper, it is interesting to note that the apparent importance of the 'blameworthiness' factor here is in direct contrast to the National Institute of Health and Clinical Excellence (NICE) view that priority setting within health care should not take account of 'deservedness'. Principle 10 in NICE's document on social values states that discrimination against patients with conditions that are, or may be, self-inflicted should be avoided based on the findings of the citizen's council who, we are told, 'rejected the notion of 'deservedness' in priority setting within the NHS'(NICE, 2005). It is worth noting that the conclusions reached by NICE's citizen's council would appear to conflict with results of other studies which have solicited the views of the general public about the issue of responsibility and blame (see, for example, Ratcliffe 2000). It may, therefore, be premature to conclude that the population that the NHS serves do not wish personal responsibility to be taken into consideration in the prioritisation of resources.

On the other hand, some of the apparent desire to 'punish' those in some way responsible for their own demise may be an artefact of the prioritisation exercises used. For example, whilst respondents in this study were asked to accept that the various fatalities *would* be prevented and that all programmes cost the *same* amount of money (see figure 1 above), it seems plausible that certain types of fatalities were considered more difficult and costly to prevent than others. Thus, the lower weight attached to the prevention of, for example, suicides may in part reflect that respondents felt

nothing much could be done to prevent those deaths. Hence, resources directed towards suicide prevention would essentially be wasted and 'efficiency' considerations may be overtaking considerations of 'deservedness'. Similarly, respondents stating that health care resources (such as transplant organs) ought not to be directed towards smokers and alcoholics may also be incorporating efficiency arguments into their thinking (as they perceive the likely health gain to be lower). It will be important that any future research into priority setting within the NHS disentangles efficiency considerations from other criteria.

Another important methodological issue we raise here is how PTO responses ought to be aggregated. This is an area that we believe has been under-researched to date and previous studies have aggregated PTO data in a number of different ways. For example, in one study that set out to test the significance of age and severity of illness in social values for health care, Nord et al (1996) fixed the number of 20-year olds to be treated at 10 and asked respondents to 'set' the number (X) of 10 year olds that would have to be treated to make the two programmes equivalent in social value. The *median* equivalence number of 10 year olds – deemed to be just as good as treating ten 20 year olds- was reported at 9.5 (see table 2 on page 106) and the relative value of treating 10 year olds compared to 20 year olds taken to be 10/9.5. Alternatively, respondents in the Damschroder et al study (2004) were asked for their equivalence value between curing 100 quadriplegics or X patients with foot numbness. Individual ratios of X/100 were computed and log transformed because they were 'heavily skewed' (i.e. geometric means were taken). Neither study reported results using alternative aggregation methods.

It is not, however, our intention here to criticise the aggregation techniques used elsewhere. Rather, we merely flag up that the 'aggregation issue' exists and highlight the problem of comparing the results of studies that have used different aggregation methods. This is important as a number of tasks involved in value elicitation exercises require the aggregation of ratio data. In fact, the 'aggregation issue' also came up in our recent work on the Social Value of a QALY (SVQ) study (Baker et al, 2008,) not only in relation to estimated weights from PTO responses (but see Robinson, 2007) but also to the estimated WTP for a QALY using the SG/CV approach (see Loomes 2008). We believe this is a general area that requires further attention and look forward to hearing the views of HESG/NHESG members on this topic (or any other!).

References

Baker R, Bateman I, Donaldson C, Jones-Lee M, Lancsar E, Loomes G, Mason H, Odejar M, Pinto Prades JL, Robinson A, Ryan M, Shackley P, Smith R, Sugden R, Wildman J, Weighting and valuing quality adjusted life years: preliminary results from the Social Value of a QALY Project, 2008.

Baron J, Ubel PA, Types of inconsistency in health-state utility judgments. *Organisational behaviour and human decision process* 2002 89: 1100-18

Chilton, S**.,** Covey, J., Hopkins, L., Jones-Lee, M.W., Loomes, G., Pidgeon. N. and Spencer, A. (2002). Public Perceptions of Risk and Preference-Based Values of Safety, *Journal of Risk and Uncertainty*, 25:3, 211-232

Damschroder LJ, Roberts TR, Zikmund-Fisher BJ, Ubel PA, Why people refuse to make tradeoffs in person trade off elicitations: a matter of perspective? Medical Decision making, 2007, 27: 266-280

Damschroder LJ, Baron J, Hershey JC, Asch DA, Jepson, C Ubel PA, The validity of person tradeoff measuremenmts: drandomized trial of computer elicitation versus face –to-face interview. Medical Decision Making 2004, 24:170-180.

Loomes G (on behalf of the SVQ team) The feasibility- or otherwise- of estimating the monetary value of a QALY, paper presented at HESG, Norwich Jan 2008

National Institute of Health and Clinical Excellence, Social value judgements. Principles for the development of NICE guidance, December 8th 2005.

Nord, E, Street A, Richardson, J, Huhse, H, Singer, P 'The significance of age and duration effect in social evaluation of health care', *Health Care Analysis*, 1996; 4: 103-111.

Ratcliffe J. Public preferences for the allocation of donor liver grafts for transplantation, Health Economics, 2000;137-148.

Robinson A (on behalf of the SVQ team) Do members of the public attach more weight to some QALYs than others?, conference presentation iHEA, Copenhagen 2007.

Schwappach DLB. The equivalence of numbers: the social value of avoiding health decline: an experimental web-based survey, BMC medical inform Decision making 2002; 2.3

Appendix A: Descriptions of the 15 railway fatalities.

L • Adult rail passenger killed on train • Collision accident caused by signal failure • No other people hurt or killed at the same time	**G** • Adult rail passenger killed falling from station platform • Passenger tripped on uneven/ unrepaired platform No other people hurt or killed at the same time
S • Adult rail passenger killed on train • Derailment caused by vandalism to the track • No other people hurt or killed at the same time	**Q** • Adult commits suicide jumping from station platform • No other people hurt or killed at the same time
H • Adult rail passenger killed on train • Struck whilst leaning out of window while train moving • No other people hurt or killed at the same time	**E** • Adult commits suicide on railway line • Railway line close to psychiatric institution - fencing inadequate • No other people hurt or killed at the same time
R • Child trespasser killed on railway line • Taking shortcut across the track - inadequate fencing or climbed through hole • No other people hurt or killed at the same time	**K** • Adult driver killed in car struck by train on level crossing • Zigzagged around an automatic half barrier • No other people hurt or killed at the same time
A • Child trespasser killed on railway line • Involved in act of vandalism – climbed over or made hole in fencing • No other people hurt or killed at the same time	**D** • Adult driver killed in car struck by train when on level crossing • Signals or barriers failed • No other people hurt or killed at the same time
N • Adult trespasser killed on railway line • Taking shortcut across the track - inadequate fencing or climbed through hole • No other people hurt or killed at the same time	**F** • Track worker killed on railway line • Had not received enough training or similar failure by rail system • No other people hurt or killed at the same time
B • Adult trespasser killed on railway line • Involved in act of vandalism – climbed over or made hole in fencing • No other people hurt or killed at the same time	**C** • Track worker killed on railway line • Had received enough training but knowingly broke safety procedures • No other people hurt or killed at the same time
M • Adult rail passenger killed falling from station platform • Under influence of alcohol • No other people hurt or killed at the same time	

Appendix B: The flowchart through the PTO procedure (red indicates prefer L, blue prefer R)