
Published paper

Fowkes, A.S. (1992) *How Reliable is Stated Preference?* Institute of Transport Studies, University of Leeds. Working Paper 377

UNIVERSITY OF LEEDS
Institute for Transport Studies

ITS Working Paper 377 *ISSN 0142-8942*

September 1992

HOW RELIABLE IS STATED PREFERENCE?

A.S. Fowkes

ABSTRACT

FOWKES, A.S. (1992). How reliable is Stated Preference? *ITS Working Paper 377*, Institute for Transport Studies, University of Leeds, Leeds.

Agencies involved with transport operations have a need for traveller valuations of the various attributes of the transport alternatives. Observation of revealed market place behaviour in order to make Revealed Preference estimates of these valuations have great limitations. As an alternative, analysis of travellers' Stated Preferences has gained in popularity, albeit with worries as to its reliability. This paper seeks to make the case that Stated Preference surveys can be reliable if carried out in a reasonably competent way. Validation against Revealed Preference methods is discussed, followed by an explanation of why Stated Preference can be better than Revealed Preference. The importance of testing designs is stressed. A method which provides consistency checks and independently derived estimates of attribute valuations is presented. Orthogonality of designs is discussed. The Scale Factor problem and Taste Variation are discussed, with advice as to how to overcome any difficulties.

KEY-WORDS:

Contact: Tony Fowkes, Institute for Transport Studies (Tel: 0532 335340).

HOW RELIABLE IS STATED PREFERENCE?

1. INTRODUCTION

It is perfectly natural and understandable to put greater faith in data or actual decision making rather than data on what respondents say they would do faced with various hypothetical choices. The methodology for the analysis of actual data on discrete choices is well developed, and is generally known as Revealed Preference analysis. By observing choices actually made between alternatives having known characteristics we can impute relative valuations, such as that between time and money, - the so-called value of time. Unfortunately, the average information content for each such choice is meagre, so that even where estimates of relative valuations can be derived at all, they have wide confidence intervals around them. Attempts to increase the sample size quickly lead to the inclusion of respondents facing rather different choice contexts, which muddies the water to an extent that the additional sampling cost yields little in the way of improved estimates.

The alternative of using responses from hypothetical choice scenarios has had a chequered history, and rightly attracts much suspicion. Fortunately, it was decided to try out such methods against Revealed Preference, and a hybrid approach called Transfer Price, as part of the UK Department of Transport Value of Time Project (MVA, ITS, TSU, 1987). To the surprise of many, including myself, a version of these hypothetical techniques, by now known as Stated Preference, clearly outperformed the others. The Department therefore accepted the use of Stated Preference for the main surveys, and accepted the results therefrom. DOT acceptance has been very influential, such that British Rail and local councils now regularly use Stated Preference results when preparing investment cases which either have to go to DOT for approval, or which seek a DOT grant, eg. 'Section 56'. This paper attempts to review the ways in which the reliability of Stated Preference results has been increased, relative to other methods.

2. RP v SP VALIDATION

One major purpose of the DOT Value of Time study's survey of North Kent Commuters was to validate the SP technique against the RP technique previously used. The general finding was that the error bands on the RP results were sufficiently wide to include the SP results and a lot else too. In other words, the fact that the overall VOT estimates by RP and SP were not significantly different was not particularly convincing evidence given that only rather unexpected SP VOTs would have been found to be significantly different. Similar conclusions from RP/SP comparisons were drawn for Tyne Crossing data by Wardman (1991) and for Dutch data by Bradley and Gunn (1990).

Much more powerful evidence, though, has been provided by Wardman's analysis of subgroups in the North Kent sample. The results are shown in Table 1. While there are no (even near) cases of statistically significant difference between the RP and SP results, it can clearly be seen that the SP and RP results vary in the same direction. The North Kent SP experiment is now regarded as a relatively poor one compared to current practice, yet its results are far from being random.

Table 1
RP v SP Comparison: North Kent Commuters Values of Main-Mode In-Vehicle Time (p/min)

		RP	SP
SEX	MALE	3.43	3.13
	FEMALE	2.87	2.70
AGE	16-24	2.48	2.61
	25-44	3.29	3.02
	45+	4.23	3.18
INCOME	-£7000	2.45	2.53
	£7000-11000	3.72	3.06
	£11000+	4.18	3.20

Source: Wardman, JTEP (1988)

Note: The largest statistic for a test of difference between RP and SP estimates in any of these 8 cases is 1.01.

Wardman continued his analysis by comparing the predicted choices from the SP model with both the actual choices and the predicted choices from the RP model. Individuals' choices were predicted by entering into a generalised cost formulation the times and costs faced in practice. Each individual was assigned to the mode having least generalised cost. Table 2 shows the outcome of the comparisons of SP and RP choices with actual choices. From the SP model 77% of choices are correctly predicted, compared to 71% from the RP model, and 53% by randomly allocating individuals to modes according to the proportions using each mode.

Table 2
Comparison of Actual and Predicted Modes: North Kent Survey

	SP	RP
Train correctly predicted	387	411
Coach correctly predicted	134	69
Train incorrectly predicted	126	191
Coach correctly predicted	32	8
% correctly predicted	77	71

Source: Wardman, JTEP (1988)

3. BOUNDARY VALUES

Usually, we have only one RP observation per individual. We need to interview/observe a group of individuals all facing a similar choice, eg. mode choice for the journey to work. More than this we need data which is informative. By that I mean that the individuals in our sample should, taken together, have faced a range of 'boundary values' for their choices.

In order to explain the concept simply, let us restrict ourselves to just two attributes X and Y (taking values X_i, Y_i respectively), with utility, U, computed as a weighted sum:

$$U = aX + bY$$

Then, if we compare two transport alternatives i=1 and i=2, we have

$$U_1 - U_2 = a\,(X_1 - X_2) + b_1\,(Y_1 - Y_2)$$

At the point of indifference between the alternatives $U_1 = U_2$ and so

$$\frac{b}{a} = \frac{X_1 - X_2}{Y_2 - Y_1}$$

where b/a is the boundary value of Y expressed in terms of X. In the absence of random effects, an individual whose value of Y in terms of X is greater than b/a will prefer the alternative with greatest Y, and vice versa. We will denote boundary values of Y expressed in terms of X as B(Y:X).

To look at a concrete example, suppose the two alternatives are journeys by train and by coach, and that the only considered attributes of these alternatives are cost and time. This last statement implies that there is no 'Alternative Specific Constant' in favour of either train or coach - they are judged solely on their cost and journey time.

U = Cost + (VOT) Time (in pence, say)

where VOT = Value of Time (in pence/min, say)

$$B(\text{Time:Cost}) = \frac{\text{Train Cost - Coach Cost}}{\text{Coach Time - Train Time}}$$

Because it will be commonly used, we will rename B(Time:Cost) as BVOT.

As a numerical example, suppose that the following are the rail and coach times and costs for Leeds-London (round trip):

	Train Cost	=	£40
	Coach Cost	=	£20
	Train Time	=	260 mins
	Coach Time	=	460 mins
BVOT = B(Time:Cost) =	$\frac{4000-2000}{460-260}$	=	10p/min

Hence individuals with values of time greater than 10p/min will choose the quicker/dearer alternative (train), while those with lower values of time will choose the slower/cheaper alternative (coach).

All else equal, the design should present boundary values closely either side of the true relative valuation. In saying this, it should be noted that boundary values in a design will vary with any third and subsequent attributes. For example, suppose that in the Leeds-London example we were to assume that there was some fixed benefit, A, of travelling by train as compared to travelling by coach, regardless of how long the journey takes. Let us suppose that the population can be divided into two groups having preferences for train over coach, all else equal, of A_1 and A_2 respectively. The boundary value expression becomes:

$$\text{BVOT} = \text{B(Time:Cost)} = \frac{\text{Train Cost - Coach Cost - A}}{\text{Coach Time - Train Time}}$$

which in the Leeds-London example, with the two sub-groups of the population described above gives:

$$\text{BVOT (group 1)} = \frac{2000 - A_1}{200} \text{ p/min}$$

$$\text{BVOT (group 2)} = \frac{2000 - A_2}{200} \text{ p/min}$$

This is illustrated in the top line of Table 3 for the cases where ASC = 0, 2 or 5. It can therefore be seen that the previously found BVOT, in the absence of an ASC, of 10p/min is shown in the ASC = 0 column. If a traveller is willing to pay £2 extra to travel by train rather than coach, for times and costs equal, then we say the ASC in favour of train is £2. The corresponding BVOT for the situation considered earlier is now 9p/min (see Table 3, line A, under ASC = 2). Similarly, if the ASC were £5 the BVOT would be 7.5p/min. The relationship of BVOT with ASC for this situation is displayed graphically in Figure 1 as line A. Note that ASC can be negative, in which case coach is favoured to train, for times and costs equal.

In Stated Preference experiments we ask for more than one choice, and we should ensure that, whatever the value of 'third' variables (such as ASC here), respondents should be faced with a suitable range of BVOTs. If the situation described previously is taken to be represented on card (or 'scenario') A, then cards B to F in Table 3 show how suitable ranges of BVOTs can be achieved. We can judge their adequacy either by reference to some values of ASC which are deemed to cover a suitable range of what is likely to be encountered in the sample, for which we have taken ASC's of 0, 2 and 5; or graphically as in Fig.1.

Figure 1: Boundary Value Map

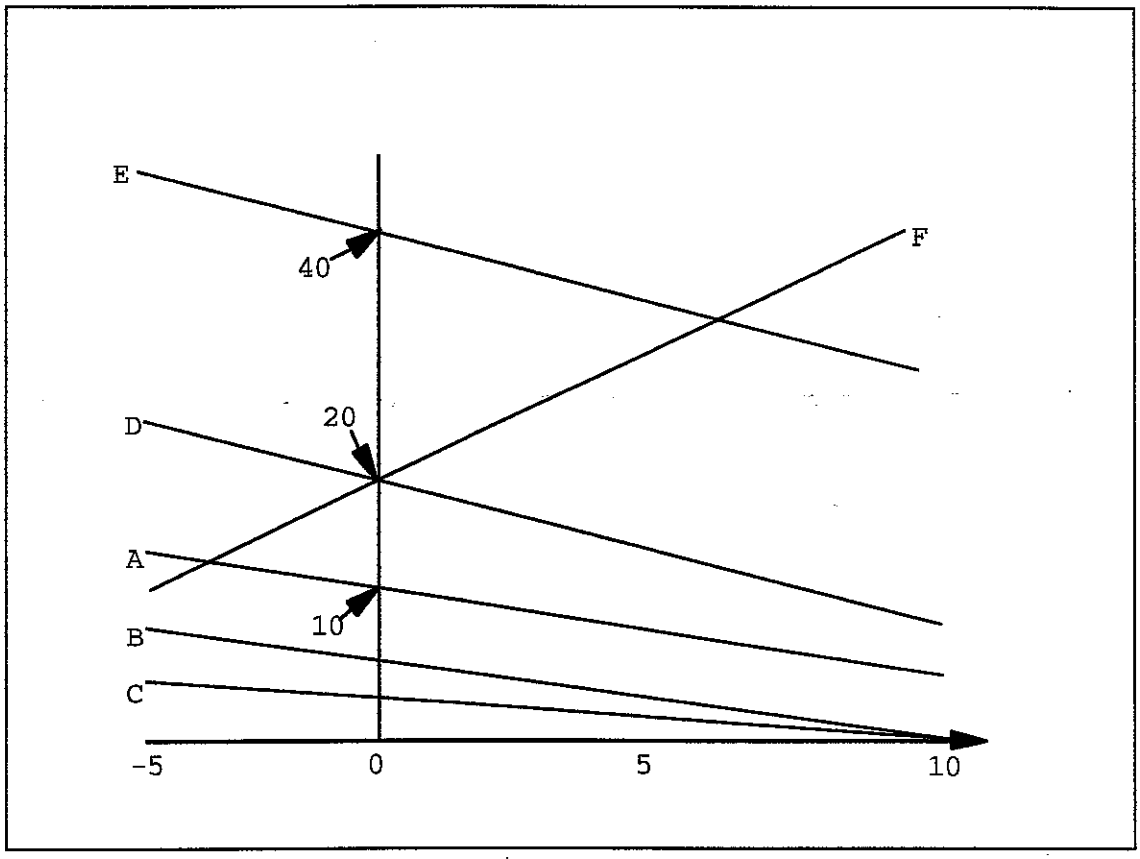

Table 3: Effect of Alternative Specific Constant on Boundary Values of Time (BVOT)

Card (Scenario)	Train Cost	Coach Cost	Train Time	Coach Time	BVOT ASC=0	BVOT ASC=2	BVOT ASC=5
A	40	20	260	460	10.00	9.00	7.50
B	60	50	260	460	5.00	4.00	2.50
C	60	50	210	610	2.50	2.00	1.25
D	40	20	210	310	20.00	18.00	15.00
E	60	20	360	460	40.00	38.00	35.00
F	40	50	360	310	20.00	24.00	30.00

The values shown for cards B to F are not meant to be optimal, merely first attempts at obtaining a suitably good coverage of boundary values. I have endeavoured to choose attribute levels which illustrate the principles of design, whilst restricting myself to just two cost levels and two time levels for each mode. It is generally held that use of few (preferably nice round) attribute levels helps respondents, particularly in ranking exercises. My own view is that such considerations should be clearly secondary to the attainment of adequate boundary values. My forbearance in Table 3 is merely to demonstrate that good boundary values do not always imply unaesthetic attribute levels.

Cards B and C attempt to suitably subdivide the area below the BVOT line for card A, which I shall call BVOT(A). Where a valuation can be taken to be positive (eg. the value of time), it seems sensible to work in ratios rather than absolute differences. Hence card B keeps the same times as card A but halves the cost difference. In the absence of an ASC this has the effect of

halving the BVOT (from 10.00 to 5.00). Since the ASC is taken as a subtraction from the Train minus Coach cost difference, other BVOTs do not all fall by 50%, but instead fall as ASC increases, with the same slope as BVOT(A). Algebraically, BVOT(B) = BVOT(A) - 5.

Card C further halves the BVOT for zero ASC (now to 2.50), but does it by retaining the same cost difference as in card B, but doubling the time difference. Consequently, this time all BVOTs are halved, ie.:

$$BVOT(C) = 0.5 * BVOT(B)$$

Cards D and E do much the same, but in the opposite direction, ie. for zero ASC they successively double BVOT. Card D retains the same cost difference as card A, but halves the time difference, so that all BVOTs double. Card E retains the same time difference as card D (albeit with different levels) but doubles the cost difference. This gives BVOT(E) values for our range of ASCs between 35 and 40 p/min. While this seems perfectly high enough for most current UK applications, Figure 1 does show that BVOT(E) falls away as ASC increases. Since high ASC's are likely to occur with high VOT's, reflecting the influence of income, this may not be regarded as adequate. Card F has therefore been included, which has an upward sloping BVOT line. The trick is to switch the ordering of both costs and times between the modes. This preserves the trade-offs, but can look odd to clients. If they become worried you can tell them that it is a device for checking consistency of responses, and indeed it may be effective in that role. Respondents very rarely get worried by such things, taking it all in their stride. The effect of card F can be judged from Figure 1.

In general there will be 'third' variables which will cause uncertainty for us when 'designing in' desired boundary values, either because the population's valuation of the 'third' variables is unknown or because different subgroups of the population are in any event expected to have different valuations. To the extent that the inclusion in the design of a suitably chosen range of boundary values has been understood and accepted by practitioners, it is presumably this difficulty which has so impeded its use thus far. The problem is at its greatest when utilising orthogonal fractional factorial designs for SP experiments, and most practitioners have felt safer sticking to these. The present author has been happy to use non-orthogonal designs, as discussed below, and has therefore been more free to choose appropriate boundary values, as we shall now see.

4. BIN ANALYSIS

When designing Stated Preference surveys it is often desirable to 'design in', for a given choice, what may be called 'fixed boundary values', by which I mean that all but two attributes have levels which are equal for this choice. This has the immediate advantage that the boundary value between the two attributes with unequal levels is known to the designer, and is not a function of the valuation of some third variable. By specifying two fixed boundary values in the design, respondents can be split into 3 'bins', ie. those below the lower boundary value, those between the two boundary values, and those above the higher boundary value. In addition, there is the potential for irrational response (below the lower boundary value but above the upper boundary value) which is useful in spotting respondents who have either misunderstood the questionnaire or not taken it seriously.

In general, if we specify n fixed boundary values we can place respondents into one of n+1 bins, or describe them as irrational. By plotting the frequency distribution as a histogram we can see the shape of the relative valuations implied by the responses to these choices (which, of course, may

only be a subset of the SP experiment). This need not correspond to the bell-shaped distribution assumed by conventional logit modelling!

In practice there usually is a bell shape - but possibly more than one, which indicates the presence of taste variation and suggests attempts at segmentation. Commonly we find a bell shape with a lump in the top bin, which we usually interpret as a mixture of respondents who feel they must have the attribute being valued (at whatever cost), those who have misunderstood the SP exercise, and those who are trying to bias the exercise. In all three cases the advice is to run logit models without these people. The justification in respect of the first group is that, although they may value the attribute very highly, in practice they will not be able to afford to pay the indicated amount for it and so, in effect, will not be choosing either of the alternatives offered - neither is acceptable.

Besides enabling the modeller to check on the quality of responses, Bin Analysis has another major advantage in that it provides some (admittedly rough) results that can be readily understood by the client and are not dependent on the 'black box' of logit modelling. As an illustrative example, suppose that in Table 3 we know that ASC = 0 (as would be the case, for example, if the alternatives were not 'TRAIN' and 'COACH' but two possible services for a given mode, say train). Suppose the proportions choosing the option labelled 'TRAIN' were as follows for each card:

A	B	C	D	E	F
0.30	0.55	0.80	0.14	0.05	0.16

we could then deduce the following distribution of values of time:

VOT (p/min)	
Below 2.5	20%
Between 2.5 and 5.0	25%
Between 5.0 and 10.0	25%
Between 10.0 and 20.0	15%
Between 20.0 and 40.0	10%
Over 40.0	5%

The above figures are determined as follows. The 80% of respondents favouring the faster/dearer option on card C mean that only 20% have a VOT below BVOT(C), ie. below 2.5p/min. Similarly the 55% choosing the faster option on card B means that 45% have a VOT below BVOT(B), namely 5p/min, of which we know that 20% are below 2.5p/min, leaving 25% between 2.5p/min and 5p/min. The rest of the distribution is determined similarly, except that the two cards with BVOTs of 20p/min (cards D and F) have had their percentages averaged to give 15% having a VOT above 20p/min.

5. BOUNDARY VALUES FOR RP ANALYSIS

We cannot always assume that each individual will be facing a trade-off. For instance, although train is generally a faster mode than coach, it may be that the coach stop is much more convenient for some individuals than is the rail station. In that way COACH TIME could be (say 5 minutes) below TRAIN TIME and BVOT is negative (–6p/min). Negative boundary values indicate that one alternative dominates the other in the attributes being considered - here coach is both quicker and faster. Learning that an individuals' value of time is greater than –6p/min is unlikely to be

much help to us.

We therefore seek experimental situations where choices are not dominated. This frequently rules out car versus train RP mode choice studies, since in many situations respondents will regard the car as quicker and cheaper. This is not only because marginal motoring costs are very low relative to average motoring costs, but also because motorists may genuinely not consider variable maintenance costs when making mode choices. A third aspect of the problem is that when responding to surveys motorists usually only include easily quantified costs such as petrol costs, tolls and parking charges.

If we manage to avoid dominated choices, we still require that the choices exhibit a good range of (positive) boundary values. This can be a problem where, to continue our example above, cost and time differences between alternatives are positively correlated. This can easily occur in practice as, for each mode, longer journeys are more costly and take more time. If fares were charged at a simple rate per mile and speeds were constant for all journey lengths for a given mode, we would have:

TRAIN COST	=	MILES * TRAIN FARE PER MILE
COACH COST	=	MILES * COACH FARE PER MILE
TRAIN TIME	=	MILES/TRAIN SPEED
COACH TIME	=	MILES/COACH SPEED

$$BVOT = \frac{\text{MILES*}\ (\text{TRAIN FARE PER MILE} - \text{COACH FARE PER MILE})}{\text{MILES*}\left[\left(\frac{1}{\text{COACH SPEED}}\right) - \left(\frac{1}{\text{TRAIN SPEED}}\right)\right]}$$

The MILES term cancel, and all the other terms are assumed constant so there is just one BVOT for all journey lengths. Our data will tell us merely what percentage of our respondents' values of time lie on each side of the one and only BVOT.

A particularly successful mode choice RP design was carried out in North Kent in 1983 as part of the DOT Value of Time Study. Rail and coach commuters into central London were interviewed. By design and some good fortune, the time and cost differences came out negatively correlated, largely due to the range of access distances to the suburban rail stations, and to the coach stops. Consequently, there was a great variation in the degree to which the door to door rail time was faster by rail than by coach. Hence there was a good range of BVOTs in the data. The possible correlation effect referred to above was avoided by sampling commuters making journeys all of similar distances, such that the cost difference between rail and coach was similar for all respondents. Nevertheless, the 95% confidence interval on the overall value of time estimates was $\pm$ one third, indicating that even a good RP experiment with about 1000 responses still yields rather imprecise estimates. Naturally, values of time disaggregated by sex or by income will be even more imprecisely estimated.

6. ORTHOGONALITY

If we have a product with two attributes of interest, in addition to its price, then it is clearly undesirable that their levels should be highly correlated in the experimental design. For example, if we wished to find monetary valuations for the availability of a telephone on a train and the availability of a buffet on a train, it would clearly be nonsense to present options where every train that had a telephone had a buffet and vice versa. We would then have perfect collinearity and we could only obtain a valuation for the joint presence of telephone and buffet, and could not deduce

the separate effects. As is well known, high levels of collinearity will reduce the accuracy with which the separate effects can be determined. This had led, quite sensibly, to the use of orthogonal experimental designs, by which we mean that the attribute levels are chosen such that there is zero correlation between the attributes.

The point at which it is here argued that this goes too far is when the cost attribute is also made orthogonal, particularly when what are required are monetary valuations. This is because these monetary valuations are derived as the ratio of the various attribute coefficient estimates to the coefficient estimate on the cost variable. Statistically, some particular form of correlation can help improve the accuracy of the monetary valuations. This is because the formula for the variance of a ratio includes, in its numerator, a covariance term. Thus if higher costs are associated with 'better' levels of one of the attributes (as is rather sensible) then the precision of estimate of the monetary valuation will be improved. This can be visualised by noting that overestimates of the coefficient of cost will tend to be associated with overestimates of the coefficient of the other attribute, so that the ratio is largely preserved.

The present advice is, then, for determining monetary valuations we should ensure orthogonality (or at least low correlation) amongst the non-cost attributes, but should deliberately choose cost levels to give good boundary values. It should be noted that it will usually be necessary to 'favour' one attribute at the expense of others. Additionally, however, even if the actual coefficients are required (eg for forecasting or to estimate elasticities) rather than relative valuations, experience suggests that non-orthogonal designs concentrating on achieving good boundary values can perform very well.

7. IMPORTANCE OF TESTING DESIGNS

Using a Stated Preference design where the highest boundary value of time is 3p/min is clearly not sensible for business travellers whose values of time will almost certainly be over 10p/min. All respondents will choose the more expensive and quicker option. This can be seen from an analysis of the boundary values, but most cases will be much less clear out. The only way to tell whether a design is likely to be adequate is to simulate responses and try to recover assumed parameter values. This is not quite so easy and straightforward as it may sound, and it can be an expensive exercise.

Our advice is to draw up a set of tests involving wide enough ranges of parameter values to cover those likely to prevail for any subgroup to be sample (unless these subgroups can be separated out with a screening questionnaire and subsequently given an SP design customised for that group). Furthermore, we strongly recommend that the assumed parameter values for the various attributes should not be kept in fixed relation (ie perfectly correlated). While it may be sensible for high values of one attribute to tend to be associated with high values of another attribute (due, say, to the influence of income or wealth) we should try some low-high combinations to check that the design can cope with these. This is particularly important if we are using a non-orthogonal design, as may be appropriate when we wish to devote most estimation effort to one particular relative valuation.

By checking that our designs can recover (a suitable range of) assumed parameter values in various combinations, we allow ourselves much more freedom in design and can go forward to the survey with much greater confidence than otherwise. In particular, we can justify a departure from conventional orthogonal designs.
Table 4 shows an example of an experimental design used in 1987 to study the valuation of

overcrowding on ECML trains. A secondary objective was to measure the valuation of having to adjust your departure time from what is ideal. Adjustments of 1 hour or 2 hours were considered, half the sample being asked about having to travel earlier and the other half about having to travel later.

'Crowding' had 4 'levels' in the design: PLENTY OF SEATS, TRAIN FULL BUT YOU GET A SEAT, STAND 30 MINS and STAND 60 MINS. Choices 1, 2, and 3 offer the choice between a FULL train and PLENTY of seats for additional payments of 50p, £1, £2 respectively. Hence the 'boundary values' for FULL against PLENTY for these three cards are 50p, £1 and £2 respectively. Provided that the true value lies in that sort of range we should be able to obtain an efficient estimate of it. Our estimate of the value of FULL versus PLENTY for the whole sample turned out to be 91p.

Seventeen sets of assumed values for DEP TIME, FULL, STAND30 and STAND60 were put through a computer program which created 1600 observations, each incorporating its own random error term. These 17 sets of synthetic data were then each analysed in the same way as proposed for the actual data. This procedure produced the 17 sets of estimates listed in Table 5 alongside the assumed values. Note that the estimates manage to track the assumed values over large ranges.

Table 4: Experimental Design for ECML Overcrowding Survey, 1987

	OPTION A			OPTION B		
CHOICE NO	FARE	SEATING	DEP TIME	FARE	SEATING	DEP TIME
1	A	FULL	A	+50p	PLENTY	A
2	–£1	FULL	A	A	PLENTY	A
3	A	FULL	A	+£2	PLENTY	A
4	–50p	STAND30	A	A	PLENTY	A
5	A	STAND30	A	+£1	PLENTY	A
6	–£2	STAND30	A	A	PLENTY	A
7	A	STAND30	A	+£5	PLENTY	A
8	A	STAND60	A	+£2	FULL	A
9	A	STAND60	A	+£5	FULL	A
10	A	STAND60	A	+£10	FULL	A
11	A	STAND60	1HR	+£5	PLENTY	A
12	A	STAND60	1HR	+£10	PLENTY	A
13	A	STAND60	1HR	+£20	PLENTY	A
14	A	FULL	2HR	+£1	FULL	A
15	A	FULL	2HR	+£5	FULL	A
16	A	FULL	2HR	+£40	FULL	A

Notes:

DEP TIME is specified as changes to the actual journey
'A' denotes that the variable is at the same level as for the actual journey made
The levels for the SEATING variable are explained in the text.

Table 5: Tests of the Experimental Design for ECML Overcrowding Survey

ASSUMED				ESTIMATED			
DEP TIME	FULL	STAND 30	STAND 60	DEP TIME	FULL	STAND 30	STAND 60
1.00	50	100	150	1.07	44	112	180
1.00	150	300	500	0.91	152	303	511
1.00	250	750	1500	0.98	236	695	1491
1.50	100	250	500	1.52	102	264	505
2.00	75	150	400	1.84	67	143	405
2.50	100	300	500	2.58	105	289	523
3.00	50	200	300	2.71	57	234	342
3.00	200	400	600	2.87	211	384	619
3.50	100	250	500	3.46	103	254	504
4.00	100	400	750	4.00	86	412	759
5.00	50	150	250	5.19	45	152	247
5.00	300	500	1000	4.69	283	521	991
8.00	150	300	750	7.44	200	291	813
10.00	50	100	200	10.53	47	96	202
15.00	150	350	1000	15.54	168	357	988
20.00	100	250	500	18.96	133	223	512
25.00	500	1000	2500	27.10	386	974	2309

Notes: The value of DEP TIME is given in pence per minute
The values of the other variables are specified in pence

In the event, our estimate of the values for our whole sample were 8.10, 0.91, 9.93 and 11.43 for DEPTIME, FULL, STAND30, STAND60 respectively. In retrospect, therefore, it would appear that the assumed values for STAND30 tended to be too low. Indeed, the fixed boundary values for STAND30 as against PLENTY, are clearly too low. From choices 4 to 7 in Table 4 these can be seen to be 50p, £1, £2 and £5. Hence the highest boundary value (£5) is well below the estimated value for the whole sample (£9.93). However, the tests did manage to recover £10 as £9.74 (see bottom line of Table 5) so maybe that wasn't too serious a problem.

There was a more serious problem with the design, which did not come to light at the testing stage. This was that the design was not able to cope with 'taste variation' in the sample, such that a small but significant minority indicated that they were willing to pay £40 not to have to adjust their departure time by 2 hours.

8. TASTE VARIATION AND SAMPLE SIZES

Conventional logit modelling is based on Random Utility theory, but the randomness is confined to an error term and does not affect the attribute parameters, ie. for individual i and alternative m with j attributes (X), denoting Random Utility by RU:

$$RU_{im} = U_{im} + \varepsilon_{im}$$

$$= \sum_j \beta_{.jm} X_{ijm} + \varepsilon_{im}$$

ie. the ß's do not depend on i.

Although this suggests that we would do well to move from conventional logit modelling to some more complex method, this is not necessarily the case. It has been shown (Fowkes and Wardman, 1988) that when we require estimates of relative variations, ie. ratios of parameters, as we usually do, then there is a general problem with using any such estimation method in the presence of taste variation. This is because our estimate of relative valuation is the ratio of two parameter estimates, 'averaged' as it were over the sample. Hence we have, as our estimate, a ratio of means. What we want, however, is to get the average relative valuation for the population being surveyed, ie. a mean of ratios. It is simple to demonstrate that, in general, a mean of ratios is not equal to a ratio of means. Our simulation work has shown that this problem is every bit as serious as using inappropriate conventional logit modelling.

Our recommendation, therefore, was that, as far as practicable, data sets should be 'segmented' during estimation, so that separate coefficients are calculated for groups of respondents with reasonably similar relative valuations. This can either be done by calibrating separate models or, with more effect, allowing one or more of the parameter estimates to take different values for each of the segmentation groups. For instance, if it is felt that relative valuations are likely to vary with the respondent's income, then the cost coefficient (if one is, as there usually is, present) can be allowed (by dummy variable techniques) to take on different values for low, middle and high income groups. Relative valuations are then computed for each income group using the appropriate cost coefficient.

In the overcrowding survey, we were unable to split out those willing to pay the £40. We tried to do it by journey purpose, income group, class of travel and so on, but to no avail. Only by removing choice 16 from the analysis could sensible estimates be obtained. The situation was studied thoroughly so that we could understand what had happened and justify what we had done. Thoroughly testing for all potential (and possibly peculiar) varieties and combinations of taste variation appears impractical at the moment.

Consideration of taste variation is a principal determinant of desirable sample sizes for SP. Without taste variation we can improve accuracy in two ways: either sample more individuals or ask a given group of individuals more questions. Perfect information from just one respondent would be better than slightly inaccurate information on thousands of respondents. However, in practice we would still want to average over several individuals since the mechanism by which the error terms is generated is somewhat vague and so day-to-day variation, for instance, would be missed by just surveying one respondent on one day. Where population sizes are small, as may be the case for freight decision makers, quite small sample sizes of, say, 4 or 5 respondents may suffice provided they are asked for sufficient choice responses to enable a reasonably accurate model to be fitted.

More usually, we will wish to allow for taste variation, as advised above, by segmenting the population into groups likely to have similar values. The conventional method then assumes negligible taste variation within segments. Because of the arbitrariness of some of this apriori segmentation we would usually wish to see sample sizes of, say, 20 to 30 for each segment, and this has become something of a 'rule of thumb'. Note that this is not affected by how many SP

choices we ask of each respondent. The result is that with the degree of segmentation now sometimes specified, we can be led to a desire for large SP sample sizes. For example, if separate segments were defined for the cross-classification of 5 income groups, 2 sexes, 3 journey distances and 3 journey purposes, we would already be up to 20x5x2x3x3 = 1800 even assuming that the sample could be evenly distributed over the segments.

9. THE SCALE FACTOR PROBLEM

A possibly very serious problem with Stated Preference methodology is that the errors people make when responding to SP surveys may not be the same as they would make in real life. This is important since logit modelling returns parameter estimates scaled relative to the estimated variance of the error term. For given relative valuations all the coefficients will be larger if there is little unexplained error variance than if there is large unexplained error variance. This does not matter when estimating relative valuations, since these are obtained as the ratio of estimated coefficients and so the scaling factors cancel out and have no effect. However, for forecasting and for estimating elasticities, we do need the estimates of the coefficients themselves, and not just their ratios.

It is probably helpful to think of the job of the scale factors to be to correctly weight the two modelled influences on utility, namely:

- the deterministic part,
- the random error.

Suppose that for a particular journey, our estimated relative valuation of time in terms of money (the value of time) suggested that the disutility of travelling by the slow, cheap coach was less than the fast, expensive train. The deterministic part of the model would say that, all else equal, everyone would choose coach. We would adjust this model by the average preference people have for travelling by train instead of coach (or vice-versa). Whatever this adjustment, it would still leave either everybody travelling by coach, or everybody travelling by train.

What is needed is the random error which will ensure that when there is really a big difference in disutility between the modes then one mode really will take virtually all the traffic, while when the difference in disutilities is very small, practically half will choose each mode. If the deterministic difference in utilities is 100, then a random error of $\pm$ 2 will have no effect and 100% will go by the better mode. If the deterministic difference is 2 and the random error is $\pm$ 100, then virtually equal shares will be predicted for the two modes.

The problem is, then, to obtain scaled parameter estimates in the correct scale to the error terms. There are two ways in which this can be achieved. Firstly, Revealed Preference (RP) data deals with actual choices so the actual level of error relative to the deterministic part can be estimated directly. Secondly, if we have good information on market share by each mode in known base conditions, it should prove possible to find an adequately good scale factor by choosing that which reconciles the model with this external information.

It will therefore be appreciated that the scale factor problem is at its most acute for SP studies, where there is no RP data also available, nor adequate mode share base data. It has been argued, particularly by Mark Wardman, that SP errors will tend to be larger than actual (or RP) errors, due to the hypothetical nature of the questions. A counter-argument to this would be that the attributes of both chosen and rejected mode are known exactly in SP, whereas in real life (and RP) they will only be known imperfectly. However, the balance of the evidence to date suggests that the former

case is true, namely that SP errors are too large, although not greatly so.

The implication of SP errors being larger than they ought is to give them overdue weight when forecasting, relative to the deterministic part. For example, if the deterministic part suggests that the choice should be train, and correct error terms would yield a prediction of 90% rail, 10% bus, the overly large (SP based) error might give something like 80% rail, 20% bus. Remember that with no error we would predict 100% rail, 0% bus, whereas with overwhelmingly large error we would predict 50% rail, 50% bus. An approach which we currently favour is to average the deterministic and probabilistic forecasts, which in the above example would be just right.

Where we have an RP and SP performed on 'comparable' data, we might rescale the SP coefficient to look like those of the RP model. This would involve running identical RP and SP models, determining the correct residual and scaling factor, and then rescaling more complicated SP models accordingly. In this way use of SP would enable forecasts to be made taking into account a wider range of factors than could be treated in the RP, and with a much greater level of accuracy.

The general method currently in use for obtaining forecasts and elasticities from SP data is to pivot on a 'known' elasticity. For example, if we assume that we know the price elasticity, then a journey time elasticity can be found, with a few assumptions, from the value of time. Since the value of time is a relative valuation (ie obtained from the ration of two scaled coefficients, so that the scale factor cancels) there is no problem.

An alternative approach, which is just becoming practicable, is to ask sufficient questions of each respondent so that individual models can be calibrated. For further discussion and investigation of the Scale Factor Problem see Wardman (1991).

10. CONCLUSION

Valuations of attributes are required in order to justify investment expenditures. Conventional economic techniques have not supplied much help in doing this. Revealed Preference analysis using disaggregate models has provided some estimates, but only rather imprecise ones at great expense and in specialised circumstances. The desire by the transport industry, in particular, for attribute monetary valuations has provided finance for a considerable research effort which led to the development of Stated Preference methods. Being borrowed from other disciplines, there exists scope for improving these techniques for the situations in which they are now being used. This paper presents an exposition of recent developments which have been incorporated in recent studies. The availability of sufficient computer power to enable thorough simulation testing of designs has meant that these new methods can be shown to be improvements and gain acceptance from those involved in investment appraisal.

Much research work is currently under way, particularly regarding computerisation of SP techniques. This allows the prospect of what is termed 'Adaptive Stated Preference' (eg Fowkes and Tweddle, 1988) whereby the attribute levels offered are adjusted according to earlier responses. In the terms of the discussion in this paper, this allows us to choose good boundary values to incorporate in the design at each stage, on the basis of past responses. In this way the computer can 'hunt down' the true valuation and confirm it by offering boundary values close by on either side. The method is still in its infancy but has enormous potential. Other current work is attempting to computerise the design procedure for SP (Holden, Fowkes and Wardman, 1992), based on the principles discussed earlier.

10. REFERENCES

Bradley, M.S., and Gunn, H.F. (1990) "A Stated Preference Analysis of Values of Travel Time in the Netherlands", Paper presented to TRB Annual Meeting, Washington

Fowkes, A.S., and Tweddle, G. (1988) "A Computer Guided Stated Preference Experiment for Freight Mode Choice", PTRC proceedings published as *Transportation Planning Methods* (P306), pp295-305, Planning and Transportation, Research and Computing, London

Fowkes, A.S., and Wardman, M. (1988) "The Design of Stated Preference Travel Choice Experiments with Particular Regard to Inter-Personal Taste Variations", *Journal of Transport Economics and Policy 22* (1), pp27-44

Holden, D.G., Fowkes, A.S., and Wardman, M. (1992) "Automatic Stated Preference Design Algorithms", PTRC Summer Annual Meeting, published as *Transportation Planning Methods*, PTRC London, pp.153-166

MVA Consultancy, Institute for Transport Studies (University of Leeds) and Transport Studies Unit (Oxford University) (1987) "*The Value of Travel Time Savings*", Policy Journals, Newbury

Wardman, M.R. (1988) "Comparison of RP and SP Models of Travel Behaviour", *Journal of Transport Economics and Policy*, (22), pp.71-91

Wardman, M.R. (1991) "Stated Preference Methods and Travel Demand Forecasting: An Examination of the Scale Factor Problem", Transportation Research 25A, pp.79-89