Monograph:
Fowkes, A.S. (1998) The Development of Stated Preference Techniques in Transport Planning. Working Paper. Institute of Transport Studies, University of Leeds , Leeds, UK.

Working Paper 479

ITS

Institute of Transport Studies
University of Leeds

Published paper

Fowkes, A.S. (1998) *The Development of Stated Preference Techniques in Transport Planning.* Institute of Transport Studies, University of Leeds, Working Paper 479

Working Paper 479

Revised November 1998

THE DEVELOPMENT OF STATED PREFERENCE TECHNIQUES IN TRANSPORT PLANNING

A S Fowkes

Institute for Transport Studies
University of Leeds

THE DEVELOPMENT OF STATED PREFERENCE TECHNIQUES IN TRANSPORT PLANING

This work was sponsored by the Engineering and Physical Sciences Research Council

Document Control Information

Title	:	The Development of Stated Preference Techniques in Transport Planning
Author(s)	:	A S Fowkes
Ref No	:	479
Version	:	2
Date	:	2.11.98
Distribution	:	
Availability	:	Unrestricted
File	:	k:\common\asf\reports\wp479pt1\pt2\pt3\pt4
Authorised by	:	A S Fowkes
Signature	:	Alan

1. INTRODUCTION

This paper looks at the development of Stated Preference (SP) techniques in transport. It is a mix of history, review, and explanation of key techniques and arguments. It attempts to avoid controversy. All those mentioned have played a part in the development or understanding of the techniques available today. With respect for these efforts we need to identify promising areas for the development of improved methods for the future.

In large measure, this paper reports what was written at the different stages of development of SP methods. Extensive quotes are included. While nothing that is now seen as plainly wrong has been left unremarked, there has not been any intention to critically assess each contribution. On some matters the jury is still out, whilst more generally it is true to say that SP design is a compromise which inevitably involves accepting some deficiencies to gain other benefits.

This paper arises out of an EPSRC funded project into Stated Preference design, and so the review concentrates on design matters. Naturally, problems with estimation using SP data have implications for the design of SP experiments, and so some attention is given to estimation where this is thought relevant.

By compressing so much into such a small space, the paper will be heavy going for many readers. Do not despair. It is strongly recommended that no more than a single subsection be attempted at a sitting.

Thanks to those who have helped, particularly with the arduous typing, and apologies to those whose work I have left out, misrepresented or maligned or whose name I have misspelt etc. Having taken two years to get this far, the thought of a further revision is too horrific to contemplate at the moment, but just in case I would be happy to receive comments, corrections, typos, improvements to references, additional references etc. Enjoy.

2. ORIGINS

2.1 Origins of Stated Preference

Researchers from many different disciplines have contributed to the development of Stated Preference methods. Perhaps the earliest documented relevant works relate to experimental economics. Swanson (1988) describes the following:

> "Experimental economists are concerned with testing the validity of assumptions that underlie normative models of behaviour. Kagel and Roth (1995) provide an extensive review of the field, and identify what might be the first application of Stated Preference. This was a study by Thurstone in 1931 (Thurstone, (1931)), who tried to estimate indifference curves experimentally by asking people to make choices between different combinations of coats, hats and shoes. Although he claimed success, Wallis and Friedman (1942) criticised his work in terms that have been used many times to attack SP. They said that people could not know how they would make such choices in reality, and their responses would be systematised to give plausible but spurious results. Nine years later, Rousseas and Hart (1951) tried to address this in a follow up study in which they asked people to choose between breakfast menus consisting of different combinations of bacon and fried egg. To introduce a note of realism, each subject was 'obliged to eat all of what they chose'. They also claimed success."

According to Wardman (1987), the origins of Stated Preference methods can be traced back to studies in the area of mathematical psychology in the 1960's. This work looked at how individuals combined information in the process of decision making. The paper by Luce and Tukey (1964) can be said to have begun the process, and introduced the name 'Conjoint Measurement'. The word 'conjoint' can just be taken to mean 'united', and by this Luce and Tukey meant that the alternatives in the decision could be viewed as the weighted combination of the various aspects, or attributes, of these alternatives. These ideas were taken up by economists, the paper by Lancaster (1966) being particularly influential.

Research followed into methods of collecting and analysing observed preference data (e.g. Anderson, 1970) and developing the theory of conjoint measurement (e.g. Krantz, Luce, Suppes and Tversky, 1971; Krantz and Tversky, 1971). The methods were taken up by market researchers as described by Wardman (1987);

> "Marketing research was quick to exploit the potential of these new techniques to forecast individuals' choices amongst consumer products. The paper by Green and Rao (1971) is commonly cited as the start of the use of SP methods in this field and the 1970's witnessed a large growth of interest. New techniques were developed and applied, such as the trade-off method (Johnson 1974) and methods using combinations of data (Akaah and Korgaonkar 1983; Green 1984) whilst an increasing number of algorithms for analysing responses were presented (Carroll 1972; Johnson 1973; Srinivasen and Shocker 1973) and tested (Carmone, Green and Jain 1978; Green and Srinivasen 1978; Wittink and Cattin 1981). Cattin and Wittink (1982) estimated that over 1000 commercial applications had been carried out in the decade up to 1980 in the US.
>
> "SP techniques were not adopted as quickly in transport economics, particularly in academic circles where they were regarded with some scepticism, and early applications were conducted by market researchers; for example, by Davidson (1973) in forecasting the demand for a new air service and by Johnson (1974) who examined preferences between the speed, seating capacity, price and warranty period of new cars. The early

transport applications in the UK were undertaken by transport consultancy agencies (Hoinville and Johnson 1971; Steer, Davies and Gleave 1981) whilst in the US their use was encouraged by public bodies such as the New York State Department of Transportation (Donnelly, Howe and Deschamps 1976; Eberts and Koeppel 1977; Koeppel 1977)."

2.2 Conjoint Measurement

It is generally agreed that 'Stated Preference methods', as discussed in this paper, arose out of market research techniques, termed 'conjoint analysis', developed in the United States in the early 1970's (see e.g. Kroes and Sheldon, 1988). In these early studies, the aim was to find utility weightings (or 'part worths') that were consistent with rank orderings provided by respondents offered option differing in terms of various attributes. These estimated utility weightings were required for each respondent separately.

Wardman (1987) briefly described the techniques of the early 1970's:

> "The techniques used in marketing to analyse ordinal SP responses, and which we have termed marketing models, include MONANOVA, LINMAP, PREFMAP and Johnson's trade-off algorithm. MONANOVA (Kurskal 1965) and PREFMAP (Carroll 1972) are monotonic regression methods. The procedures involve minimising badness of fit to obtain utility weights which best reproduce the rankings supplied. Johnson's trade-off algorithm (Johnson 1973) is similar to MONANOVA except that the input data are rankings supplied in two factor evaluations across a number of evaluations of several attributes. These estimation techniques are restricted to ranked data whilst MONANOVA is restricted to the part-worth function model. LINMAP (Srinivasen and Shocker 1973) is a linear programming method which also produces utility weights which minimise badness of fit. The input data is of paired comparison form and thus if n options are ranked, n(n-1)/2 paired comparisons are entered into the model. The technique is not restricted to either rankings or to the part-worth function model.
>
> "In the survey by Cattin and Wittink (1982), MONANOVA was found to be the technique most commonly used by commercial agencies for the analysis of ordinal data, although MNL was becoming increasingly popular. MONANOVA seems to be the most popular estimation technique in marketing research in general. In contrast with econometric models, marketing models are typically calibrated at the individual level."

Topics of concern were: (i) that the ranking might contain inconsistencies, such that no set of utility weightings would be consistent with the reported ranking and (ii) that several rather different sets of utility weightings might be consistent with the reported rankings. What was required was a statistical error theory, such that the 'deviance' of the data from the fitted model could be minimised and a best model found. However, no satisfactory statistical error theory emerged.

Louviere, (1988a) summarised the position as he saw it:

> "Conjoint analysis was popularised as a tool for the practical analysis of rank-order consumer judgement data by Green and Rao (1971) and Green and Wind (1973). The theory that underlies the design and analysis of rank-order judgement experiments was developed by several writers (for example, Luce and Tukey, 1964; Kruskal, 1965; Tversky, 1967), and is summarised in Krantz *et al.*, (1971). Unfortunately, the theory

and practice of rank-order judgement analysis are somewhat unrelated, since the methods of analysis (a) are not based on the theory, (b) do not have a statistical error theory, and therefore (c) cannot be used to test the adequacy of the theory.

"The axiomatic theory of rank-order conjoint analysis is called "Conjoint Measurement" (Krantz *et al.,* 1971). This theory requires real ranking data to satisfy a large number of ordinal conditions before one can conclude that a particular utility specification is appropriate for scaling (that is, estimating) part-worth utilities from an individual's judgement data. Most individuals are not perfectly consistent in their rankings, and therefore there is error in their data. However, Conjoint Measurement has no error theory on which to base statistical tests of part-worth parameters or competing utility specifications.

"Consequently, most practical and academic researchers who analyse ranking judgements assume that individuals' rankings are generated by a strictly additive (no non-additivities or interactions) function of the unknown part-worth utility measures. Part-worth utilities are estimated by least-squares procedures (for example, MONANOVA) that optimise the fit between observed and predicted rankings, *assuming* that an additive utility specification is correct. "Badness-of-fit" statistics known as "stress" measures are used as an index of how well additive or other specifications fit the observed rankings.

"Unfortunately, "stress" measures are closely related to the quantity $(1 - R^2)$; and it is well known that R^2 is an unreliable measure of the adequacy of conjoint models (not to mention other models). Dawes and Corrigan (1974), Wainer (1976) and Anderson and Shanteau (1977) are among those who have demonstrated that (a) conjoint experiments ensure high goodness-of-fit or low badness-of-fit measures, (b) many possible specifications can produce approximately equivalent fit measures, and (c) wrong specifications can produce "better" fit measures than "right" specifications in real, fallible data.

"In particular, factorial-type experiments guarantee that main effects or other simple specifications will account for most of the variance in judgement data, even when wrong. This happens because "true" utility functions are conditionally monotone in each attribute, and the joint combination rule can be well approximated by functions that predict "higher overall utility corresponds to more high part-worths" and "lower overall utility corresponds to more low part-worths". Conjoint models mimic these conditions very well."

These are harsh words against the methods used in these early studies. Nevertheless, estimates of (part-worth) utility weighting were derived for each individual, permitting distributions of valuations to be derived, rather than just an estimate of the mean valuation. To see an example of how this was done, let us consider the technique called MONANOVA (Monotone Analysis of Variance) as used by Steer Davies and Gleave Ltd (1981), with results published in Sheldon and Steer (1982).

2.3 Case Study - Steer Davies and Gleave Ltd Service Frequency and Through Trains Report 1981

The technique involved presenting respondents with alternative packages, each described by attributes set at particular levels. The option of presenting paired comparisons was rejected, and accordingly respondents were each asked to rank 10 alternatives. Since the responses were the ranks, i.e. ordinally scaled, the MONANOVA technique was selected for estimation.

The design of the survey incorporated 3 attributes each at 3 levels. This gave $3^3 = 27$ possible combinations that might be presented to respondents. This is called a *full factorial design.* A Latin Square orthogonal synmetic statistical design was used. This is an example of a fractional factorial design. This reduced the design, so as not to overload the respondents, to 9 alternatives, i.e. a square with side = 3. Let A1, A2, A3 be the three attributes and L1, L2, L3 denote the three levels of each, then the Latin Square looks like:

	A2 : L1	A2 : L2	A2 : L3
A1 : L1	A3 : L1	A3 : L3	A3 : L2
A1 : L2	A3 : L2	A3 : L1	A3 : L3
A1 : L3	A3 : L3	A3 : L2	A3 : L1

i.e., each level of attribute A1 has its row, each level of attribute A2 has its column, and each level of attribute A3 appears once in each row and one in each column.

In the SDG experiment, the 3 attributes were FREQUENCY, JOURNEY TIME and FARE. Each had the same 3 levels, namely BAD, CURRENT and GOOD. Working across the rows of the Latin Square, and then down, the 9 combinations forming the alternatives presented to the respondents are as shown in Table 2.1. Note, however, that these were shuffled before being presented to the respondents.

Table 2.1 : Latin Square design as used in Steer, Davies, Cleeve Ltd (1981)

Alternative	(A1) Frequency	(A2) Journey Time	(A3) Fare
1	BAD (L1)	BAD (L1)	BAD (L1)
2	BAD (L1)	CURRENT (L2)	GOOD (L3)
3	BAD (L1)	GOOD (L3)	CURRENT (L2)
4	CURRENT (L2)	BAD (L1)	CURRENT (L1)
5	CURRENT (L2)	CURRENT (L2)	BAD (L1)
6	CURRENT (L2)	GOOD (L3)	GOOD (L3)
7	GOOD (L3)	BAD (L1)	GOOD (L3)
8	GOOD (L3)	CURRENT (L2)	CURRENT (L2)
9	GOOD (L3)	GOOD (L3)	BAD (L3)

The main feature of Latin Square designs is that the levels of the three attributes are orthogonal, i.e. statistically uncorrelated. This does not mean that the attributes themselves are uncorrelated in the real world (e.g.. services with longer than average journey times, i.e. BAD, might have lower than average fares, i.e. GOOD, in order to compensate and maintain patronage) or that respondents' valuation of the attributes are uncorrelated (e.g. travellers might be prepared to put up with high fares more easily if the journey time was good). This last point is at the heart of the matter, since our choice of design will be influenced by how we think the attribute valuations are correlated.

The reason we often choose uncorrelated designs, as in Table 2.1, is that each attribute is orthogonal to the others, i.e. knowing that A1 = BAD tells me nothing about the levels of A2 and

A3 being offered. Each level of A2 has equal choice of occurring with any given level of A1, and the same goes for A3. This allows respondents to have a free hand in showing their opinion of a given level of A1, regardless of their option of the level of A2 and A3. If Table 2.1 were altered such that the JOURNEY TIME level (column A2) was always the same as the FREQUENCY level (column A1), i.e. BAD always with BAD, CURRENT always with CURRENT etc., then we would have perfect correlation in the design and we would not be able, at the estimation stage, to disentangle the two effects. If respondents ranked the GOOD FREQUENCY/GOOD JOURNEY TIME alternative the highest, we would not know whether they liked the high frequency, the low journey time, or both.

Orthogonality of attribute levels clearly has some desirable aspects, but the Latin Square is not free from correlations. By definition, if we know the levels of any two attributes, then the level of the third is known. For example, if respondents were tolerant of poor frequency as long as journey times were good, or of poor journey times as long as frequency was good, but would be very resistant to them both being poor (i.e. BAD) then the only way they would have of showing this would be to rank alternative 1 as the worst. If the fare level for that alternative had been good, then we might have been able to spot this interaction effect, but since the fare is also BAD, the fact that this alternative is ranked worst is no surprise, and tells us very little. It certainly does not tell us of an interaction between FREQUENCY and JOURNEY TIME.

The above method permits designs for three attributes at any number of levels - the more levels the bigger the square. However, the number of levels does need to be the same for each attribute. This restriction can be overcome by considering Latin Squares with missing rows. For example, suppose that attribute A1 has 3 levels, but that attributes A2 and A3 have 4 levels each (L1 to L4). We can then consider a 4 by 4 Latin Square with one missing row. Below we have taken row 4 to be missing, but show what it would have been:

	A2 : L1	A2 : L2	A2 : L3	A2 : L4
A1 : L1	A3 : L1	A3 : L2	A3 : L3	A3 : L4
A1 : L2	A3 : L4	A3 : L1	A3 : L2	A3 : L3
A1 : L3	A3 : L3	A3 : L4	A3 : L1	A3 : L2
Missing	A3 : L2	A3 : L3	A3 : L4	A3 : L1

The '12 alternative design' is produced by spelling out the first 3 rows, in the same way as was done in Table 2.1. With one row missing, the remaining rows are still orthogonal to columns and the levels of A3, because each row still contains all 4 levels of A2 and all 4 levels of A3. However, we have lost the Orthogonality been A2 and A3, since each column no longer contains each level of A3 (since the fourth row is missing). Situations like this, where designs incorporate only partial orthogonality are commonplace in statistics, but in transport there has been a strong tendency, supported by little thought), to stick with completely orthogonal designs. At first sight this would appear to place considerable restrictions on the size and shape of designs, but in fact the problem has been heavily researched and suitable experimental designs catalogued e.g. Addelman (1962), Hahn and Shapiro (1966). Wiley (1977) considered ways of avoiding dominated or dominating alternatives in the designs, i.e. 'selecting Pareto Optimal Subsets'. A condensed version of the Hahn and Shapiro tables was reproduced as Appendix A of Kocur et al (1982) and has often been referred to in the transport profession as the 'cookbook'. This source was not available to SDG when they drew up Table 2.1, but they were aware of the general literature on experimental design (e.g. Box, 1952, and Winer, 1962).

Considering Table 2.1, the usual response variable in statistics would be a measurement, such as a preference rating, which could be analysed by regression techniques. Since SDG only asked for a ranking, a technique such as MONANOVA had to be used. The procedure transforms preference data such that a 'best' model can be developed to explain the observations. In the present context this is done by estimating part-worths or preference weights w_{ij} of this model:

$$U_k = \sum_i \sum_j \frac{k}{\delta_{ij} w_{ij}} \tag{2.1}$$

where w is the part-worth (or preference weighting) of level j of attribute i and

$$\delta_{ij}^{k} = \begin{cases} 1 & \text{if level 1 of attribute i is present in option (package) k} \\ 0 & \text{otherwise} \end{cases}$$

In the Service Frequency and Through Train Study, respondents were asked to rank nine different service packages. The MONANOVA program starts by producing an initial estimate of the part-worths w_{ij} and checks whether these estimates are capable of reproducing the observed rank orders using the model of equation (2.1) above. If the observed ranking is not reproduced MONANOVA searches over all monotone ascending transformations of the data and selects the one that most closely reproduces the ranking. The criterion for closeness of fit is the square root of the difference between the modelled and observed ranking divided by an appropriate scaling factor.

The resultant part-worths, w_{ij}, are averaged over all respondents within the market segment of interest. The next stage is to transform into differences; here six differences comprising low to medium and medium to high levels for each of the three variables. These difference measures are then normalised by dividing by the magnitude of the change in the variable in the options as offered to the respondents. For example, fares were altered from the medium level by +10% to give the high and low fare levels. Hence to normalise the difference between medium and low fares they divided by 10%, i.e. 0.1.

In order to arrive at journey time elasticities, a pivot elasticity (usually for price) is necessary, where ratios of demand elasticities are assumed equal to the relative size of the estimated preference weightings. We are not aware of any convincing arguments to justify taking ratios of MONANOVA preference weightings to give ratios of elasticities and therefore have severe doubts on this score, quite apart from the choice of pivot elasticity.

Our view of MONANOVA itself, is that it may have been appropriate to the early days of microcomputers, but that computing power is now sufficiently inexpensive to require the use of a technique to optimise the choice of part-worths, rather than merely finding part-worths that do not contradict the data. We find this extra source of uncertainty to be unsupportable in current conditions, and for this reason reject the continued use of MONANOVA.

Besides the experiment involving service frequency, seen in Table 2.1, a separate experiment was carried out regarding through trains. This fourth attribute, at only two levels (have to change, through train) was incorporated within the 9 alternative design by means of an asymmetric orthogonal design derived by 'collapsing' the allowed for 3 levels of the fourth attribute to the 2 that were required. Table 2.2 shows the design, and indicates where the collapsing was introduced as (L3 $\rightarrow$ L2). This gave twice as many occurrences of level 2 than of level 1 for attribute A4.

Table 2.2 : Asymmetric orthogonal design, as used by Steer, Davies, Gleave Ltd (1981)
(see text)

Alternative / Attribute	A1	A2	A3	A4
1	L1	L1	L1	L1
2	L1	L2	L3	L3 → L2
3	L1	L3	L2	L2
4	L2	L1	L2	L3 → L2
5	L2	L2	L1	L2
6	L2	L3	L3	L1
7	L3	L1	L3	L2
8	L3	L2	L2	L1
9	L3	L3	L1	L3 → L2

No consideration was given to incorporating a good range of boundary values, but we will leave that topic for later (Section 3.6). Rather, the ranges were determined for each attribute around its current level. In relation to the experiment in Table 2.1, consideration of the current understanding of elasticity measures suggested that fares and journey times were roughly equally important to customers, while frequency was of less importance. On this basis fares and journey times were, initially, specified as ±20% of current levels (for bad and good) with frequency ±45%. Following piloting, the range of fare charges was reduced to ±10%, as this attribute was otherwise dominating the choices.

The researchers lay great stress on their use of what they termed 'The Journey Planning Game' (described below) to overcome what they saw as 'the information dilemma'. In order to properly explain the effect of the change in levels of service that the respondents were to choose between, it would be necessary to 'educate' respondents in such a way that they could no longer be considered to be representative of the travelling public. On the other hand, they felt that if no explanations were offered, respondents would be unlikely to be able to understand the implications of the choices presented to them.

The journey planning game was developed to overcome the problems thought to be posed by the information dilemma. They outlined three key features:

(i) respondents should be asked to reconsider their travel behaviour within the context of the set of real activities and actual journey(s), not on the basis of hypothetical possibilities;

(ii) the interview procedure should allow for a replication of information channels (e.g. looking at timetables, asking enquiry clerks etc.) used by the respondent to understand the characteristics of the travel options available;

(iii) the interviewer should first establish the extent to which the respondent has knowledge of the transport options and then use these levels of knowledge as controls to describe the other options which the respondent has to appraise.

The practical manifestation of this was that the interviewer first established the information constraints under which the passenger had elected to operate and then produced equivalent information for each of the alternative packages the respondent had to rank, e.g. replicating the enquiry and exchange between the passenger and the clerk at a BR station. Respondents were asked to re-plan the journey they were currently undertaking, together with its various constraints (e.g. travelling with children, having to be at a meeting for 11.00 etc.).

In discussing the merits of the above procedure, let us go through the three points listed in the above in turn. Firstly, we have great sympathy for point (i), i.e. that the trade-off exercise should be set within the actual constraints faced by travellers for the journey on which they were intercepted. It would be unproductive, say, to ask holidaymakers to pretend to be businessmen, or for others to be asked to imagine they were travelling with an arbitrary number of children. Clearly, the people who best know about such circumstances are the people in those circumstances.

However, the above need not be taken as an unbreakable rule. Respondents might be asked about other recent journeys they have made, but again subject to their constraints at that time. Respondents might also be asked to consider completely hypothetical means of transport, e.g.. re-opened rail services, in which uncertain circumstances we might be content to set the context as that of a 'typical' trip in some sense. Moreover, some 'constraints' are self made in the sense that meeting times are arranged in the light of knowledge of the current timetable. Hence, 'having to be somewhere by 11.00' might not mean that a revised train service that got you there at 11.05 was any worse than that actually obtaining.

With regard to (ii) we can see some advantage in replicating the information channels available to the respondent, but in practice feel that this would be more than offset by the consequential interviewer bias. There is no way that the hesitancy to queue and approach a BR enquiry clerk, for example, can be replicated by the interviewer who has already been in conversation with the respondent for several minutes. Some respondents will be eager to please and will seek out much more information than they otherwise would. Others will be impatient with the interview and, discerning minimal penalties for error, will seek less information than they would otherwise and might assume any items not explicitly mentioned to be unchanged from the present service. In short, we feel that this is an unwarranted complication.

Turning to (iii), which may be characterised as working within the respondents' level of knowledge, we see no benefit from so doing. If we wish respondents to rank options, we should give them all the attributes for each option, and not rely on them to ask. We feel this may partly explain why the SDG results tended to show greater values for worsenments compared to improvements. If people are reasonably happy with their present service they would be unlikely to find out all the aspects of an improved service, whereas a worsened service which prevented them from doing something they presently do would lead them to make a fuller investigation of the alternatives available to them. Even if they do not so enquire it may still be the case that the worsened service forces them to see the worsenment, while an improved frequency could be ignored if there was already a conveniently time train.

2.4 Case Study : Cranfield Work on Reliability

Stated preference methods were used by researchers at the Centre for Transport Studies, Cranfield Institute of Technology, whilst investigating passenger attitudes to lateness (Benwell and Black, 1984, 1985). This was a pioneering study in the area of user valuation of reliability. Respondents were offered a trade-off exercise between money and minutes late, the latter being presented as a distribution over 10 trains. Initially it had been desired to present a trade-off exercise between journey time and minutes late, but this alternative was rejected because it was found at the feasibility stage that respondents found some difficulty in discriminating between journey time and late time. Experience from other surveys, however, suggests that this should not have been a serious problem.

Respondents were presented with seven cards, each containing a change in fare from the present, and the lateness for each of 10 trains, Table 2.3. The idea was that respondents should consider the uncertainty in arrival time, as well as the (statistically) expected lateness, when placing the cards in rank order.

Table 2.3 : The Cranfield Lateness trade-off analysis

Card No.	Change in Price Over Current Fare (£)	The Distribution of Lateness (cards show number of minutes late)										Total Minutes Late
1	-5	21	23	26	0	0	0	36	33	0	0	139
2	0	2	0	13	8	5	0	0	0	0	4	23
3	0	33	0	24	0	0	0	0	0	0	0	57
4	+5	8	0	0	4	8	2	0	0	13	0	35
5	+10	7	0	0	0	0	0	0	0	5	0	12
6	-5	0	0	25	0	0	34	0	0	0	0	59
7	+5	0	33	0	0	0	0	0	0	25	0	58

There were 15 combinations of the seven cards which correspond to positive values of lateness. However, it is reported (Banwell and Black, 1984, page 9) that "a close examination of the combinations given by respondents shows that less than 10% correspond exactly with one of the 15 combinations that a rational application of decision rules would imply". Since the researchers have deliberately set up a third attribute varying over the cards, namely the dispersion of lateness, there is definitely no case for this use of the label 'rational'. The researchers chose to ignore any systematic response to the dispersion of lateness, and merely allocate those (90%) 'irrational' respondents to that one of the 15 'rational' rankings that was closest to their own ranking.

The result was a bimodal distribution of the estimated value of lateness, which even persisted when broke down into first and second class travellers. Although the bimodal distributions are not implausible, the shape is somewhat unexpected and there must be a possibility that the estimation method had worked to produce a misleading result. Without the detailed data it is not possible to investigate this further.

Notwithstanding the above comments, it is nevertheless a lesson that can be taken from the Cranfield work that relative value distributions should be plotted out and presented in reports. It is certainly true that many researchers in this area do try to plot results, but when more than two attributes are varying this requires that all but two are set to some value. Because of correlations of valuations in the population (e.g. those with a high value of time also having high valuations of many other things, due to the influence of income) this may often not be very satisfactory. Perhaps for this reason, plots of relative valuations are rarely presented in reports.

A second strand of estimation used in the Cranfield Study was to apply regression analysis. Ordinary Least Squares regression gave average values of lateness not too dissimilar from hose from the method discussed above. The preference ranking number (one to seven) for the cards was regressed against fare difference and total minutes late. It was appreciated that this was rather a crude exercise, and so further runs were carried out using logit and probit analysis. The researchers report that these 'superior estimation techniques' did not give significantly different results.

2.5 Priority Evaluator

In the late 1960s a series of small scale studies was undertaken by Social and Community Planning Research (SCPR) on behalf of the Highway Economics Unit of the Ministry of Transport (see Hoinville and Courtenay, 1978). The complete project formed an experimental piece of research designed to test a particular method of establishing priority preferences. They examined the relative importance of factors determining the choice of mode used for the journey to work. The underlying aim of the project was to arrive at a method for deriving values placed on savings in journey time.

The method that SCPR tested (Hoinville and Berthoud, 1969), termed the Priority Evaluator, was developed from a research 'game' used by the Institute of Research in Social Sciences, University of North Carolina, as part of a study to see what kind of city people wanted. In one game, the respondent was told that he had won a house in a television show competition. It was not yet built and he was given \$2000 to spend on 34 different amenities. These were priced mostly between \$500 and \$50. After he had made these choices, and the results noted, the sum was increased to \$3000 and he was asked to choose again.

The basic idea that SCPR took from the North Carolina game was the allocation of fixed resources between a range of costed alternative choices. Typically there will be a range of factors on which expenditure can be made, and for each factor a range of levels of expenditure that can be made. As the number of factors and number of levels rise, the number of possible choices that exhaust the available budget quickly escalates. The challenge then turns on how to estimate relative valuations from such data. In general, such estimation will require that the available budget is full exhausted. In any event, responses can be expected to vary if there is an option of 'banking' some of the budget rather than spending it amongst the factors being considered. If the total budget is to be exhausted, the number of possible choices will increase, at first, as the budget is increased, but then decrease until once the budget is just sufficient to buy everything on offer there is only one possible choice, i.e. 'buy' it all. Hence, careful choice of budgets to be used in the game can keep the number of possible choices down to an acceptable level.

Hoinville and Berthoud identified two basic approaches to the analysis of Priority Evaluator (PE) responses. The first is to treat responses as a pattern of choices and to use a form of cluster analysis to determine the main types of pattern which emerge. The second, and more traditional approach, is to look at aggregate choices compared to a random distribution of choices. This second method was used by SCPR. Where 'improvements' were bought by more people than would be expected by chance it was inferred that the aggregate relative valuation of this factor was higher than the relative price of that factor implicit in the game, and vice versa. A number of simplifying assumptions were then made so that a relative value index could be derived as:

$$RV(X1) = \frac{RP(X1) \,.\, AC(X1)}{EC\,(X1)}$$

where X1 is one of the positions for one of the factors in the experiment
RV(X1) is the relative value index of X1 compared to its base

RP(X1) is the relative price index of X1 compared to its base
AC(X1) is the number of respondents at position X1
EC(X1) is the number of expected choices at position X1.

The SCPR experience can be summarised as follows. Firstly, if respondents are allowed too great a budget, the estimated relative valuations will approach the relative prices designed into the game. This is because respondents having radically different relative valuations will quickly spend all they are allowed on their preferred factor. They are then forced to spend their remaining budget on factors of secondary importance. In the extreme; when the budget is just sufficient to buy everything, respondents will have to do that, and their relative valuations, determined as above, will exactly match the relative prices designed into the game.

Secondly, SCPR noted a 'dustbin' effect, such that low priced improvements were bought with left-over expenditure, after the main choices had been made, and possibly as the only means of exactly exhausting the budget.

Thirdly, when the budget is small, some improvements can be undervalued if they are offered in highly priced increments. This is because respondents appear reluctant to use (virtually) their total budget on just one factor. This is a particular manifestation of the more general problem of the 'blocked' nature of permitted expenditures preventing fine tuning. One other manifestation of this is where respondents are allowed to revise their earlier expenditures on being awarded an increase in budget. They now feel they can afford a highly priced improvement and are willing to forego the cheaper improvement they had bought out of their previous lower budget. When analysed in the manner discussed above, this causes some element of 'leap frogging' in the relative values as wealth increases.

In the 1980's the MVA Consultancy developed the Priority Evaluator method (Copley, Bouma and de Graaf, 1987). Estimation was by a maximum likelihood disaggregate logit program, and so represents a considerable advance, but most of the problems discussed above remain and new ones arise.

2.6 Mainstream Development

Out of the above range of approaches a mainstream approach began to develop in transport. The essence of current mainstream methods can be seen in Bates and Roberts (1983), which reports their recent experience. They note that calibration of discrete choice models requires data involving trade-offs, which are often extremely difficult to observe.

> "To calibrate a discrete choice model successfully requires a certain distribution of the sample of respondents with respect to the crucial variables in the model. To put it crudely, if every respondent is in such a position that his preferred option is overwhelmingly superior to all the other options, it will be very difficult to make an accurate estimation of the trade-offs which we believe to be generally taking place. To do this, we rely on a reasonable sample of 'marginals', that is, people who are on the borderline between one option and another. In many travel contexts, this is extremely difficult to achieve.
>
> "With this problem in mind, there is currently a considerable amount of interest in whether a technique that has been widely used within market research - one of presenting the respondent with a number of hypothetical 'packages' - represents a valid way of obtaining data on trade-offs. Although the technique has been used under a variety of

guises, we will refer to it by the generic term 'stated preference', since the essence is that the respondent is asked to state his preference among a number of hypothesised options, rather than the revealed preference data standard in economic work, where the only acceptable information is considered to be the actual choice observed by the analyst."

The paper goes on to describe an experiment involving long distance travel in the Netherlands. For the mode choice experiment there were two attributes (cost and time) for each mode. Each of these four attributes was entered into the design at only two levels: high and low, with numeric values substituted. There were 10 choices to be made (called here 'options') and the design is shown in their Table 3, reproduced here as Table 2.4. Nothing is said on the origins of the design. It is not a Latin Square, and it does not appear in Kocur et al (1982). However, the attribute levels appear well chosen, as can be seen from the final three columns of the table. These show the estimated probabilities from models calibrated on the whole data set, and separately for those respondents who actually use rail and for those who actually use car. The sample of rail users were much more likely to choose TRAIN all else equal, i.e. for each of the 10 options presented. This is not surprising, but this effect is often overlooked when designing SP experiments. More importantly, though, we note that for both rail and car users separately, the range of options gives rise to a large range in the estimated probabilities of choosing train.

Table 2.4 : Estimated probabilities of choosing train given alternative data on times and costs

Option	CC	TC	CT	TT	Rail sample	Car sample	Combined
1	L	L	L	L	0.72	0.50	0.60
2	H	L	L	L	0.91	0.78	0.84
3	L	H	H	L	0.73	0.53	0.62
4	H	H	H	L	0.91	0.80	0.85
5	L	L	L	H	0.40	0.21	0.29
6	H	L	L	H	0.72	0.47	0.58
7	L	H	H	H	0.41	0.23	0.31
8	H	H	H	H	0.73	0.50	0.60
9	L	L	H	L	0.86	0.76	0.81
10	H	H	L	L	0.82	0.55	0.67

(Note : CC = car costs, TC = train costs, CT = car times, TT = train times, L = low value, H = high value).
Source : Bates and Roberts (1983)

It is worth discussing this last point in more detail. Taking first the rail sample, the probability of choosing train ranges from 0.4, when train is slow (but cheap) and car is both cheap and fast, to 0.9 when train time is fast and car is expensive. For the same cases the probability of a member of the car sample choosing train varies from 0.2 to 0.8. Had the high and low numeric values (i.e. the 'levels') for train time and car cost been closer together, then we can deduce that there would have been less spread in the estimated probabilities, i.e. less trading. This is analogous to the situation of trying to find the gradient of a straight line relationship subject to a small error term. If we were to take two observations close together on the line, and measure the change in height, we would find a relatively small value which may be heavily contaminated by the effect of the error term. The solution is to take your readings of height at two places on the line as far apart as possible. In that way we can get the biggest gain in height possible, influenced only slightly by the error term which when divided by the known interviewing difference, will give the slope to the greatest possible accuracy.

This argument says that we should not choose levels for our High and Low that are too close together. However, in the case of logit modelling, it does NOT follow that the levels should be placed are far apart as possible. This is because we are estimating probabilities, which have to lie on the (0, 1) interval. By making the differences between the High and Low levels too great we would run the risk of having *all* respondents making identical choices on each scenario. This would effectively reduce our sample size to just 10 observations (albeit repeated many times). Given the extreme nature of the data, only a very poorly calibrated model could be hoped for.

Returning to our straight line analogy we might write:

$$Y_i = a + \beta X_i + \varepsilon \qquad (2.2)$$

where ε is distributed Normally, with mean zero and variance σ^2. We estimate β as $\hat{\beta}$ by taking measurements Y_1 and Y_2, where $X = X_1$ and X_2, and calculating

$$\hat{\beta} = \frac{Y_2 - Y_1}{X_2 - X_1} \qquad (2.3)$$

Bates and Roberts had 4 explanatory variables (attributes) and a response variable on the semantic scale, R:

R = 1 Definitely choose train
R = 2 On balance, choose train
R = 3 Indifferent
R = 4 On balance, choose car
R = 5 Definitely choose car.

This response was converted into a probability of choosing train, P, as follows:

$$P = 1.1 - 0.2R \qquad (2.4)$$

In order to avoid predicting P values outside the (0, 1) interval, the Berkson-Theil logit transform was used to give the dependent variable, Y, for the regression:

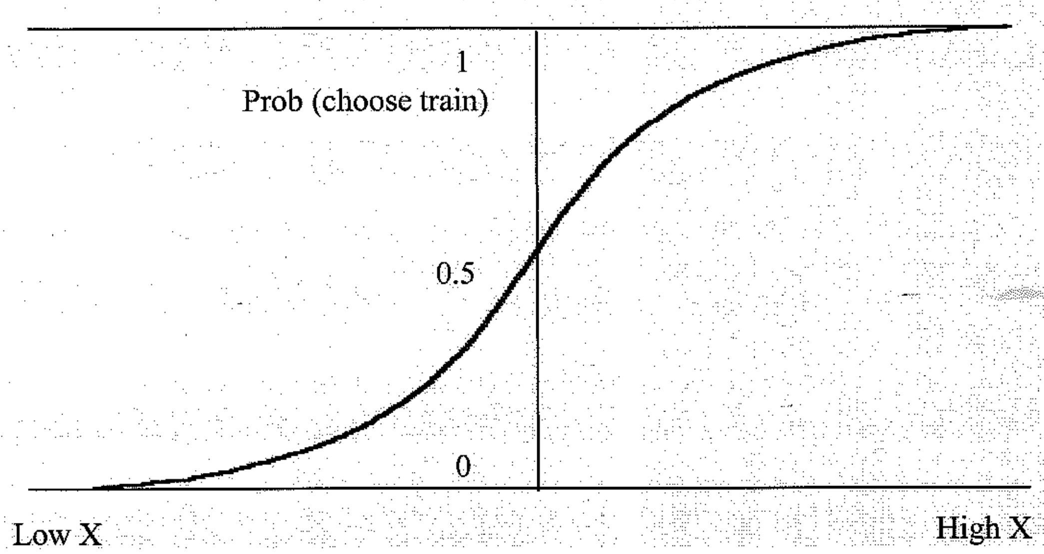
$$Y = \log\left(\frac{P}{1-P}\right) \qquad (2.5)$$

The regression equation is:

$$Y = a_r + b_r\,COST_r + b_c\,COST_c + c_r\,TIME_r + c_c\,TIME_c + e \qquad (2.6)$$

where ε is again Normally distributed with mean zero and variance σ^2. However, because of transformation (2.5), the error term for a probability (P_i for when $X = X_i$) is no longer Normal. The earlier argument, with regard to straight line estimation, of having the X values are far apart as possible, no longer holds. Instead, doing this just pushes us further and further into the areas where the probability is either very near unity or very near zero. (See Fig. 2.1)

Fig. 2.1 Illustrative diagram of a cumulative probability distribution such as the logit



Points in these areas tell us very little about the rates of trade-off in equation (2.6), for example the value of time in train, (c_r/b_r).

Just how far we should spread the high and low values is a matter this project will be discussing elsewhere. This was not a matter that had been much further investigated in the interim. Attention was more focused on improving the information content of choices offered to respondents. We now turn to look at developments in SP design after 1983.

Where data is collected in the form of discrete choices, i.e. (0,1) data, it is not possible to calculate y for each response. One possibility is to use 'aggregate SP' data (Louviere and Hensher, 1982), i.e. take P to be the proportion of respondents choosing option 1 on a particular scenario. The problem is then that the total number of observations available for analysis is just the number of scenarios presented to each respondent. This will rarely be sufficient to enable to model to be well calibrated. The number of observations available for calibration could be increased by a factor k if respondents were split into k groups, possibly facing a different set of scenarios - but not necessarily, and P calculated for each scenario within each group, using the following amended formula. –

$$P = \frac{r + 0.5}{n - r + 0.5}$$

where r is the number of respondents choosing option 1 for that scenario in that group, and n is the number of respondents to that scenario in that group.

A second, and much more popular, possibility is to use a maximum likelihood discrete choice logit estimation procedure. Several pieces of software were written to facilitate this. A notable early piece of software was MIT's MLOGIT (written by Chuck Manski, modified by Moshe Ben-Akiva, and documented in Howe and Liou, 1975), but by the time of the early SP studies discussed here David Hensher's BLOGIT (documented by Crittle and Johnson, 1980) was preferred - in the UK at least.

A further development was possible where each scenario contained several alternatives which the respondent was asked to rank. This data was handled by the 'ordered' or 'exploded' logit model (Beggs, Cardell and Hausman, 1981; Chapman and Staelin, 1982). If the scenario contains m alternatives and the alternative preferred is ranked first, then this alternative is taken as preferred over all the others. Then the alternative ranked second is taken as preferred over the m-2 alternatives not ranked 1 or 2. And so on, until the alternative ranked m-1 is taken as preferred to the alternative ranked m. In effect, m-1 choice sub-sets have been constructed, where we know the preferred option in each, and so we have generated m-1 observations for each respondent. Naturally, with many alternatives to rank, the respondent might become fatigued. This might either affect the middle rankings, where the choices might be expected to be less clear cut than at the extremes, or else respondents might work their way down from rank 1 and become fed up before reaching the end of the exercise. In either case, the general advice is to place most reliance on the four, or so, most preferred alternatives and curtail explosion after these. For design, it would seem more reasonable not to ask for large numbers of alternatives to be ranked in the first place.

2.7 Conclusion

The rest of this paper deals with post 1983 developments in Stated Preference design in Great Britain. Conjoint Analysis techniques have continued to be developed elsewhere (see Wittink and Cattin, 1989; Wittink, Vriens and Burheme, 1994; Oppewal, 1995; Carroll and Green, 1995; Carmone and Shaffer, 1995; Huber, 1997; Swanson, 1998).

3. STATED PREFERENCE DEVELOPMENTS IN TRANSPORT, 1983 ONWARDS

3.1 Introduction

We have seen in the previous section, how SP methods were introduced into transportation analysis. From about 1983, several practitioners in transportation began to feel able to suggest improvements to the methods, and the assumptions of the "received wisdom" came under close scrutiny. The availability of computer software to calibrate reasonably large disaggregate choice models greatly facilitated these developments.

3.2 Louviere, Hensher and Woodworth, 1982/3

Hensher (1994) writes:

> "Although it is always difficult to pinpoint the major events which heralded in the beginning of a widespread interest in SP methods, the motivation seems to have evolved from a number of applications in which the behavioural response involved an alternative which was either not currently available (e.g. Louviere and Hensher 1983, Hensher 1982) or where there was difficulty in assessing substantially different attribute mixes associated with existing alternatives to those observed (e.g. Kocur et al 1982, Hensher and Louviere 1983, Bradley and Bovy 1985, Louviere and Kocur 1983). An important paper by Lerman and Louviere (1978) demonstrated the theoretical links between revealed preference and stated preference models.
>
> "Prior to the paper by Louviere and Henser (1983), the emphasis had been on judgmental tasks in which a respondent was asked to rate or rank a number of attribute mixes associated with a particular choice context. The modelling of this data using standard regression-based estimation procedures required simulation of choice environments in order to predict market share. Louviere and Hensher showed how a preference experiment (i.e. a number of alternative mixes of attributes) could be extended to incorporate choice experiments in which an individual chooses from among fixed or varying choice sets, enabling estimation of a discrete-choice model and hence direct prediction of market share. Stated *choice*-experiments are now the most popular form of SP method in transportation and are growing in popularity in other areas such as marketing, geography, regional science and tourism. The papers by Louviere and Hensher (1982) and Louviere and Woodworth (1983) have become the historical reference sources for stated choice modelling in transportation.
>
> "The introduction of stated choice modelling using the set of established discrete-choice modelling tools routinely applied with revealed preference data widened the interest in SP-methods. For the first time travel behaviour researchers could see the benefit of stated-preference data in enhancing their travel choice methods. This I would argue was the major watershed which after 10 years has results in widespread acceptance of SP methods in practice in transportation."

Although there is little acknowledgement of where the work going on in the UK prior to this period, some of which was mentioned in Section 2, it is undoubtedly true that 1982/3 was the watershed in the acceptability of Stated Preference techniques in transport, and that the work of

Louviere, Hensher and Woodworth at that time had the greatest intellectual rigour. Regarding design issues, Louviere and Woodworth (1983) write (p352) that:

"Optimum designs for estimating the parameters of stochastic multiattribute choice models are not obvious and little work has been directed towards this problem because the optimum design requires *a priori* knowledge of the "true" choice probabilities. Additionally, most individuals working in the choice area are interested primarily in modelling issues and only secondarily in design issues. We believe discussion of design issues is important, particularly if such a discussion can lead to improvements in data collection and analysis"

Having drawn attention to the usual impracticability of using what we now term a full factorial design, in which each level of each attribute is combined with each level of every other attribute they propose using a fractional factorial design:

> "A fractional factorial design that enables one to estimate all two-way interactions probably would permit a fairly strong rejection test because if the MNL (Mutlinominal Logit) model is true, all of the interactions between the alternatives should be nonsignificant (because of the Independence from Irrelevant Alternatives (IIA) Property)".

On the basis of experimental results, which they report in an appendix, they conclude:

> "Thus, we suggest that for both practical and academic applications, the orthogonal, main effects, fractional factorial design plans are an efficient choice for the design of discrete choice or resource allocation studies. If rejection tests for MNL or other models are required, one could use design plans which allow as many relevant two-way interactions as possible to be tested".

Some empirical studies are discussed, but no guidance is given as to how to determine efficient attribute levels for the design, and this issue is not really treated as important. The same can be said of Hensher's 1994 Survey of Stated Preference practice (Hensher, 1994).

3.3 Concerns over whether designed-in orthogonality persists into the estimation data

Although much care is often taken in ensuring that Stated Preference designs are orthogonal, much less attention is devoted to ensuring that this orthogonality is not lost before the estimation stage, which is where it is really needed. Hensher (1994) summarised the position:

> "Hensher and Barnard (1990) have made a distinction between design-data orthogonality (DDO) and estimation-data orthogonality (EDO) in order to highlight that DDO is not always preserved in model estimation. This is very important for the most common procedure in travel behaviour modelling of estimating an MNL model with three or more alternatives on the individual response data, namely pooling all data (i.e. number of individuals in the sample by number of stated choice replications per individual) across the sampled population, but *not* aggregating the response data within a sampled individual. Estimation orthogonality using individual data and discrete choice models requires that the *differences* in attribute levels be orthogonal, not the absolute levels. Techniques such as MNL estimated on individual data require the differencing on the attributes to be *the chosen minus each and every non-chosen.* Since the chosen

alternative is not known prior to design development, it is not possible to design an experiment which has DDO, and which also satisfies EDO (Hensher and Barnard 1990).

> "The innovative method proposed by Louviere (1988) for overcoming EDO is not feasible where individual data are applied in estimation. The Louviere method defines a base alternative and derives all attribute combinations from a given *difference of attribute levels* satisfying an orthogonal-difference design. It is however suitable when the choice responses are aggregated within each individual's set of replications to derive choice proportions for each alternative. In this case, logit regression is a suitable estimation method, which does not require any further differencing in estimation. Transportation modellers have tended to opt for the preservation of the individual discrete-choice responses, and hence (without realising it in most cases), accepting some amount of correlation."

3.4 The value of Orthogonality

To what extent should designers of SP experiments be worried about departure from orthognality? This issue has been of concern to researchers in Britain, but appears to have been ignored elsewhere. However, possibly due to worries of the type just discussed in 3.3 above, there are signs of change. Hensher (1994), for example, writes

> "One of the important issues in statistical design is *orthogonality,* which ensures that the attributes presented to individuals are varied independently from one another. This property of zero-correlation between attributes enables the analyst to undertake tests of the statistical contribution of main effects and interactions, and is promoted as a major appeal of SP data compared to RP data. There is a view that although this is a desirable property, it is not a necessary condition for useful SP modelling. RP modellers have had to live with some amount of correlation, and have suitable tests for multicollinearity to identify when correlation is a problem. Mason and Perreault (1991) show in a cross-sectional context that fears about the harmful effects of collinear attributes often are exaggerated. Indeed the major benefit of SP methods is the ability to capture the response to diverse attribute combinations which are not observed in the market. One suspects that this is the dominating reason for the popularity of SP methods in transportation".

Mason and Perreault are in fact quite clear on the basis of their work that:

> "Collinearity per se is of less concern than is often implied in the literature; however the problems of insufficient power are more serious than most researchers recognise. Hence, at a broader conceptional level our results demonstrate that issues of collinearity should not be viewed in isolation, but rather in the broader context of the power of the overall analysis".

Steckel, Desarbo and Mkajan (1990) consider departures form orthogonality for the purpose of adding realism to the experiment, but seek to minimise the level of introduced correlation subject to specified restrictions regarding realistic attributes levels.

Green and Srinivasan (1990) confirm that "the presence of interattribute correlations per se does not violate any of the assumptions of conjoint analysis", but go on to use the analogy of multiple regression analysis to claim that interattribute correlation "increases the error in estimating preference parameters" and so advise that "interattribute correlations should be kept to a minimum (but they need not be zero)".

Johnson, Mayer and Ghoose (1989) suggest that negative correlations among the attributes (e.g. where each alternative has some attributes better than average and some worse then average) can reduce predictive ability for noncompensatory models (e.g. conjunctive and disjunctive models).

If researchers know of no other design issues other than that of maintaining orthogonality, then it is little surprise if great store is set on maintaining it. Once other issues arise, such as the inclusion of a good range of boundary values (discussed in Section 3.6 below) in a attempt to reduce the standard errors of estimated parameters, and so increase the power of any tests, then there can be a trade-off against orthogonality. Fowkes and Wardman (1988) comment as follows:

> "Most SP experiments have been based on full or fractional factorial designs using orthogonal arrays; thus the attributes are independently distributed. Catalogues of such designs are readily available (Cochran and Cox, 1957, and Kocur *et al.*, 1982). This approach makes it possible to avoid problems of multicollinearity, such as may be encountered with market place data. However, the disbenefit of having some correlation between the attributes need not be large, and may easily be outweighed by the disbenefits of slavishly adhering to orthogonal designs. We set out below advice on five important aspects of SP experiments, which on occasion will conflict with the desire for orthogonality".

These aspects were:

- (i) Realistic attribute levels
- (ii) Plausible combination of attribute levels
- (iii) Avoiding small variations in attribute levels
- (iv) Incorporating a good range of boundary values
- (v) Building in the ability to test the rationality and integrity of SP responses.

In order to appreciate the weight of the argument here we must turn to consider some of the development of the ideas underlying the above.

3.5 The UK Department of Transport 'Value of Time' Project and UK Science and Engineering Research Council 'Business Travel' Project.

The U.K. Department of Transport 'Value of Time' project began in 1980, funded partly by the British Railways Board and undertaken by a consortium consisting of The MVA Consultancy; the Institute for Transport Studies, University of Leeds; and the Transport Studies Unit, University of Oxford. The final report was published as MVA/ITS/TSU(1987). The key findings were reported to the 1986 PTRC Summer Annual Meeting, some being included in the Proceedings as Bradley, Marks and Wardman (1986).

Phase I of the project was envisaged as a review and identification of areas needing further study. However (MVA/ITS/TSU, 1987):

> "During Phase I, it became clear that the traditional empirical source of value of time (using choice models based on 'Revealed Preference' (RP) data) was unlikely to be

satisfactory for resolving a number of the problems of interest. Chiefly, this is because of the relatively large sample sizes needed to achieve sufficient accuracy of estimates, and the difficulty and cost of obtaining such samples in many choice situations".

These statistical aspects were studied in the project's Working Paper 7, which is available as Gunn (1981). In that paper, Gunn included a section on 'efficient design' where he says:

> "The general principles of survey design and sample size assessment can be described in simple terms as follows. We suppose that a survey is to be conducted in which observations of a variable Y are to be made at N points corresponding to different values of a variable X - say X_1, X_2, .., X_N - and that a model is to be fitted in which Y is to be related to X by a relationship involving unknown parameters. Suppose that the distribution of Y, given X and the true values of the unknown parameters, is known. Then, in anticipation of the results of the survey/experiment, we can write down general forms for the estimators of the unknown parameters, and hence of the fitted model, and also write down general forms for the variance-covariance matrix of the parameters and thus also of function of these parameters, including the fitted model. If our intention is to maximise the accuracy of estimation of some function of the fitted parameters for given sample size, or to minimise sample size for some required accuracy of estimation, we can refer to these general forms to indicate the relationship between the amount of data, the location of this data and the consequent accuracy of the fitted parameters."

After giving some examples, Gunn turns to the criteria for an optional design:

> "An overview of the various approaches that have been taken to the design problem is given by Silvey (1980). In the specific context of disaggregate models, see also Daganzo (1979). In general, the solutions depend on the objective. .....we have considered the problem of maximising the accuracy of a single parameter. However, the same approach could be used for any general function of parameters, provided that it is single valued. This does raise difficulties when there is no 'natural' choice of such a function. According to Silvey, the most commonly adopted (or at least, for theoretical exposition, most frequently postulated) is the 'criterion of D-optimality' which amounts to minimising the determinant of the variance-covariance matrix of the parameters in the model. This objective is equivalent to minimising the area of any given confidence region for the parameters, thus in some sense maximising the joint accuracy of the parameter estimates.
>
> "Figures [3.1] and [3.2] illustrate this concept with reference to a logit model framed in terms of two parameters, a and b. The model is taken from Bates et al (1978) and refers to the proportion of households owning at least one car as a function of gross household income. Figure [3.1] shows the data and the fitted model. Figure [3.2] shows the 95% confidence region associated with the estimated parameters, using the maximum likelihood estimators, for the given data set. One use of the 'optimum design' approach would be to determine 'where' (i.e. at which income points) the data should be collected for populations with similar expected relationships between car ownership and income in order to minimise the error of the fitted parameters as descried by the size of the corresponding confidence regions.

"The solution to this problem is given by Silvey, quoting from Ford (1976); for any given sample size, half the observations should be taken at one income point, and half at another. The points are given by a general formula; in the case of the relationship described these turn out to be approximately £15 and £62 for the 1972 data, I_1 and I_2 in figure [3.1]. Once again we see how crucial is the assumption that the model is correct! However, this sort of information does provide valuable insights into the relative values of taking observations at different points, providing we are reasonably cautions about the policies it advocates.

"Inference about 'values-of-time' has usually involved models with a particular form of parameter structure; typically, there has been a function relating 'utility' to observed variables by an expression such as

$$U_i = (\theta_0{}^i - \theta_1 M_i - \theta_2 T_i - \theta_3 Z_i) \qquad [3.1]$$

in which Z is some variable like comfort, 'M' denotes a money cost and 'T' denotes time in an actively, and the θ's are constants.

"In certain cases, there may be advantages in interpreting the fitted coefficient of cost variables in probabilistic choice models based on random utility theory with the dispersion parameter Λ, which is inversely related to the standard deviation of the random component of the utility function (the effect of the 'unobservables'). This corresponds to a choice of money units for the utility expression.

For logit models the relationship is $\Omega = \dfrac{\Pi}{\sqrt{6}} \; \dfrac{1}{\sigma}$

where σ denotes the standard deviation of the random component.

"Adopting this convention, we can write the general form of the linear utility function which is used in many empirical studies as

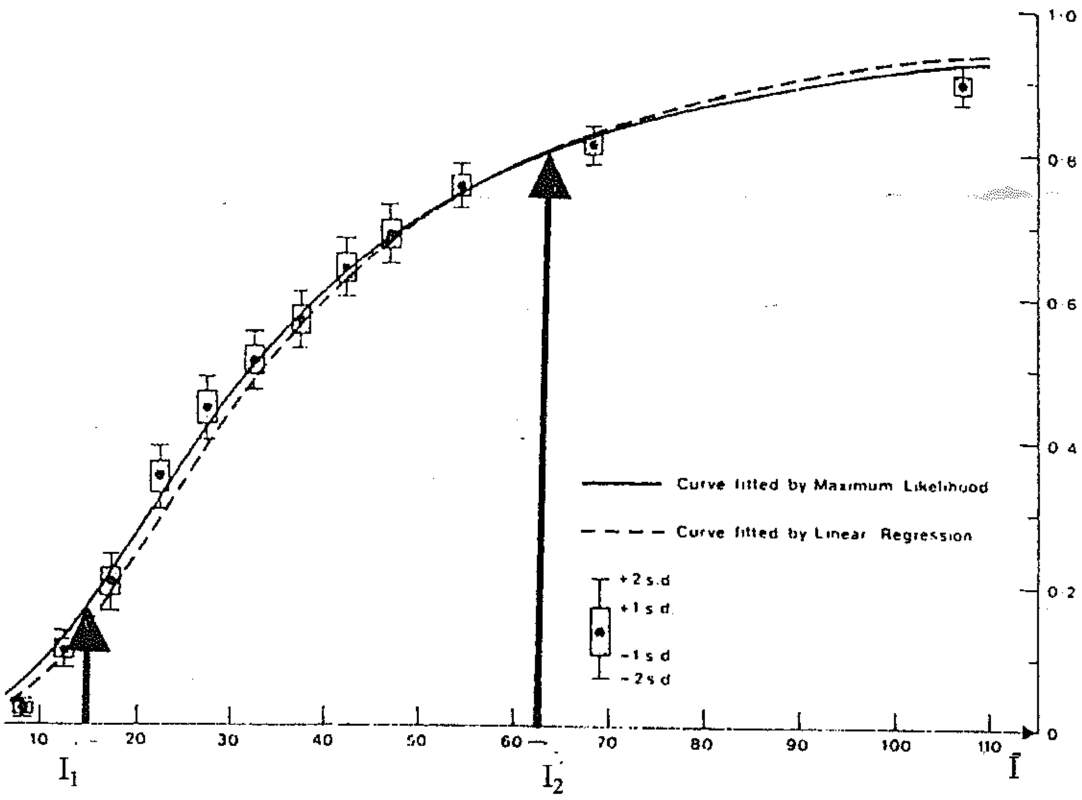
$$U_i^j = \theta_0^i - (\Omega)M_i^j - (\Omega V_T)T_i^j - (\Omega V_Z)Z_i^j + \varepsilon_i^j \qquad [3.2]$$

where θ_0^i refers to the mean of the 'unobservables', V_T to the value of time and V_Z to the value of some other variable, all now measured in money terms. ε is assumed Weibull, standard deviation 1, for 'logit' models, ('i' refers to option, 'j' to individual.)

"Returning to the notation of equation 5. In this case, there is a single function of parameters that is of paramount importance, namely the ratio θ_2/θ_1, the 'value-of-time' if circumstances are appropriate, the accuracy with which a particular design estimates this ratio forms a natural criterion of optimality.

Fig. 3.1

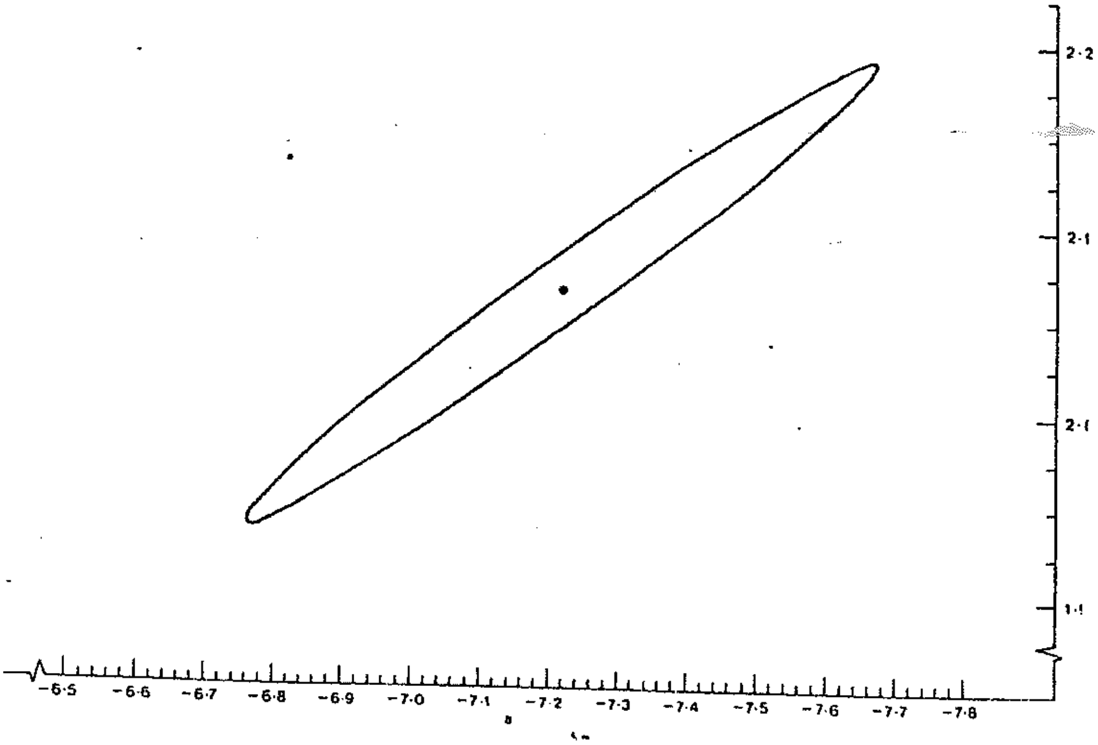
1972 FES DATA FOR P_{1+} SHOWING CONFIDENCE INTERVALS AND FITTED CURVES



Household income in £'s per week

Fig. 3.2

P_{1+} RELATIONSHIP FOR 1972 FES: 95% CONFIDENCE REGION FOR $(\hat{a}, \hat{b})$



"The Taylor series approximation for the variance of a function gives

$$\mathrm{Var}(\underline{f}(\underline{x})) \approx \left[\frac{\partial f}{\partial \underline{x}}\right]^T \mathrm{Var}(\underline{x}) \left[\frac{\partial f}{\partial \underline{x}}\right]$$

Thus if we have a general expression for the variance-covariance matrix of the fitted parameters we can approximate the variance of a function of the parameters. If the estimates have been derived as likelihood maximising solutions, such an estimate is provided by the inverse of the expectation of the matrix of second derivatives of the log-likelihood function. In the case of the hypothetical example given above, ...the criterion to be minimised is

$$VR = \left[\frac{1}{\theta_1^2} V_{22} - \frac{2\theta_2}{\theta_1^3} V_{12} + \frac{\theta_2^2}{\theta_1^4} V_{11}\right] \qquad [3.3]$$

where the θ coefficients are those fitted in a model of the form

$$\Delta U = \theta_0 - \theta_1 \Delta M - \theta_2 \Delta T - \theta_3 \Delta Z + \varepsilon$$

and V_{ij} is the covariance of θ_i and θ_j.

"To make the VR expression useful for design purposes, we need approximations to the V_{ij}, which will in general be functions of sample location (in terms of (ΔM, ΔT,ΔZ)) and sample size as well as of the unknown coefficients $\underline{\theta}$. A suitable approximation to the variance covariance matrix of the fitted coefficients for (aggregate) logit models is given in Gunn and Whittaker (1981) for the case of Poisson errors. A similar approximation for the Binomial case (we shall assume a binary choice and $\theta_0 = \theta_3 = 0$ here) is as follows:

Write $m = \Delta M$, $t = \Delta T$

Define $\left.\begin{aligned} \overline{m} &= \Sigma_i W_i m_i / \Sigma_i W_i \\ \bar{t} &= \Sigma_i W_i m_i / \Sigma_i W_i \end{aligned}\right\}$ $i = 1,.., n$, the no. of points at which observations are taken

where $W_i = n_i\, p_i\, (1-p_i)$ $\qquad$ n_i = no. of observations taken at point i, defined by (m_i,t_i)

and $p_i = [1 + \exp(-\theta_1 m_i - \theta_2 t_i)]^{-1}$ $\qquad$ (assuming $\theta_0 = \theta_3 = 0$ for illustration)

$$\text{set} \quad \begin{aligned} V(m,m) &= \Sigma_i W_i (m_i - \overline{m})^2 \\ V(t,t) &= \Sigma_i W_i (t_i - \bar{t})^2 \\ V(m,t) &= V(t,m) = \Sigma_i W_i (m_i - \overline{m})\,(t_i - \bar{t}) \end{aligned}$$

Gunn comments as follows:

"Equation [3.3] allows us to say a number of things about the conditions necessary for 'value of time' measurement, as well as providing actual quantitative information about accuracy for any proposed design (i.e. selection of points at which to experiment), and determining the relative trade-off between the number and location of the experimental points and the survey effort to apportion to each.

"Firstly, we can see that the larger is (θ_1^2), the more accurate our measurement (other things being equal). Having identified (θ_1) as being inversely related to the standard deviation of the random component of the utility function, in money terms, we can interpret this as saying that conditions in which the 'representative' component (i.e. that which is made explicit) dominate the total utility expression will be most favourable for accurate value-of-time measurement. In other words, where the model explains little of the variation, measurement will be poor.

"Secondly, it is easy to see that the term $[V(m,m)V(t,t)-(V(m,t))^2]$ will be zero if M and T are linearly related. In other words, in such conditions VR would be infinite: no measurement is possible if 'time' and 'cost' are perfectly correlated, and the less they are correlated the better.

With this notation, an approximation to the variance-covariance matrix of the fitted θ_1 and θ_2 coefficients is

$$\mathrm{var}\,(\theta_1,\theta_2)=\frac{1}{\left[V(m,m)V(t,t)-(V(m,t))^2\right]}\begin{bmatrix} V(t,t) & -V(m,t) \\ -V(m,t) & V(m,m) \end{bmatrix} \quad [3.5]$$

We can now write

$$VR=\frac{1}{\theta_1^2\left[V(m,m)V(t,t)-(V(m,t))^2\right]}\left[V(m,m)+2\left[\frac{\theta_2}{\theta_1}\right]V(m,t)+\left[\frac{\theta_2}{\theta_1}\right]^2 V(t,t)\right] \quad [3.6]$$

Note that $\dfrac{\theta_2}{\theta_1}$ is our estimate of the value of time.

"Equation [3.6] allows us to say a number of things about the conditions necessary for 'value of time' measurement, as well as providing actual quantitative information about accuracy for any proposed design (i.e. selection of points at which to experiment), and determining the relative trade-off between the number and location of the experimental points and the survey effort to apportion to each.

"Firstly, we can see that the larger is θ_1^2, the more accurate our measurement (other things being equal). Having identified θ_1 as being inversely related to the standard deviation of the random component of the utility function, in money terms, we can interpret this as saying that conditions in which the 'representative' component (i.e. that which is made explicit) dominate the total utility expression will be most favourable for accurate value-of-time measurement. In other words, where the model explains little of the variation, measurement will be poor.

"Secondly, it is easy to see that the term [V(m , m) V (t , t) – (V (m , t))^2] will be zero if M and T are linearly related. In other words, in such conditions VR would be infinite: no measurement is possible if 'time' and 'cost' are perfectly correlated, and the less they are correlated the better.

"Thirdly, we can see in general terms that VR contains a term linear in the V(.,.) divided by one quadratic in the V(.,.). Broadly speaking, accuracy will come from maximising the V(.,.). From the definitions of the V terms we can see that such a maximum will occur as a compromise between two opposing trends: terms such as $(m_i - \bar{m})$ and $(t_i - \bar{t})$ will suggest placing the experimental points as far apart as possible to maximise the expression; however, at extreme points the W_i will tend to zero (p_i will tend to zero or unity, so p_i (1-p_i) will tend to zero) and so a compromise will occur. (The one dimensional example given above produced a solution roughly at the points of inflection, and this may generalise. Using eq (3.6) together with symmetry arguments should lead to a straightforward, if tedious, solution for the optimal design in the general case.)

"Finally, we can see that the optimal design/accuracy of measurement depend on the level of the value-of-time. It is more sensible to consider the ratio of the standard error of measurement of the vot to its absolute level in this case: $\dfrac{\sqrt{VR}}{\theta_2 / \theta_1} = RSE$

We obtain

$$RSE = \sqrt{\frac{1}{\theta_2^2\left[V(m,m)V(t,t) - V(m,t)^2\right]}\left\{\left[\frac{\theta_1}{\theta_2}\right]^2 V(m,m) + \left[\frac{\theta_1}{\theta_2}\right] 2V(m,t) + V(t,t)\right\}} \quad [3.7]$$

For very small values of time, the expression in curly brackets is dominated by V(m,m), whereas for large values of time the V(t,t) expression dominates. Different design strategies will be appropriate for different values"

This work by Gunn has been very influential in our own thinking, and so we quote it extensively here so as to properly acknowledge the debt. Equation 3.3 can be confirmed by reference to standard statistical texts (e.g. Kendall and Stuart, 1969). Gunn's advice was primarily intended as guidance for sampling strategies in Revealed Preference studies, and he arguably did all that was required in that regard. However, the same ideas apply to the design of SP experiments, where the designer can choose the M and T values offered to respondents, and the combinations in which they appear. In section 3.10 we shall take up the story more then a decade later, when equation 3.6 was (effectively) differentiated with respect to quantities that could be controlled in the design, in order to find their optimal settings. Gunn's feeling seemed to be, based on the times he mentions it, that the data should be placed around the points of inflection of the logit function, but the mathematical support for this is weak. It was quite clear to Gunn, however, that all the data should not be clustered very close to the p = 0.5 point, nor be for very high or low p values.

Phase II, of the UK DOT Value of Time Study, was delayed till mid-1983 and was a pilot phase. Since Stated Preference techniques had not previously been much used in this country, a validation of SP against Revealed Preference techniques on the same respondents was proposed.

Business travel was excluded from the Value of Time Study, but a separate parallel study was funded by the Economic and Social Research Council at this time, and conducted by ITS Leeds.

The first SP experiment to be designed for the Value of Time study was for North Kent Commuters, and it was proposed initially to follow the methods adopted by the MVA Consultancy for work in the Netherlands (Bates and Roberts, 1983). However, at the meeting called to determine the design to be used, the present author proposed that candidate designs should be tested to check that respondents with differing values of time would be expected to answer in different ways, thereby betraying their value of time, to a greater or lesser degree, with any design failing this test being rejected and an improved design sought. This proposal was adopted. The mode choice experimental design was composed of 5 difference attributes, each at a high and a low level. Bates (1984) takes up the story:

> "The next problem is to assign values to the high and low points for each variable. This turns out to be quite a difficult task, since two potentially conflicting criteria must be satisfied. In the first place, the values (here, the difference in attributes for rail and coach) must be realistic, so that they lie reasonably within the range of the respondents' experience. In the second place, the alternatives offered must not be such that one mode will dominate another for the majority of respondents; if, for instance, given a chosen set of values for the eight alternatives, the great majority of respondents choose rail in every case, very little information will be provided by the survey.
>
> "In order to circumvent this problem, we first came up with a range of cost and time differences which seemed reasonable on the basis of the level of service currently offered by the two modes in the North Kent corridor. We then used the standard generalised cost formulation, in which walking and waiting times were valued at double the in-vehicle rate, and calculated the difference in generalised cost for each alternative on the basis of a range of assumed values of time, from 0.5 to 8.0 pence/min. The cost and time differences were then adjusted within the sensible ranges so that no one mode dominated at any of the chosen values of time. In fact, since coach tends to be a slower but less expensive mode, the effect was that at low values of time, most (but not all) alternatives suggested a preference for coach, while the opposite was true at high values.
>
> "In this way, the following sets of values were chosen for the five variables:

Variable	Low	High
DC	40p	100p
DWK	2 mins	6 mins
DWT	3 mins	10 mins
DIVT	-25 mins	-40 mins
DOTH	5 mins	15 mins

> "Figure [3.3] shows the range of values in the costs and in-vehicle time plane, together with the implicit values of time. The modal choice at each of these points will be determined according to the individual's value of time. If there is no modal bias, then the line of equal utility goes through the origin. Because of the way the differences are expressed (Rail - Coach), points to the right and above the line of equal utility will tend

to be coach-preferred, and conversely. The steeper the line of equal utility, the higher the value of time.

"A number of observations can be made about this figure. Two possible lines of equal utility are shown in the figure - illustrating a value of time of 8 p/min, and other of 0.5 p/min. When no other time attributes are taken into account, it appears that persons which a value of time of 8 p/min will choose rail for all replications, while persons with a value of 0.5 p/min will choose coach for all replications. Since replications 5 and 7 should normally be chosen in preference to replications 2 and 4, the greatest amount of information comes from the relative preference between replications 6 and 8 and replications 1 and 3. Indifference between these two sets will imply a value of time of 4 p/min."

The SP experiment described above, and conducted in Summer 1983, maintained orthogonality whilst paying some attention to the other points mentioned. The employee survey, described by Fowkes, for the Business Travel project abandoned orthogonality altogether. The context had been thought too difficult for SP, but in the event the survey, conducted in the first half of 1984, was unexpectedity successful. Some description of the design issues is contained in Marks and Fowkes (1986) and reproduced here:

"It was hoped that respondents would answer the ranking exercise by trading differences in cost against differences in time away from home, the inconvenience of start times and any other perceived differences between the services offered by the 4 modes. The experiment was designed by setting the start times and total journey times, which together determined the finish times. Levels of the time and cost variables were chosen so that the data would identify a reasonably wide range of time valuations. An orthogonal design was not considered possible because of the constraints imposed by the following two 'real life' considerations:

i) Travel times by first and second class rail should be equal, unless we were to complicate the analysis by having frequent first class only trains.

ii) The cost of first class rail should be about 50% greater than the cost of second class rail, as is usually the case during the business peak. As in (i) we wished to keep our hypothetical options as close as possible to travellers' actual experiences.

"In order to ensure the experiment could identify a wide range of values of time; 'iso-utility' or boundary values of time were calculated for each modal comparison. An iso-utility or boundary value of time is the value of time at which an individual would be indifferent between a given pair and modes. Table [3.1] contains the 'iso-utility' values of time for the experiment calculated assuming the utility derived from modal attributes other than cost and time is zero. The table shows that there is a wide range of boundary values. The intention was to allow for a wide range of in-vehicle values of time, together with a wide range of variability in valuations of factors other than cost and time. The effect of these others factors is captured by Alternative Specific Constants (ASCs) included in model calibrations, where they represent the utility gain (or loss) of, say, flying as opposed to travelling by first class rail, assuming the costs and times are identical for both modes.

"Attribute values were primarily chosen so that choices between air, and first and second class rail covered a wide range of boundary values of time. Travel by car was not expected to be a serious option for most respondents because of the length of the hypothetical journey. Any aversion to the use of car means the boundary values given in columns 1, 4 and 5 of Table [3.1] are biased upwards."

Fig. 3.3

Experimental design - pink cards; costs vs in-vehicle time
(for the RED cards, each point is shifted 15 mins to the left)

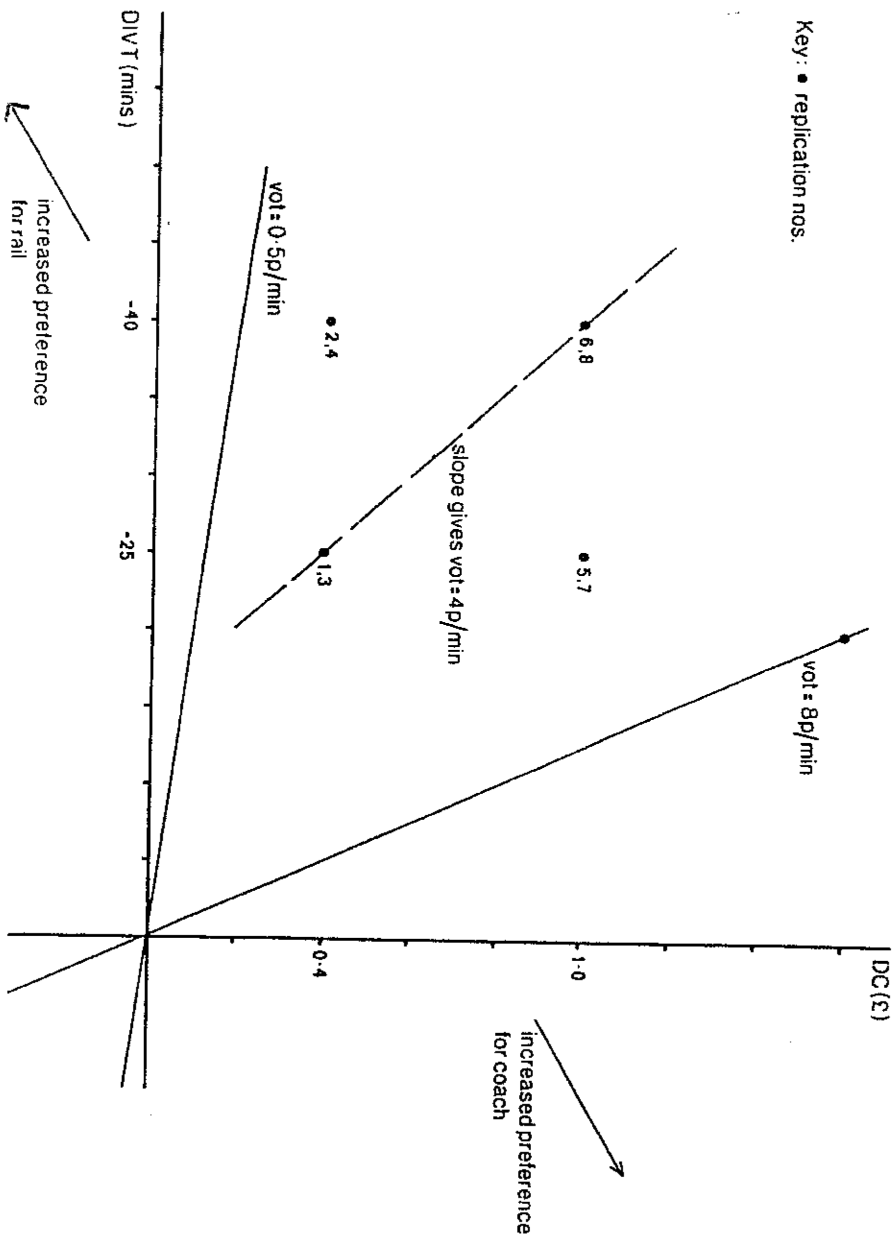


Table 3.1 : Boundary values of time (£/hr)* from SERC 'Business Travel' project, employee survey (Marks and Fowkes, 1986)

Question No.	Air vs car	Air vs 1st class rail	Air vs 2nd class rail	1st class rail vs car	2nd class rail vs car
1	11.4	2.5	15.0	23.3	6.7
2	14.3	15.0	17.5	20.0	40.0
3	15.0	80.0	110.0	2.0	4.0
4	17.1	40.0	60.0	8.0	0.0
5	10.0	-15.0	2.5	43.3	20.0
6	11.3	-8.8	1.3		
7	13.3	20.0	40.0	10.0	0.0
8	15.0	-17.5	5.0	47.5	25.0
9	13.8	1.3	8.8		
10	11.1	15.0	17.5	-20.0	-40.0
11	13.3	5.0	30.0	17.5	5.0
12	13.8	10.0	22.5	17.5	5.0

* Negative values of time occur whenever the choice is dominated by one mode i.e. the cheaper option is the faster. Positive infinite values of time occur whenever there is no difference in travel times and so; all else being equal; one would chose the cheaper mode.

Bates and Roberts (1986) summarised the experience of SP gained during the value of Time Project:

> "As well as deriving values of time from observed travel choices (the traditional "Revealed Preference" approach), the study has placed considerable reliance on "Stated Preference" data, where respondents are presented with hypothetical situations and asked to assess them in various ways: the reasons for doing this have been discussed elsewhere (see for example Bates (1983)). The experience of this approach built up during the study has been generally favourable: the results have been comparable with those obtained by the more traditional approach, and the variations in values of time obtained from the analysis have accorded well with expectations.
>
> "In the course of the study, six major surveys were carried out and analysed, covering most combinations of purpose, mode and context (urban vs inter-urban)..........

"In all cases, analysis consisted of relating the response variable to the kinds of utility formulations discussed ...... earlier ......, making use of the segmentation approach. Leaving aside, the question of Transfer Price, which has been previously discussed (Broom et al. (1983), Gunn (1984)), there were three kinds of response variable:

a. RP choice vector ((0,1) variable)

This applied to the West Yorkshire, North Kent and Tyne Crossing studies, where we collected information about the actual choice made by respondents.

b. SP 'rating' response

This was used in the North Kent, Tyne Crossing and Long Distance Coach/Rail surveys. Each individual is presented with a series of "replications", and on each replication he has to choose between two alternatives according to a five-point semantic scale. As was described in Bates (1984), the responses are transformed into an estimate of the utility difference between the two alternatives, which can then be regressed on suitable formulations for the functional form of the utility difference.

c. SP 'ranking' response

This was used in the Urban Bus and Long Distance Car surveys. Each individual is presented with a series of alternative journeys and asked to rank them in order of preference. This set of ranks is then "exploded" into a sequence of discrete choices in the manner described by e.g. Chapman & Staelin (1982).

"More details about the actual variables used in each of the four studies carried out within the final Phase of the project are given in Bradley, Marks & Wardman (1986)."

3.6 Boundary Values and Taste Variation

Some of the SP designs used in the Value of Time Study were non-orthogonal, and most paid careful attention to the implied 'boundary' values of time, which determined the response given the respondent's actual value of time. We have seen that the employee SP design in the SERC Business Travel project also incorporated these principles. However, these principles were not yet widely accepted, and ways of implementing them were still being gradually improved. In this sub-section we will set out some of the main underlying theory and look at some early applications.

One important point which must always be considered is that the assumption of the logit model is that all respondents have identical parameters (i.e. identical tastes) and that the error term just reflects what might be termed 'day to day' variability, possibly due to omitted variables. Of course, in practice the error term can cover some part of variation in parameters between individuals, but the theory says it should not, and econometric difficulties could arise, principally because of the independent variables being correlated with the error term, causing heteroskedasticity.

If it were possible to get enough data for each individual to permit models to be calibrated to that individual, then the problem would not arise, although there might then be discussion of how best to combine individuals to provide aggregate values. The approach in transport has been, instead, to calibrate logit models on groups of people: possibly sub-samples split before the SP is administered, perhaps on the basis of a screening process; possibly split after the survey on the basis of reported characterises, such as age, sex, income etc.; or split within the modelling by means of allowing different parameters on the basis of such characteristics. This last method was known as 'segmentation' during the Value of Time study. The advantage of the first method, pre-screening, is that a customised SP experiment can be presented to each group, e.g.. long-distance travellers can be given choices involving long distance journeys, and short distance travellers given choices involving short distance journeys.

In order to consider the problem more fully, we might set out the theoretical position as follows, where RU is random utility, X_{ijm} are attribute levels observed for individual i, attribute j and mode m and $\varepsilon 1_{mk}$ and $\varepsilon 2_{mi}$ are two components of error. The first error term depends only on the individual and the second varies over every choice (k). The term Random Utility is used to emphasise that the underlying truth is considered random, rather than being an unknown function which has to be estimated using a statistical technique involving error terms.

$$RU_{imk} = \sum_j \beta_{ijm} X_{ijmk} + \varepsilon 1_{mk} + \varepsilon 2_{mi} \tag{3.8}$$

with of the β being the parameters to be estimated, and k the observation number.

At its simplest, normal practice is to assume that the parameters do not vary from person to person, (i.e. using a dot to indicate dimensions in which there is no variation):

$$\beta_{ijm} = \beta_{\bullet jm} \qquad \text{for all i} \tag{3.9}$$

and that the second error term does not exist, i.e.. there is no variability from person to person, just from observation (choice) to observation:

$$\varepsilon 2_{mi} = 0 \qquad \text{for all i} \tag{3.10}$$

This last point is important as it lies at the heart of a debate on the determination of standard errors, and consequently the calculation of t ratios and the conduct of significance tests. The usual assumption is that all observations are subject to the same random error, regardless of which respondent was responsible for that choice. It is usual to report t ratios on that basis. It has sometimes been suggested (e.g. Louviere and Woodworth, 1983) that choices made by the same respondent may be in some way similar, and so should not be treated as independent observations counting as one extra degree of freedom. At its extreme, this argument is clearly nonsense, since the number of respondents is clearly not the sample size. Estimates can be derived for individual respondents, proving that their error terms do vary from choice to choice. Current best practice is discussed in Ouwersloot and Rietveld (1996).

Returning to the coefficients, if the parameters were thought to vary as a function of some variable, e.g. income, then exploratory modelling might pick this up. For example, suppose (Fowkes and Wardman, 1988) that parameter β_{ijm} was a linear function of an arbitrary variable Z, with no taste variation in the other coefficients i.e.:

$$\beta_{ijm} = \begin{cases} \alpha_{om} + \alpha_{1m} Z_{im} & \text{for } j = 1 \\ \beta_{\bullet jm} & \text{else} \end{cases} \tag{3.11}$$

Then

$$RU_{imk} = \alpha_{om} X_{i1mk} + \alpha_{1m} Z_{imk} X_{i1mk} + \varepsilon 1_{mk} + \varepsilon 2_{mi} \tag{3.12}$$

which could successfully be modelled with an X_1Z interaction term.

Alternatively, particularly if Z is only known for each respondent from a tick-box question, i.e. grouped, dummy variables, DZ_g, can be set up such that

$$DZ_g = 1 \qquad \text{if } Z = g, \qquad \text{else zero} \tag{3.13}$$

whence interaction dummy variables XDZ_g can be formed. This is the method referred to as 'segmentation' earlier.

For any of the above methods to work, respondents must have been posed a sufficiently 'broad' set of SP questions, so as to provide trade-offs involving relative valuations covering the full range occurring in the population. In Fowkes and Wardman (1988), following Fowkes (1985), we referred to these trade- off values as 'Boundary or equi-utility values'. Thankfully, it is the former name that has stuck. In that paper we set out the case as follows (quotation marks omitted for clarity):

> The process of choosing the precise trade-offs to be presented in the SP experiment is both important and non-trivial, particularly given the assumed presence of inter-personal taste variations. The prime objective is to offer choices that will permit model parameters to be determined accurately.
>
> The choice of the particular attribute values must take into account the relative valuations at which individuals would be indifferent between options. In order to achieve a satisfactory design, the set of these boundary or equi-utility relative valuations should cover a reasonable range of potential variation in taste and uncertainty as to the likely average value. We might then think of each choice as deciding on which side of the implied boundary point an individual lies, so that the set of such choices presented to an individual will place him in one of the ranges between adjacent boundary values. In recent SP applications in transport the importance of these considerations does not seem to have been fully perceived.
>
> Consider the model:
>
> $$U_m = \beta_c COST_m + \beta_x X_m \tag{3.14}$$
>
> where COST is the monetary cost, say in pence, and X is some attribute of mode m. Let us now define a boundary relative valuation of X in terms of money as
>
> $$B(X{:}COST) = (COST_1 - COST_2)/(X_2 - X_1) \tag{3.15}$$
>
> In the absence of random effects, an individual whose value of X is greater in money terms than B(X:COST) will prefer the alternative with the largest amount of X, and vice versa. With a random error, if all respondents have monetary values of X equal to B(X:COST), we should find 50 per cent of respondents choosing each option.
>
> In order to obtain an accurate estimate of the respondent's relative valuation, we must present sufficient boundary values to make the inter-boundary value distance acceptably small. It will usually be thought desirable to have boundary values closer together where we are expecting to find actual values. This will not 'force' these values to be returned by the estimation, but will imply a lesser accuracy for values more sparsely covered.
>
> Often, we will have a third attribute varying, say Y, that is,

$$U_m = \beta_c COST_m + \beta_x X_m + \beta_y Y_m \tag{3.16}$$

One way of proceeding is to assume that the parameter of this attribute is some multiple of one of the other parameters. For instance, the value of walking time has conventionally been assumed in UK practice to be twice the value of in-vehicle time.

Suppose we decide to assume that

$$\beta_y = k\beta_x \tag{3.17}$$

so that we feel able to choose for test purposes low, medium and high values for the unknown k. Generalising (3.15), we have

$$B(X{:}COST) = (COST_1 - COST_2)/[k(Y_2 - Y_1) + (X_2 - X_1)] \tag{3.18}$$

Thus the boundary values are now a function of the factor k. We require our design to be satisfactory in the range of potential values of k, so we inspect the set of boundary values in turn for k low, medium and high. Some 'peculiar' boundary values will no doubt now result, but this is not important, provided that, for each of the three values of k in turn, there are sufficient, and sufficiently well spaced, boundary values to "cover" individuals' likely monetary valuations of X. A simple computer program will make checking this an easy task, but much trial and error may be required before acceptable values are found.

To take the above example further, we will usually wish to derive estimates of the relative monetary valuations of both X and Y, and so we could repeat the above exercise with X replaced by Y and vice versa, checking the boundary values B(Y:COST). However, it may be preferable merely to check the boundary values for k, thus ensuring that this ratio can be adequately recovered by the estimation process.

This discussion has assumed that we have restricted our specification of the hypothetical choices to the presentation of attribute levels, and that all else is assumed equal for each alternative. We may, however, wish to differentiate between different 'sorts' of alternatives; respondents will then be expected to react to the attribute levels presented in the context of their past experience and prejudices concerning each sort of alternative. Statistically, we then need to use Alternative Specific Constants (ASC) in the model in order to allow for a general preference for some sorts of alternatives rather than others, in situations where the attribute levels for each alternative are identical.

We can illustrate this point by considering the distinction between what we have termed 'within mode' studies and 'between mode' studies. In the 'within mode' case we will present respondents with descriptions of different journeys by, say, bus in terms of various attributes (time taken, fare, etc.) and ask which they prefer. They will have no reason to prefer one alternative to another, except on the basis of the attribute levels given. Thus there will be no justification for including ASCs in the model. In a 'between mode' study, on the other hand, we might describe alternatives by the same attributes, but additionally specify that one alternative is, say, train, and the other, bus. For all attributes for which levels are not specified (for example, comfort), respondents will take into account their own perception of the different modes. Since it may not now be true that 'all else is equal' for the two alternatives, an ASC should be included.

Mathematically, with an ASC present, we have, for alternatives 1 and 2:

$$\begin{aligned} U_1 &= ASC + \beta_c COST + \beta_x X_1 + \beta_y Y_1 \\ U_2 &= \qquad\quad \beta_c COST + \beta_x X_2 + \beta_y Y_2 \end{aligned} \tag{3.19}$$

where ASC is the 'coefficient' of a 0-1 variable taking the value one for alternative 1 and zero for alternative 2. Boundary values can now be derived as:

$$B(X{:}COST) = \frac{COST_1 - COST_2 + (ASC/\beta_c)}{k(Y_2 - Y_1) + (X_2 - X_1)} \tag{3.20}$$

$$B(k) = \frac{[\beta_c (COST_1 - COST_2)/\beta_x + (ASC/\beta_x) + (X_1 - X_2)]}{(Y_2 - Y_1)} \tag{3.21}$$

In (3.20) the term (ASC/β_c) is merely the ASC expressed in monetary units, while in (3.21) the term (ASC/β_x) is the ASC expressed in the units of X. Since the ASC will not be known in advance, separate boundary values should be constructed to encompass all likely values.

3.7 British Rail InterCity East Coast Main Line Overcrowding and Departure Time Preference Study

Much of the above methodology was adopted in a 1987 study for British Rail in conjunction with PhD research at ITS, Leeds. Due to commercial confidentiality restrictions, the work was not fully published at the time. What follows is largely taken directly from (the cleared for publication sections of) Wardman and Fowkes (1987).

The trade-offs amongst attributes can be seen in Table 3.2 which presents the experimental design. In all cases except scenarios 11 to 13, the trade-off is between two variables only with the remaining variable constant across options. As an example, the boundary values of FULL in relation to BAY (scenarios 1 to 3) are: 50 pence, £1, and £2. We would ideally like to include more boundary values for this and the other attributes but are constrained by the total number of comparisons included in the design. The task required of the respondent is simplified given that generally only two attributes vary across options.

Although we have moved away from zero correlations between attributes, it would be unwise to have very high correlations. Testing the design using synthetic data will indicate, amongst other things, whether there are problems for the accuracy with which the relative values can be estimated as a result of any collinearity present in the design.

3.7.1 Testing the Experimental Design

Fowkes and Wardman (1988) also recommend the use of synthetic data and computer simulations to test the adequacy of the design in terms of whether it allows the accurate estimation of known underlying utility weights. Such a procedure is of use even if the implied boundary values have been carefully taken into account and it was used in developing our design before conducting the survey.

Table 3.2 : The Experimental Design

	OPTION A			OPTION B		
SCENARIO	FARE	SEATING	DEPTIME	FARE	SEATING	DEPTIME
1	Actual	Full	Actual	+£0.50	Bay	Actual
2	-£1.00	Full	Actual	Actual	Bay	Actual
3	Actual	Full	Actual	+£2.00	Bay	Actual
4	-£0.50	Stand 30	Actual	Actual	Bay	Actual
5	Actual	Stand 30	Actual	+£1.00	Bay	Actual
6	-£2.00	Stand 30	Actual	Actual	Bay	Actual
7	Actual	Stand 30	Actual	+£5.00	Bay	Actual
8	Actual	Stand 60	Actual	+£2.00	Full	Actual
9	Actual	Stand 60	Actual	+£5.00	Full	Actual
10	Actual	Stand 60	Actual	+£10.00	Full	Actual
11	Actual	Stand 60	1 Hour	+£5.00	Bay	Actual
12	Actual	Stand 60	1 Hour	+£10.00	Bay	Actual
13	Actual	Stand 60	1 Hour	+£20.00	Bay	Actual
14	Actual	Full	2 Hours	+£1.00	Full	Actual
15	Actual	Full	2 Hours	+£5.00	Full	Actual
16	Actual	Full	2 Hours	+£40.00	Full	Actual

Notes : DEPTIME is specified as changes to the actual journey and 'Actual' denotes that the variable is the same as for the actual reference journey.

The tests of the design in Table 3.2 were based on a synthetic data set of 1600 observations, that is the number of observations which would be generated by 100 respondents. Choice is deemed to be based on random utility (RU) which, for alternative i, could be expressed as:

$$RU_i = \alpha_{1i}X_{1i} + \alpha_{2i}X_{2i} + \ldots\ldots + \alpha_{ni}X_{ni} + \varepsilon_i \qquad (3.22)$$

The simulation process assigns individuals to that option with highest random utility, that is the choice data is taken to be discrete. The deterministic component of random utility ($\Sigma\alpha_{ki}X_{ki}$) comprises the part-worth utilities associated with each attribute whilst the stochastic element (ε_i) represents unobservable and unmeasurable effects. Since we can control for the variables influencing choice, the errors in an actual SP experiment can be taken to represent any discrepancy, for whatever reason, between stated and actual preferences. The error term also contains any inter-personal taste variation, that is where the α_{ki} vary across individuals, which is not explicitly allowed for at the modelling stage in practice.

The aim is to test the adequacy of the design across a wide range of underlying utility weights. This is done by comparing actual and estimated relative values for a number of different values of interest. The actual relative values are constant across individuals in any particular comparison rather than explicitly introducing taste variation into the estimation process. The latter would make the interpretation of the results of the tests more difficult. The effect of taste variation at the estimation stage is a separate issue. Although the tests are based on discrete choice data, and our SP experiment does not restrict the analysis of the data to this form since a five point categorical response scale is used, the form of the response scale is largely immaterial with regard to the extent to which the SP experiment accurately serves its purpose.

Given the discrete choice between the two alternatives, calibration of the model provides estimates, in the form of scale transformations, of the utility weights of equation (3.22). The test process then examines whether the actual relative valuations can be accurately recovered from the synthetic data.

The indirect utility function was specified in difference form and includes dummy variables to represent FULL, STAND30 and STAND60. These are estimated in relation to the omitted category of BAY. The latter is assigned a zero value. In the actual experiment, departure time variations are to be analysed in terms of their effect on the amounts of early and late time incurred by an individual, that is the discrepancies between ideal and actual departure times which may involve having to depart earlier or later than desired. The tests reported here implicitly assume that everyone is initially departing at their ideal time and that variations from the ideal time have constant unit value.

The results of the tests are reported in Table 3.3. The BLOGIT program of the Australian Road Research Board (Crittle and Johnson 1980) was used to calibrate the discrete choice logit model. The left hand columns of Table 3.3 list the relative values used in each test and the right hand columns contain the estimates derived from the synthetic data.

Although the standard errors associated with the estimated relative valuations are not given, the estimated values are sufficiently close to the actual values across a wide range for use to conclude that the experimental design performs well and is capable of accurately revealing individuals' underlying preference structures from their responses to the hypothetical trade-offs presented. The error component of random utility has been kept relatively small so that any large discrepancy between estimated and actual values can be taken to result from a shortcoming in the design rather than a large range in which the estimated value can lie.

3.7.2 Stated Preference Models

Table 3.4 contains the results of an SP model which was calibrated on the 598 individuals who have purchased a standard single/return, day return, executive or saver ticket, who were making journeys of more than one hour and who had supplied information on their actual and desired departure times.

The model specifies dummy variables to represent the effects of FULL, STAND30 and STAND60. These are estimated in relation to the omitted category of BAY. The latter is omitted to avoid perfect collinearity and a singular matrix. FULL represents the amount of time spent in a full train, that is the dummy variable is multiplied by the individual's actual journey time, since the disutility of a full train can be expected to vary according to the length of time spent on a full train.

EARLY represents the amount of early time involved, that is a departure before the desired time. Similarly, LATE specifies the amount of time after the desired departure time that the journey was or would be made. Although a departure time variation which moves the individual closer to the desired departure time will generally be beneficial, any amount of early or late time incurs

Table 3.3 : Tests of the Experimental Design (Relative Values)

ACTUAL				ESTIMATED			
DEPTIME	FULL	STAND30	STAND60	DEPTIME	FULL	STAND30	STAND60
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 DEPTIME is given in pence per minute. The values of the other variables are specified in pence. Note, however, that values for FULL as pence per minute are subsequently presented.							

disutility. The estimated values of early and late time represent the disutility in money terms of having to depart one minute earlier/later than desired.

If option A requires a one hour earlier departure than for the actual journey made, and the individual travelled 30 minutes earlier than desired, the early time associated with option A is 90 minutes. Given that option B does not vary the departure time, it would involve the actual amount of early time of 30 minutes. Late time would be zero for both options. If instead option A introduced a departure time one hour later than for the actual journey, option A now incurs 30 minutes late time whilst option B, which has not changed, involves the 30 minutes early time of the actual journey. Where there are no departure time variations, each option will contain the amount of early/late time applicable for the actual journey and the difference between the two options is clearly zero. Where the individual travelled at the desired time, the effect on early/late time of departure time variations is exactly as specified in the experiment.

The coefficients are made generic, that is they have the same value for both options, since there is no reason why the coefficients should vary by option in this exercise. The variables can therefore be specified in difference form in this binary choice context. Although the value of standing can be non-linear in the model specified, FULL, EARLY and LATE have constant unit values. Subsequent analysis justified the latter by showing that the value of time spent in a full train varied little according to the amount of travel time and that the values of EARLY/LATE were insensitive to the amount of early/late time involved.

Table 3.4 : An Overall Stated Preference Model - Includes £40 Scenario

VALUE OF EARLY (p/min)	11.21		(15.43)
VALUE OF LATE (p/min)	11.57		(15.84)
VALUE OF FULL (p/min)	1.71		(10.18)
VALUE OF STAND 30min	£13.64		(19.31)
VALUE OF STAND 60min	£12.43		(27.24)
RHO BAR SQUARED		0.23	
OBSERVATIONS		7412	
OPTION A		2227	
OPTION B		5185	

Notes : t ratios are given in brackets. The t ratios are not adjusted to allow for repeat observations per person.

The coefficients in Table 3.4 are all of the correct sign, denoting that more of any of the variables reduces utility, and they are all highly significant. The values derived appear somewhat high. This is because the sample contains individuals who have not paid for the train fare themselves, particularly business people travelling in the course of their work. Those who have not fully accounted for cost in the choice between travel options, because they do not pay for the fare themselves, will have relatively low cost coefficients and hence high relative values.

The most noticeable feature of the results presented in Table 3.4 is that the value of STAND 30 exceeds that of STAND60. This is clearly implausible. On inspection of the responses supplied by individuals, the problem was traced to scenario 16 of Table 1, where some 10% of respondents stated that they would be willing to pay £40 to avoid departing 2 hours earlier/later. Since there are no boundary values of a two hour departure time variation between £5 and £40, this was working through to influence the derived values and, in particular, to cause the estimated value of STAND30 to exceed that of STAND60. It was suspected that those willing to pay £40 to avoid departing 2 hours earlier/later would not be paying for the rail ticket themselves. However, when those who did not pay for their own ticket were omitted, 5% were still apparently prepared to pay £40 in this situation. The calibrated model for these individuals also had a value of STAND30 which exceeded that for STAND60.

Following detailed analytical work that located the fault at scenario 16, this scenario was omitted for all individuals. This reduced the data set to 7092 observations and the results of a model calibrated on this data are given in Table 3.5. Despite the fewer observations in the model, the t ratios of the estimates are higher. This suggests that error has been removed from the model although the choices are no better explained since there is no change in Rho Bar Squared.

Table 3.5 : An Overall Stated Preference Model - Omit £40 Scenario

VALUE OF EARLY (p/min)	7.62		(16.28)
VALUE OF LATE (p/min)	8.57		(17.31)
VALUE OF FULL (p/min)	1.52		(12.99)
VALUE OF STAND 30min	£9.93		(22.39)
VALUE OF STAND 60min	£11.43		(36.95)
RHO BAR SQUARED		0.23	
OBSERVATIONS		7092	
OPTION A		1938	
OPTION B		5154	

Notes : t ratios are given in brackets. The t ratios are not adjusted to allow for repeat observations per person.

The value of STAND60 now exceeds that for STAND30 whilst the other values, particularly EARLY and LATE, are somewhat changed. It would appear that the value of STAND30 in Table 3.4 is too high, rather than the value of STAND60 being too low. This is confirmed by some analysis which was conducted to examine this issue further and to determine whether the estimates in Table 3.5 could be taken to be reliable. This involved a simulation exercise, the results of which are presented in Table 3.6.

In the previous simulation exercise reported in Table 3.3, very few (and in many cases no) individuals paid £40 to avoid a 2 hour change in departure time. Thus the problem apparent here never arose. Table 3.6 gives three models based on synthetic data. The first row lists the actual values which the simulation process is required to recover.

Table 3.6 : Simulation of the Problem

	TIME	FULL	STAND30	STAND60
ACTUAL VALUES	6.00p/min	£3.00	£10.00	£12.00
A) ESTIMATED VALUES	6.24p/min	£2.78	£10.14	£11.83
B) 10% PAY £40	8.33p/min	£4.04	£17.30	£13.06
C) OMIT £40 SCENARIO	5.61p/min	£2.94	£10.25	£12.27

Model A reports the results of a simulation along the lines of those previously conducted. It can be seen that the calibrated model has little difficulty in accurately recovering the actual values. Model B simulates the problem under consideration by amending the choices of 10% of individuals so that they would pay £40 in the relevant scenario. This could be interpreted as an error in the SP responses or as taste variations whereby this amount would actually be paid. The effect on the values derived is clear and, as in the comparison of the models presented in Tables 3.4 and 3.5, the estimated value of STAND 30 exceeds that of STAND60. The estimated values in model B are all somewhat different to the actual values. Model C omits the scenario which contains the £40 cost increase. It can be seen that the problems otherwise apparent are avoided and that the actual values can still be accurately estimated.

It is not possible to identify whether the responses of those who state that they are prepared to pay £40 to avoid departing 2 hours earlier/later are subject to serious error or reflect an actual willingness to pay such an amount. However, it seems sensible to omit the scenario containing the £40 cost increase because the responses to it influence the value of STAND30 in a drastic fashion, because they seem to affect the estimated values of variables other than departure time variations to which they relate and because the experimental design still performs well when this scenario is omitted. That scenario is consequently of no use to the calibration and would best have been excluded at the design stage.

3.7.3 Bin Analysis

By designing the 16 scenarios such that in 15 cases either seating condition or departure time was constant in both options, and by providing a range of boundary values, it became possible to conduct what Fowkes (1991a) has termed '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' i.e.. those below the lower boundary values, 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."

Table 3.7 shows an example of a bin analysis, relating back to Table 3.2 and concentrating just on the valuation of STAND30 versus BAY (i.e. Stand for 30 minutes and then have a bay shared by at most one other person, as against being in that latter situation for the whole time). The sample is all 515 respondents, and not the screened data set used for the logit modelling.

Table 3.7 : Example of Bin Analysis - Willingness to pay to avoid standing 30 minutes. East Coast Main Line travellers, 1987.

WILLINGNESS TO PAY	%	DIFFERENCE Δ%	MID-PT W	VALUE WΔ%
Less than 50p	4.8	4.8%	25p	£0.01
Less than £1	13.5	8.7%	75p	£0.06
Less than £2	15.6	2.1%	£1.50	£0.03
Less than £5	41.0	35.4%	£3	£1.06
Less than infinity	100.0	59.0%	say £8	£4.72
			Weighted average	£5.88

Table 3.7 illustrates the transparency of the method. The actual estimate is somewhat arbitrary due to uncertainty over which monetary value to attribute to each bin. Mid-points have been used, except for the final bin which has no upper limit. This is because we had not expected such high valuations and therefore included no fixed boundary value higher than £5. It will be noted that only 41% said that they were not willing to pay £5 (i.e. for scenario 7 in Table 3.2, 41% chose A and 59% chose B). Clearly the indicated willingness to pay is on average greater than £5. In fact, after excluding 'funnies' and short distance trips, the logit model reported in Table 3.5 gave £9.93 as the value of STAND30. After disaggregating by class of travel, the value of STAND30 for the majority group (second class leisure, pay self) come out at £6.00. Hence we can conclude that the median valuation from the logit analysis is around £6.00, just as the bin analysis gave.

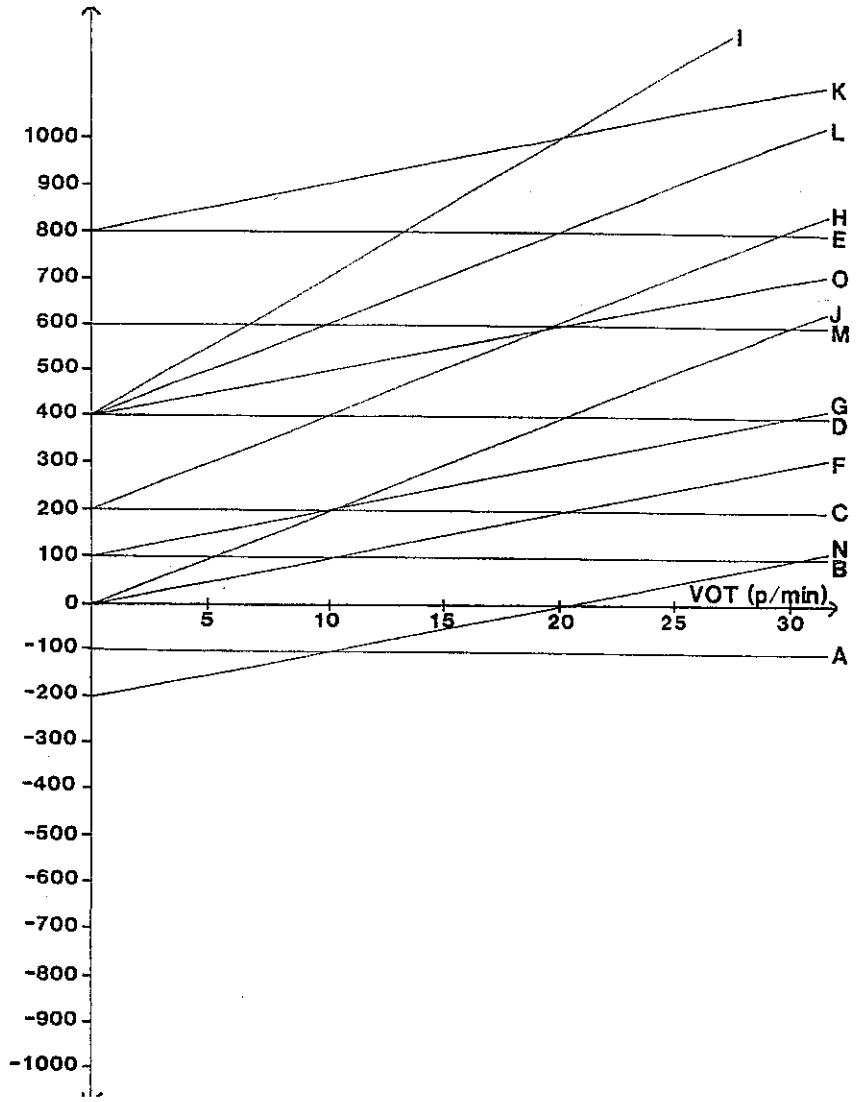
The results of this study were accepted by both BR and DoT, and formed the basis of a successful investment submission for one extra coach to be added to all East Coast Main Line Electra trains. The DoT also encouraged SP evidence in support of bids for grant and towards the opening of local railway stations and services at this time (Fowkes and Preston, 1991).

3.8 British Rail Network South East Service Quality and Provincial Overcrowding Studies

Following the success of the work reported in 3.7 above, the paper by Fowkes and Wardman (1988) was circulated by the Network South East sector of British Rail to tenderers for a Service Quality study. The MVA Consultancy teamed with ITS, Leeds to win this work. A literature review (Fowkes, 1988) was produced to consolidate and refine the new design methodology. That paper contains the first printed use of the term 'bin analysis'. It also standardises terminology as boundary values (as opposed to equi-utility values). Perhaps more importantly it contains the first printed reference to boundary value 'rays' and presents a 'boundary value map' (reproduced here as Figure 3.4), which is now often referred to as a boundary ray map or diagram. The latter has clear ancestry in Figure 3.3.

Fig. 3.4

Boundary values of travel centre improvement (pence)



An example of how desired boundary values might be incorporated into an experimental design was given in an appendix (to Fowkes, 1988) and is reproduced here. The context is an experiment designed to determine travellers' values of time and willingness to pay for an improved travel centre. The travel centre attribute entered the design at two levels: existing (E) and improved (I). The value of the improvement was not expected to be more than about 10% of the ticket price, which was never more than £50 for the (first class) travellers involved. Being first class, their value of time might be expected to be quite high. The East Coast Main Line study had found first class business travellers value of 'adjustment time' (i.e. travelling earlier or later than desired due to the lack of a train at the ideal time) of 14p/min, and the pure value of time would be expected to be a little lower than this.

The task, therefore, is to design in boundary values for the travel centre improvement which could determine (to within, say, 50p) the willingness to pay, for respondents whose values of time (for the sorts of journey in question) lay somewhere between £5/hr and £25/hr. The advice given by Fowkes (1988) is as follows:

> "To being with we would like horizontal boundary value lines at U = -£1, £2, £4 and £8. These lines are coded A to E in Table [3.8], where the cost difference is shown in pence, and a higher cost for the improvement is shown as a negative cost difference. The time differences are all zero, so that boundary values are fixed for all values of VOT.
>
> Secondly, we might wish to add 'cross bracing', and boundary value lines F to I in Table [3.8] seem suitable.
>
> Thirdly, we may wish some 'rays' through the origin. Boundary line F already fulfils this function, but only reaches 200 by VOT = 20. Hence line J would appear valuable as an addition.
>
> In practice it may simply not be possible to meet all one's desires for boundary value lines without having an excessively large number of cards. Some element of compromise will be needed, and gradual improvement attained. This process has already been done for Table [3.8], such that it is known that all the boundary value lines there listed can be obtained with a reasonably small number of cards.
>
> We may start our design by deciding the range for attribute levels. Service is either E for existing or I for improved. Cost must have an £8 difference for boundary line E, so we can choose a range such as £38 - £46. Boundary line I requires a 30 minute time difference, so we can choose a range such as 80-110 minutes. Although the argument that follows would be somewhat more general if we worked in differences, it should be easier to follow if we use absolute values in these suggested ranges. Converting to differences is straightforward.

Table 3.8 : Desired Boundary Value Lines

CODE	COST DIFFERENCE	TIME DIFFERENCE
A	+100	0
B	-100	0
C	-200	0
D	-400	0
E	-800	0
F	0	-10
G	-100	-10
H	-200	-20
I	-400	-30
J	0	-20

Boundary line E requires cards with values:

Card 1	£38	x mins	E
Card 2	£46	x mins	I

where x is as yet undetermined.

We can get boundary line B by using card 1 with:

Card 3	£39	x mins	I

and similarly for boundary line D

Card 4	£42	x mins	I

Now working 'backwards' from card 4 we can get boundary line C with:

Card 5	£40	x mins	E

As a bi-product we now already have boundary line A from cards 3 and 5. This completes the 'horizontal' boundary lines A to E.

Boundary line I requires us to use the full range of journey time, i.e.. 80 to 110 minutes. We are going to require some 10 and 20 minute differences, so this suggests that we choose x to be either 90 or 100. Let us choose 90 and take stock. We have cards:

Card 1	£38	90 mins	E
Card 2	£46	90 mins	I
Card 3	£39	90 mins	I
Card 4	£42	90 mins	I
Card 5	£40	90 mins	E

We can proceed to obtain line F from card 1 with the addition of:

Card 6 £38 100 mins I

and similarly we can obtain line J from card 6 with the addition of:

Card 7 £38 80 mins E.

Now that we have a card with 80 minutes on we can get line I with:

Card 8 £42 110 mins I.

This leaves lines G and H to be achieved. However, careful choice of the values for eight cards so far designed has given these boundary lines as bi-products:

Cards 3and 7 give line G
Cards 5 and 8 give line H.

The full set of eight cards designed to give the boundary lines A to J are then:

CARD	COST	TIME	TRAVEL CENTRE
C1	£38	90 mins	E
C2	£46	90 mins	I
C3	£39	90 mins	I
C4	£42	90 mins	I
C5	£40	90 mins	E
C6	£38	100 mins	I
C7	£38	80 mins	E
C8	£42	110 mins	I

The full set of choices involving a valuation of the travel centre improvement is set out in Table 3.9. Note that lines K to O are additional to those given as desired boundary lines in Table 3.8. Some may be of little use, but all seem to add something, and lines M and K definitely make a big improvement to the design. Figure 3.4 gives the boundary value map for this design."

Table 3.9 : Derivation of Boundary Value Lines

CHOICE	COST DIFFERENCE	TIME DIFFERENCE	LINE
C1 v C2	-800	0	E
C1 v C3	-100	0	B
C1 v C4	-400	0	D
C1 v C6	0	-10	F
C1 v C8	-400	-20	L
C5 v C2	-600	0	M
C5 v C3	+100	0	A
C5 v C4	-200	0	C
C5 v C6	+200	-10	N
C5 v C8	-200	-20	H
C7 v C2	-800	-10	K
C7 v C3	-100	-10	G
C7 v C4	-400	-10	O
C7 v C6	0	-20	J
C7 v C8	-400	-30	I

Following the difficulty experienced with the '£40' scenario in the East Coast Main Line work reported above, it was thought wise to reimpose orthogonality for the NSE designs. This tempered the freedom available for the above example, but since the area was now so much better understood, it proved possible to prepare orthogonal designs incorporating a good range of boundary values and capable, during simulation tests, of recovering an even wider range of assumed relative valuations. Due to the scale of the exercise required by NSE, a ranking approach was adopted. Since an interviewer was to be present when the exercise was completed, we did not expect the exercise to be too daunting for respondents. In fact, three separate ranking exercises were used with each respondent. A total of 8 ranking designs were produced, to allow for short and long distance and, in the case of the crowding design only, whether the journey was a commute or for leisure. For 'cleanliness' the design had just 9 cards to be ranked, whereas most designs were 10 cards and for 'crowding long distance commute' we had to resort to a 12 card design. As SP experiments go, this was one of the largest. A large amount of qualitative work underlay the designs.

While the NSE work was continuing, the Provincial sector of British Rail commissioned a study to value willingness to pay to avoid overcrowding on (a range of) their services. At this time there was much expenditure on new stock for Provincial, but the only justification being accepted for this was one of cost reduction by replacing obsolete stock. The new stock was attracting additional custom in some cases but extra stock to cater for this could not be justified without 'willingness to pay to avoid overcrowding' values acceptable to the DoT. The MVA Consultancy again led the project and the "stated Preference design..... was set up and tested by staff of the Institute of [sic] Transport Studies of the University of Leeds" (MVA and ITS, 1989). This was the first study where a full blown 'bins' analysis was presented (in the final Report; MVA and ITS, 1989) to the client. This method also allowed the data to be cleaned prior to logit analysis, by removing respondents who were markedly inconsistent in their replies. Each respondent was given 5 cards printed on a sheet with each card containing 4 options to be ranked. This self completion experiment yielded immensely rich data which supported analysis not previously dreamt of.

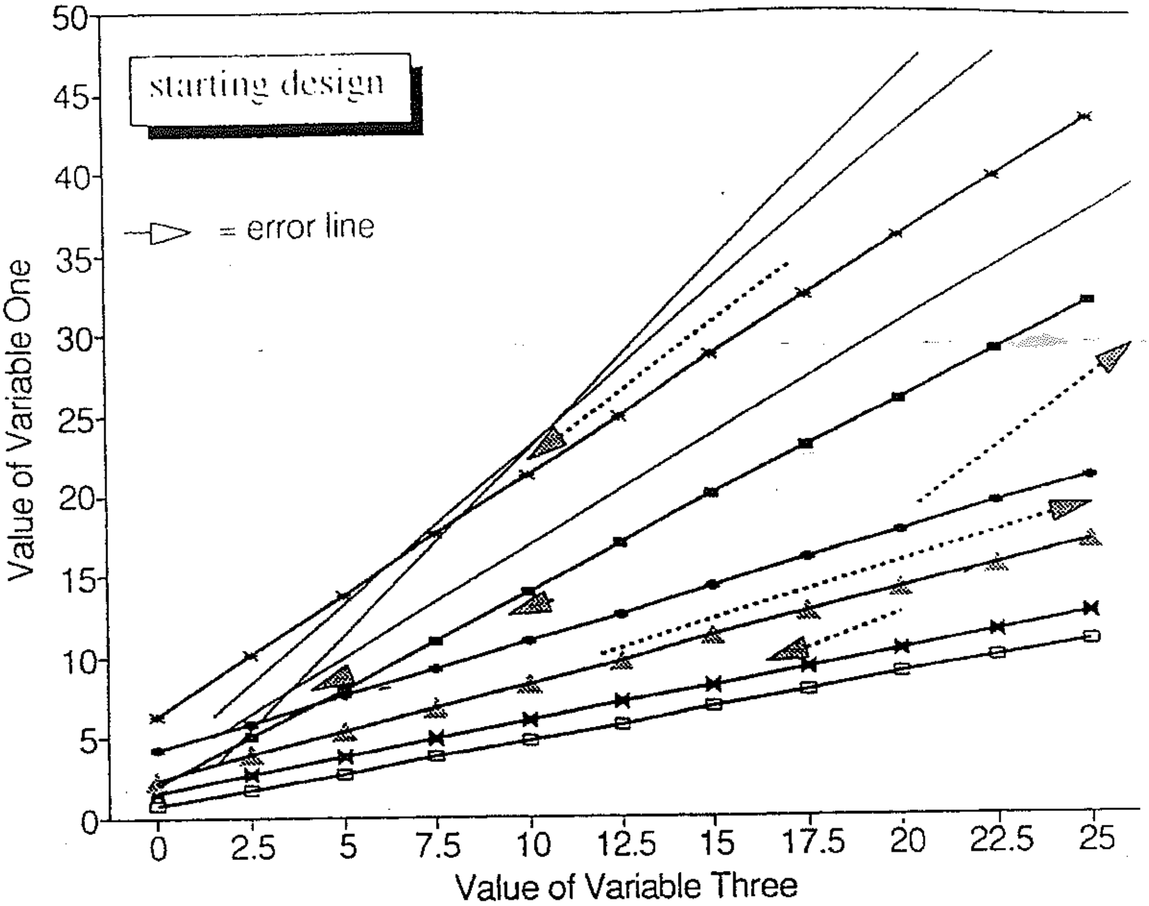
3.9 Examination of the Influence of Boundary Rays

From experience and intuition, it seemed obvious that providing a dense mesh of boundary rays around 'interesting' combinations of relative valuations was desirable. In order to test this more rigorously, Holden, Fowkes and Wardman, (1992) carried out simulations of three radically different designs. Of particular interest was design one, reproduced here as Figure 3.5a. When plotted on a boundary value map, this design was seen to consist solely of upward sloping boundary rays, with few intersections. Viewed either vertically or horizontally, there appeared to be a good spread of boundary values. However, simulation showed that the design had little control diagonally. This is another way of saying that (large) positive errors in one value were associated with (large) positive errors in the other value (and vice versa). Error lines have been drawn onto Figure 3.5a, with their tails on the assumed (i.e. target) value, and their heads pointing to the estimated value. Holden et al concluded that some intersections of boundary rays (as in Figure 3.4) would be beneficial:

> "The clue to this lies in the evident tendency in the tabulated results for design 1 for large positive errors for VOT to be paired with large positive errors for VOC, and likewise for negative errors in each. Marking the target and estimated points on the map shows that the 'error lines' between them lie between the boundary value rays. Figure [3.5a] illustrates this for a later example. Hence for a large error, the only possibility is an upward sloping error line since all the boundary value rays are upward sloping. By considering the error lines for the other designs, we could see that they are constrained in length (i.e.. magnitude of error) by the 'barriers' formed by the criss-crossing boundary value rays. Indeed, we have no cases of an error line actually crossing a boundary value ray.
>
> "It is not our contention that error lines can never cross a boundary value ray. The simulated data we use is generated from models including an error term, and real-life responses will contain various sorts of error........
>
> "Nevertheless, to all intents and purposes, error lines do not appear to cross boundary value lines. Consequently, if a client wishes to know whether travellers using a given facility would be prepared to pay at least a given value for a particular improvement to that facility, then efforts should be made to ensure that all respondents face a boundary value close to that value. At the very least, the design will ensure that the proportion willing to pay that amount can be ascertained from the responses."

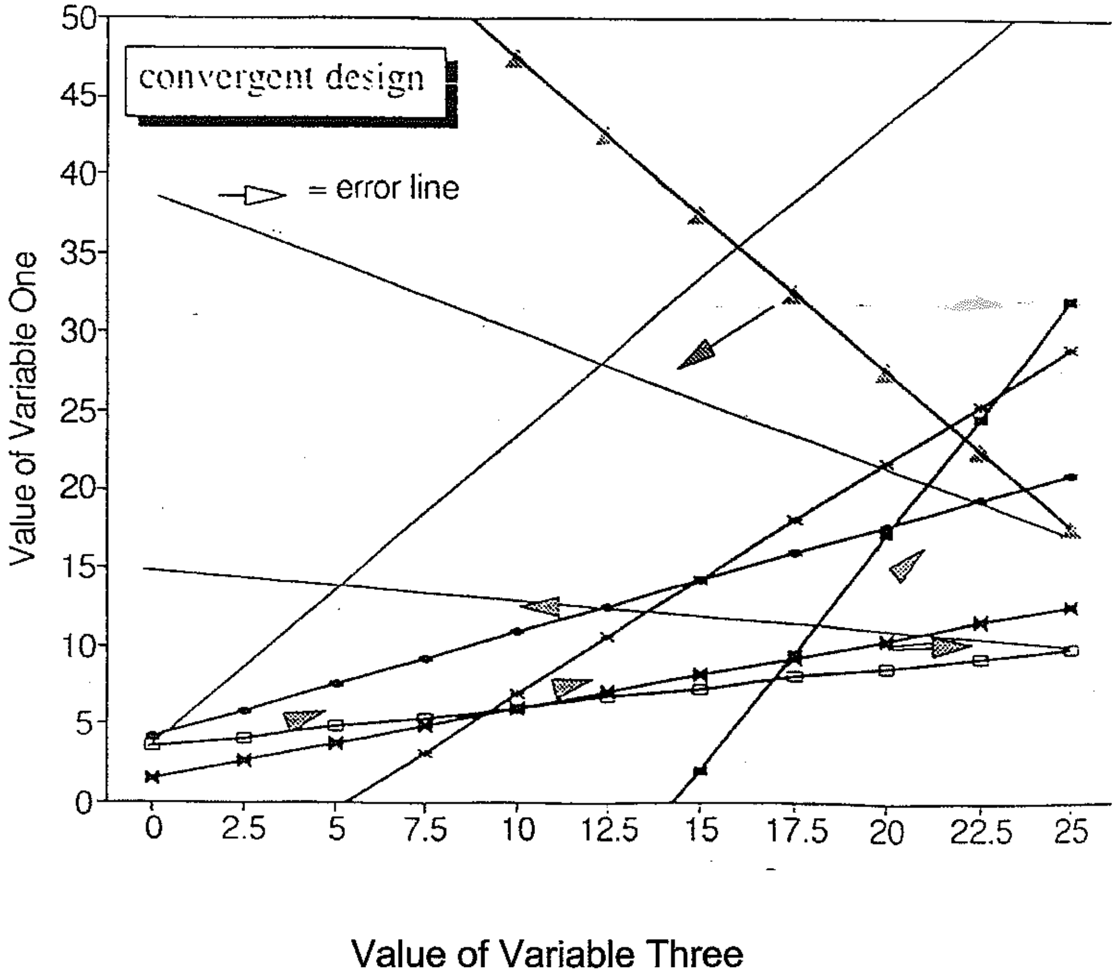
Holden et al. then proceed to devise an algorithm, which for a given range of target values, ranks the boundary rays in order of usefulness, and then replaces the worst with one chosen to 'hem-in' better the target values. This was very much experimental work rather than being a blue print for future designs, but showed considerable success. The method used was as follows:

Fig. 3.5a



starting design
= error line
Value of Variable One
50
45
40
35
30
25
20
15
10
5
0
0
2.5
5
7.5
10
12.5
15
17.5
20
22.5
25
Value of Variable Three

Fig. 3.5b


convergent design
= error line
Value of Variable One
50
45
40
35
30
25
20
15
10
5
0
0
2.5
5
7.5
10
12.5
15
17.5
20
22.5
25
Value of Variable Three

"The simulation routine used in the program takes a set of likely, or target, values and constructs a simulated sample. This sample is assumed to represent the profile of the real population. Again a utility function is used to mimic the way members of the population choose between the options. To take account of the 'unobservable' component in choice behaviour, a random number generator is used to construct a random term which proportionally adjusts the deterministic part. In this way a set of binary choice responses is collected, which can be summed and the logit of the proportion regressed on the differences in levels of variables in the design.

"The regression supplies parameter estimates along with the statistical measures, standard deviation and correlation. As the target values are expressed as a ratio to the second variable (assumed here to be the cost variable), a division is required before the estimates can be compared to the original targets. A percentage off-target figure is calculated. This is the measure which demonstrates how capable the design is of capturing particular targets. By ranking the targets in order of highest % off-target we can identify which targets require attention. The program takes the two worst estimated targets and stores the distance between the targets and their respective estimates (the length of the error line). Then it stores the mid-points of the error lines and calculates the implied gradient and intercept required to join these two midpoints. Our purpose is to reduce the area around the targets by placing a new ray across the two error lines, so that it excludes the areas into which the error lines currently veer.

"Moving to the second stage of the algorithm, it is clear that some options will produce rays that form a border of an area containing a target pair. If the ray was removed then the estimate could veer away from the target into the released area. What we term excluded area is a measure of the contribution the option is making to the efficient estimation of that target. A ray may well border more than one target area. We say that the option has a total exclusion equal to the sum of the excluded areas from regions containing targets. When each excluded area is determined, before it is added to the counter for the total exclusion of the ray, it is weighted according to the strength of feeling about the likelihood of it appearing in the population. This steers the algorithm away from removing a ray adjacent to an important target. This part of the code uses two dimensional geometry to measure the areas. For this to work, a border is required so that all areas are bounded. The borders represent the maximum and minimum values expected for the relative valuations. No area outside the borders is counted as excluded as we assume that no respondents posses values beyond this region, so a ray is not making a contribution by bordering regions outside. Again a ranking of each option's exclusion will indicate which option is to be rewritten.

"Therefore, by combining the rankings from stages one and two, we can rewrite the option contributing the least exclusion so that it excludes areas around the two poorest estimated targets. The required gradient and intercept of the new ray are decomposed into the necessary levels for inclusion in the new design. If this design returns estimates below a % off-target threshold then the scheme has converged, otherwise further iterations are required."

Figure 3.5b shows the modified version of Design One (in Figure 3.3a) after the application of the algorithm. Most of the error lines are now so short that the head of the arrow obscures the tail. In practice, we would not have such exactly known target values, and so the method would be more restricted in identifying and dealing with problem 'zones' in the map. Nevertheless, the underlying principle of enclosing areas of interest is well demonstrated.

Swansea Pearman and Holden (1993) report the development of software, called 'SPRAY' which "allows the user to draw a suitable map, enclosing areas that are likely to have a high count of respondents". Thereafter, they use an algorithm that they had developed to allocate respondents to areas and so impute their attribute valuations.

> "This is really an extension of the 'bin' analysis developed by Fowkes [1991a]. We have written software which takes the original design, and by scanning the choices made by each respondent will infer parameter valuation that best explains the choices made. Because the parameters are generated for each individual, we obtain useful information about distributions of values across the sample. However, the approach still permits analysis by the more usual probabilistic (i.e. logit) modelling approach, giving access to two complementary analytical methods."

3.10 Statistical Investigation of the potential advantages of non-orthogonality

In setting up an automatic algorithm to produce improved SP designs in the way discussed above, the question arose again of whether only fully orthogonal designs should be considered. A private communication from John Bates pointed to the theoretical work necessary, and this was followed up by Fowkes, Wardman and Holden (1993). They used a regression based approach so as to benefit from results readily available in textbooks. This meant that the dependent variable had to be continuous, rather than a discrete choice. This was nothing strange, even to disaggregate choice modelling, since respondents had often been asked to provide a rating (such as definitely choose A, probably choose A, no preference, probably choose B, definitely chose B) which had to be converted into a probability and subjected to the Berkson-Theil transformation. In aggregate work, where choices could be grouped, logits of relative frequencies of choices were anyway commonplace.

Fowkes et al worked in difference such that:

$$X_1 = \Delta COST \qquad X_2 = \Delta TIME \qquad Y = \Delta U$$

"We shall use lower case x's to denote deviations from the mean. We shall denote correlation between two variables i and j by r_{ij}, such that the correlation between time and cost is r_{12}. For an orthogonal design, of course, this would be zero, but this is <u>not</u> assumed here. Our model is now:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{3.23}$$

"Estimates of β_0, β_1 and β_2 can be derived by least squares as is shown in econometric texts such as Gujurati (1988)."

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma_\varepsilon^2}{\left(1 - r_{12}^2\right)\sum x_1^2} \tag{3.24}$$

$$\mathrm{Var}\left(\hat{\beta}_2\right) = \frac{\sigma_\varepsilon^2}{\left(1 - r_{12}^2\right)\sum x_2^2} \tag{3.25}$$

where σ^2_ε is the variance of ε. The fact that ε are assumed Weibull rather than Normal makes no material difference. The covariance between β_1 and β_2 is also given as:

$$\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)=\frac{-r_{12}\sigma_\varepsilon^2}{(1-r_{12}^2)\sqrt{\sum x_1^2}\sqrt{\sum x_2^2}} \tag{3.26}$$

The value of time, VOT, was defined as the ratio of the estimates of β_2 and β_1.

$$\mathrm{VOT}=\frac{\hat{\beta}_2}{\hat{\beta}_1} \tag{3.27}$$

They assumed that they were particularly interested in deriving an accurate estimate of VOT, and correspondingly less concerned in deriving an accurate estimate of the value of the alternative specific constant. Hence they wished to minimise the variance of VOT.

From various sources, a good approximate formula for the variance of the ratio of two estimates was known, so that in the case of VOT they had:

$$\mathrm{Var(VOT)}=\frac{\beta_2^2}{\beta_1^2}\left(\frac{\mathrm{Var}(\hat{\beta}_2)}{\beta_2^2}+\frac{\mathrm{Var}(\hat{\beta}_1)}{\beta_1^2}-\frac{2\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)}{\beta_1\beta_2}\right) \tag{3.28}$$

Note that this is equivalent to equation (3.3) above.

Substituting (3.24) to (3.26) into (3.28) gave:

$$\mathrm{Var(VOT)}=\frac{\beta_2^2\sigma_\varepsilon^2}{\beta_1^2(1-r_{12}^2)}\left(\frac{1}{\beta_1^2\sum x_1^2}+\frac{1}{\beta_2^2\sum x_2^2}+\frac{2r_{12}}{\beta_1\beta_2\sqrt{\sum x_1^2}\sqrt{\sum x_2^2}}\right) \tag{3.29}$$

which is equivalent to equation (3.6) above.

They then proceed as follows:

"The variance of VOT depends on the correlation between variables 1 and 2: the term outside the brackets increases as the correlation increases but this can be counteracted by a negative correlation coefficient operating to reduce the term within the brackets. To find the r_{12} which minimises the variance of VOT, other things equal, we differentiate equation (3.29) with respect to r_{12}

$$\frac{\partial \mathrm{Var(VOT)}}{\partial r_{12}}=\frac{\beta_2^2\sigma_\varepsilon^2}{\beta_1^2(1-r_{12}^2)^2}\left(\frac{2r_{12}}{\beta_1^2\sum x_1^2}+\frac{2r_{12}}{\beta_2^2\sum x_2^2}+\frac{2+2r_{12}^2}{\beta_1\beta_2\sqrt{\sum x_1^2}\sqrt{\sum x_2^2}}\right) \tag{3.30}$$

"By setting the above expression equal to zero, we can determine the values(s) of r_{12} which are turning points. We can find the appropriate value of r_{12} for a minimum by setting the second term equal to zero and solving. This yields either:

$$r_{12}=-\frac{\beta_1\sqrt{\sum x_1^2}}{\beta_2\sqrt{\sum x_2^2}} \tag{3.31}$$

or:

$$r_{12} = -\frac{\beta_2\sqrt{\sum x_2^2}}{\beta_1\sqrt{\sum x_1^2}} \tag{3.32}$$

> "As these are mutual inverses, only one can lie in the range (-1, 1). In either case, r_{12} will be positive if the coefficients β_1 and β_2 have opposite signs, and r_{12} will be negative if they have the same sign. This is exactly the result we are looking for but its application is clouded by the arbitrariness of the way in which we typically include variables in a model, that is, some are 'goods' and others are 'bads'. If our attributes X_1 and X_2 are TIME and COST, as proposed, then these are both 'bads'. In this case, to minimise the variance of VOT we should have r_{12} negative. We can achieve this by ensuring that our design predominantly contains combinations of TIME and COST such that slower journeys cost less than faster ones. If journey time had been presented in the design as a 'good', for example in terms of implied average speeds, then we would have wanted r_{12} positive. However, it is clear that this merely returns us to the same position, namely that slower journeys should cost less (in the design) than faster journeys."

They then presented illustrative examples and simulation results demonstrating the above findings. The paper ends with some concluding remarks:

> "We have set out our understanding so far of an issue which appears to be of some practical significance and which has certainly been under-researched. It seems that there is scope for achieving worthwhile improvements in the precision with which 'target' valuations are estimated. This can reduce the sample size necessary for a given level of precision, with associated cost savings, or else allow more confidence to be placed in the results obtained. Clearly the approach can be amended to minimise an objective function specifying the purpose of any given study. However, this is only a start and much remains to be done . The key issues which we believe require attention are:
>
> i) The extent to which worthwhile improvements in precision can be obtained without correlation problems, and indeed what level of correlation is tolerable.
>
> ii) How does a procedure to obtain more precise values of time fit in with a desire to incorporate adequate boundary values? It may be that the two objectives conflict.
>
> iii) The procedure needs to be generalised beyond the two independent variables and constant case examined here."

Ortuzar and Armstrong (1995) approach the problem in a rather different way, namely the minimisation of the confidence interval around the estimate of the value of time. It can easily be seen from their results that the width of the confidence interval is not minimised by setting the correlation between time and cost at zero.

Our own later results, from an EPSRC funded project are given in Watson et al (1996).

3.11 Conclusions

This chapter has set out the path of developments in SP design since 1983, as we see it. Advances at ITS Leeds are, of course, particularly well known to us, and form a major part of the above narrative, but we have not deliberately ignored comments on design made by others.

Despite the developments, most SP designs appear still to be based on the orthogonal cook-book methods of Chapter 2. In recent years many designers have included consideration of boundary volumes and rays, but few have been happy to dispense with orthogonal designs. The gains to be had by departing from orthogonality will be investigated in the current project.

4. ADAPTIVE STATED PREFERENCE

4.1 Introduction

The growth of computing power, especially in truly portable machines, has made it possible to make the later stages of an SP design depend on responses to earlier SP questions. This is different to what we term 'customisation', by which we mean that the SP design is varied as a whole, in response to questions asked prior to the SP experiment. For example, the experiment may be designed in differences, and related to actual times and costs reported by the respondent, giving a screen display of items and costs that would seem very familiar to the respondent. This is mere customisation, whereas this section will be looking at varying individual SP questions/scenarios in the light of earlier responses.

The first work we are aware of in adaptive SP is Johnson (1985), but we have not seen this work. The first adaptive SP we are aware of in transport is Ampt, Bradley and Jones (1987), and Bradley, Jones and Ampt (1987). These relate to the same work and we will refer to the latter paper. It is clear that more is done then mere customisation. There is a stopping criterion, up to which point additional SP scenarios were to be presented to respondents. These scenarios used a "partially pre-specified design framework, but choosing the factors to vary at random until the relative attractiveness of the options, based on current importance weights, is within a certain interval". Hence, the method is updating its estimate of the parameters and choosing scenarios that make sense for a respondent with something like those parameters. This bears some similarity with the idea of posing an SP choice containing a boundary value ray selected to pass not too far away from the current best estimate of the attribute valuations. We are not aware of any published example of large scale experiments using this method.

Work by Bradley and Daly (Bradley and Daly, 1993) showed that bias could easily be introduced into SP data when using adaptive techniques. In particular, this was found to be the case where there was taste variation present, either in the coefficients or in the residual error term. They say

> "This bias arises from the fact that the levels of the independent variables become correlated with the unmeasured components of individual preferences across the sample".

Suggestions which they put forward to address the problems they encountered included:

i. to use market segmentation to ensure that the sample is as homogenous as possible;

ii. to fit models separately for each respondent;

The use of adaptive methods in transport declined markedly after this paper, because of the problems identified and because of generally poor results obtained with adaptive designs. An exception to this was the LASP method, which is described next, which avoids the problems of bias by fitting separate models for each respondent. Use of adaptive techniques outside of Transport continues (see Carmone and Shaffer, 1995, Sawtooth Inc. 1996).

4.2 Leeds Adaptive Stated Preference (LASP)

This method was introduced by Fowkes and Tweddle (1988) in the context of the study of the attribute valuations, and hence mode choice, of freight shippers. It has always been computer based, with face-to-face interviews utilising portable PCs.

> "SP techniques appeared to be appropriate for the study of freight transport demand because of the ability to use hypothetical data, with the minimum of commercial sensitivity. Revealed Preference (RP) studies were considered to be impractical for three main reasons. Firstly, most freight rates actually paid were commercial confidential. Secondly, few firms would perceive themselves as facing realistic alternatives for general merchandise. Thirdly, we were specifically concerned to investigate the role of new technologies not yet in common use. SP methods could provide hypothetical choice scenarios, but a difficulty here was in providing reasonably realistic looking alternatives for the traffic flow in question. Clearly the alternatives available for transporting chilled meat products are quite different to those for transporting bagged fertiliser. Design of Stated Preference experiments is a complex art (Fowkes and Wardman, 1988) and it is typically the case that separate experimental designs will have to be prepared for sets of respondents facing markedly different actual choice situations."

The exact method of design of LASP was developed during that project and since. It is now usual to present respondents with 8 to 12 scenarios (or screens) each containing 4 options. The first of these is usually chosen to be similar to the respondent's current actual choice, as determined before the SP begins. This option is rated 100 and respondents are asked to rate other options in relation to this. The other options may differ in price, mode, journey time, reliability or whatever. Usually it is possible always to be able to find two options where only two attributes are varying. Respondents having difficulty in determining a rating, as is usual in the early stages, are first asked to rank the options and then very roughly suggest a strength of feeling about the difference between each adjacent pair in the ranking. Where it is thought it might help, it is suggested to respondents that if they think one option is 'twice as good' as another, they should give it twice the rating.

It should be obvious from the above that most of the information obtained comes from the rank ordering. Each respondent will be using their own rating scale and so ratings cannot be pooled over different respondents. However, provided we only calibrate at the level of the individual respondent, the rating data does provide extra information over and above the implied rankings. Over the years, intricate ways have been developed to maximise the efficiency of the calibration stage, but that is beyond the scope of this paper. It is now usually possible to obtain useful valuation information from most respondents. Sometimes the early responses are discarded if it appears that the respondent has been learning at that time. Serious inconsistencies in the data can certainly spoil the analysis.

Once one scenario has been given its ratings, these are used to set the attribute levels on the next screen. It is expected that the relative valuations built in as boundary values to the initial screen may be far away from the respondent's own valuations. Accordingly, it is deemed necessary for the boundary values to be quickly adjusted to be near those of the respondent. Without knowing these, i.e. knowing the answer before we start, it is, of course, difficult to do. Initially, the cost level is changed to see what effect this has on the rating. Options rated highly will have cost increases whereas those rated lowly will have cost decreases. The response ratings for the second screen will usually differ from those for the first screen, although the rankings may not. By noting how much the ratings change in response to the known cost change, a form of 'scaled' dead-reckoning is used to try to make the respondent indifferent between selected pairs of options on the next screen. The scaling is necessary to deal with the idiosyncrasies of individual respondent's rating schemes, and to ensure that the algorithm does not become stuck.

Respondents are allowed to be inconsistent but, up until they are, all earlier responses are allowed to limit the range within which new cost levels are chosen. Once two options are deemed to have been rated near enough equally, the level of one of the non-cost attributes is changed and the process restarts. The experiment concludes after a fixed number of scenarios (screens), or at the request of the respondent. The collected data is the form of columns of attribute levels (including mode specific dummies) followed by the respondent's rating. The LASP algorithm does not analyse this output, a combination of commercial statistical packages and specially written routines being used instead. Data for each scenario is converted into pairwise comparisons and a weighted logit model calibrated for each individual respondent. Models for groups of respondents are calculated by taking weighted averages of the individual attribute valuations. Weighting is by the inverse of the variances of the relevant coefficient estimates.

Several successful applications of LASP have been conducted and written up in commercially confidential reports. some results of the freight mode choice study mentioned above are in Fowkes, Nash and Tweddle (1991). The MVA Consultancy modified the method for a study of bulk freight (Terzis and Copley, 1992, Bates and Terzis, 1992).

5. CONCLUSION

This paper has taken a tour through the development of SP methods in transport planning. In particular, attention has been focused on the design of SP experiments. Much has been omitted but, where possible, individual authors have been left to say what they wanted at the time they wrote. In this way, it is hoped that some feel for the dynamic of the advances can be gained. It is hoped that readers will not be overawed by the material, but will appreciate the smallness of the steps by which progress is made and be prompted to make their own suggestions for improvements.

REFERENCES AND BIBLIOGRAPHY

Acito, F. and Jain, A.K. (1980), Evaluation of conjoint analysis results: a comparison of methods, *Journal of Marketing Research*, (17), pp 106-112.

Adamowicz, W.J., Louviere J. and Williams M. (1994). Combining revealed and stated preference methods for valuing environmental amenities, *Journal of Environmental Economics and Management* (26), pp 271-292

Addelman, S. (1962). Orthogonal main-effect plans for asymmetrical factorial experiments. *Technometrics,* (4), pp 21-46

Adler, G. T., Hyman, W. and Aunet, B. (1982). *Guide to forecasting travel demand and direct utility assessment,* Urban Mass Transportation Administration, Report UMTA-NH-11-0001-82-1, Washington DC.

Akaah, I.P., and Korgaonkar, P.K. (1983). An empirical comparison of the predictive validity of self-explicated, Huber Hybrid, Traditional Conjoint and Hybrid Conjoint models, *Journal of Marketing Research,* (2), pp 187-197

Ampt, E., Bradley, M.A. and Jones, P.M. (1987). *Development of an interactive computer assisted stated preference technique to study bus passenger preferences.* Paper presented to TRB Conference, Washington;

Anderson, N.H. (1970). Functional measurement and psychophysical judgement. *Psychological Review* (77), pp 153-170

Anderson, N.H. (1971). An exchange on functional and conjoint measurement. *Psychological Review,* (78), pp 457-458

Anderson, N.H. (1982). *Methods of information integration theory.* New York: Academic Press

Anderson, N.H., and Shanteau, J. (1977). Weak inference with linear models. *Psychological Bulletin,* (84), pp 1155-1170.

Axhausen, K W., Köll, H. and Bader, M. (1998), Experiments with SP and CA approaches to mode choice, *Proceedings of Seminar D, Transportation Planning Methods* I, (p423), PTRC, London, pp 41-55.

Banister, D. (1979). A method for investigating consumer preferences for certain transport policy alternatives. *Resources and Planning*, eds B Goodall and A Kirby. Pergamon Press

Bates, J.J. (1983). *Stated preference techniques for the analysis of transport behaviour.* World Conference on Transport Behaviour, Hamburg

Bates, J.J. (1984). *Values of time from stated preference data.* Proceedings of seminar HPTRC Summer Annual Meeting, Brighton

Bates, J.J. (1985). Econometric issues in Stated Preference Analysis, *Journal of Transport Economics and Policy,* (22), pp 56 - 69

Bates, J.J., (1988). Econometric issues in stated preference analysis, *Journal of Transport Economics and Policy,* (22), pp 59-69

Bates, J.J. (1994), Reflections on stated preference: theory and practice, *Proceedings of 7th International Conference on Travel Behaviour*, Santiago, Chile.

Bates, J.J., Dix, M., and May, A.D. (1987). Travel time variability and its effect on time of day choice for the journey to work, *Proceedings of Seminar C, PTRC Summer Annual Meeting, Bath,* PTRC, London

Bates, J.J., Gunn, H.F., and Roberts, M. (1978). A disaggregate model of household car ownership, *Department of the Environment, Research Report 20*

Bates, J.J., and Roberts, M. (1983). *Recent Experience with models fitted to stated preference data.* Proceeding of Seminar N, PTRC summer Annual Meeting, Brighton

Bates, J.J., and Terzis, G. (1992). Surveys involving adaptive stated preference techniques, *Proceedings of the Conference on Survey and Statistical Computing,* ed Westlake A, et al, Elsevier Science Publishers B.V.

Batsell, R.R., and Louviere, J.J. (1991). Experimental analysis of choice, *Marketing Letters,* (2), pp 199-214

Beggs, S., Cardell, N.S., and Hausman, J. (1981). Assessing the potential demand for electric cars, *Journal of Econometrics,* (16), pp 1-9

Ben-Akiva, M.E., and Lerman, S.R. (1985). *Discrete Choice Analysis: Theory and application to travel demand,* MIT Press, Cambridge, Mass.

Ben-Akiva, M.E., Morikawa, T., and Shiroishi, F. (1992). Analysis of the reliability of preference ranking data, *Journal of Business Research,* 24(2), pp 149-164

Benjamin, J., and Sen, L. (1983). *An empirical test of the utility of attitudinal measures for short-term public transportation planning.* Springfield, VA: National Technical Information Service

Benwell, M., and Black, I.G. (1984). *Train service reliability on British Rail InterCity services, Report 3, passenger attitudes to lateness.* Unpublished report to British Railways Board, Centre for Transport Studies, Cranfield Institute of Technology

Benwell, M., and Black, I.G. (1985). *User valuation of reliability in inter-urban travel: some problems of modelling and measurement,* Proceedings of seminar L, PTRC Summer Annual Meeting, Brighton

Borsch-Supan, A. (1986). *Econometric analysis of discrete choice.* Springer-Verlaag, Berlin

Bovy, P.H.L., and Bradley, M.A. (1985). Route choice analysed with stated-preference approaches, *Transportation Research Record,* (1037), pp 11-20

Box, G.E.P. (1952). Multifactor designs of the first order. *Biometrica.* (39) pp 49-57

Bradley, M. (1988). Realism and adaptation in designing hypothetical travel choice concepts. *Journal of Transport Economics and Policy* (22) pp 121-137

Bradley, M.A., and Bovy, P.H.L. (1985). A stated preference analysis of bicyclist route choice, *PTRC Summer Annual Meeting,* Sussex, pp 39-53

Bradley, M.A., and Bovy, P.H.L. (1986) Functional measurement of route choice behaviour - a study of commuter cyclists in the Netherlands, *Behavioural Research for Transport Policy*, VNU Science Press, The Netherlands.

Bradley, M., and Daly, A. (1993). *New analysis issues in Stated Preference research.* Proceedings of Seminar D, PTRC-SAM-1993, published as 'Transportation Planning Methods,' Code P366, pp75-89, PTRC London

Bradley, M.A., and Daly, A. (1994), Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data, *Transportation,* (21), pp167-201.

Bradley, M., Jones, P.M. and Ampt, E. (1987). *An interactive household interview method to study bus provision policies;* in Transportation Planning Methods, Proceedings of Seminar C, Code P290, pp 163-178, PTRC London

Bradley, M., Marks P., and Wardman M, (1986). *A summary of four studies into the value of travel time savings,* Transportation Planning methods, Proceedings of Seminar M, PTRC, London, pp 271-285

Broom, D., Lowe, S., Gunn, H. and Jones, P.M. (1983). *Estimating values of time: an experimental comparison of Transfer Price methods with the Revealed Preference approach,* PTRC Summer Annual Meeting, Brighton (not in proceedings)

Carmone, F.J. (1986). Conjoint Designer. *Journal of Marketing Research* (23), pp 311-312

Carmone, F.J. (1987). ACA System for Adaptive Conjoint Analysis. *Journal of Marketing Research,* (24), pp 325-327

Carmone, F.J. and P.E. Green, (1981). Model misspecification in multiattribute parameter estimation. *Journal of Marketing Research,* (18), pp 87-93

Carmone, F.J., and Green, P.E. and Jain, A.K. (1978). The robustness of Conjoint Analysis: some Monte-Carlo results, *Journal of Marketing Research* (15), pp 300-303

Carmone, F.J. and Shaffer, C.M. (1995), Software review: Conjoint Designer, Conjoint Analyzer, SIMGRAF, Conjoint Linmap, Conjoint Segmenter, Bridger, CONSURV, ACA, CBC, CVA, *Journal of Marketing Research,* (32), pp113-120

Carroll, J.D. (1972). Individual differences and multidimensional scaling, in *Multidivisional scaling: theory and applications in behavioural sciences*, vol 1, eds Shephard, R.N. et al, Seminar Press, New York, pp 105-155

Carroll, J.D. and Green, P.E. (1995). Psychometric methods in marketing research: Part 1, conjoint analysis, *Journal of Marketing Research*, (32), pp385-391

Cattin, P., Gelfand, A., and Danes, J. (1983). A simple Bayesian procedure for estimation in a conjoint model. *Journal of Marketing Research* (20), pp 29-35

Cattin, P., and Wittink, D.R. (1982). Commercial use of conjoint analysis: a survey *Journal of Marketing* (46) pp 44-53

Cattin, P., and Wittink, D.R. (1986). *Commercial use of conjoint analysis: An update.* Paper presented to the ORSA/TIMS Marketing Science Meetings, Richardson, TX, March 12-15

Chapman, R.G., and Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research,* (19), pp 288-301

Cirillo, C., Daly, A.J. and Lindveld (1996). Eliminating bias due to the repeated measurements problem in SP data, *Proceedings of Seminars D and E, European Transport Forum, Brunel University*, PTRC, London

Cochran, W.G., and Cox, I.G.M. (1957). *Experimental design,* John Wiley, New York

Copley, G., Bouma, A and de Graaf, E (1987) *Station investment : priorities and values* proceedings of ESOMAR conference, Montreaux, Switzerland.

Crask, M.R., and Fox, R.J. (1987). An exploration of the interval properties of three commonly used marketing research scales: a magnitude estimation approach. *Journal of the Market Research Society,* (29), pp 317-339

Crawford, H.C. (1987). *Uncertainty and the value of passenger waiting time*. Unpublished MSc dissertation, Centre for Transport Studies, Cranfield Institute of Technology

Crittle, F.J., and Johnson, L.W. (1980). *Basic Logit (BLOGIT) technical manual*, ATM No 9, Australia Road Research Board, Victoria, Australia

Daganzo, C. (1979). *Multinomial Probit,* Academic Press, New York

Daly, A. (1996). *Estimating VOT*, paper presented to course on value of time, Easthampstead, PTRC, London

Davidson, J.D. (1973). Forecasting traffic on STOL *Operational Research Quarterly* (24) pp 561-569

Dawes, R.M., and Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin,* (81), pp 95-106

de Jong, G.C., Gommers, M.A., and Klooster, J.P.G.N. (1992). Time valuation in freight transport: method and results, *PTRC Summer Annual Meeting, UMIST,* PTRC, London

DeSarbo, W.S., Carroll, J.D., Lehmann, D.R., and O'Shaughnessy J. (1992). Three -Way Multivariate Conjoint analysis. *Marketing Science.* (1), pp 323-350

DeSarbo, W.S., Wedel, M., Vriens, M., and Venkatram, R. (1992). Latent class metric conjoint analysis, *Marketing Letters,* (3), pp 273-288

Dobson, R. (1975). Towards the analysis of attitudinal and behavioural responses to transportation system characteristics. *Transportation,* 4, pp 267-290

Donnelly, E.P. (1976). Preference elasticities of transit fare increases and decreases by demographic groups. *Transportation Research Record.* (589), pp 30-32

Donnelly, E.P., Howe, S.M. and DesChamps, J.A. (1976). Trade-off analysis: theory and applications to transportation policy analysis, *Preliminary Research Report 103, Planning Research Unit, New York State Dept of Transportation, Albany, New York*

Eberts, P.M. (1977). Trade-off vs categorical judgement: a comparative analysis of two attribute scaling methods for transportation planning. *Preliminary Research Record,* 114.

Eberts, P.M., and Koeppel, K.W.P. (1977). The trade-off model: empirical and structure findings. *Preliminary Research Report 123 Planning Research Unit, New York State Dept of Transportation, Albany, New York*

Elrod, T., Louviere, J.J., and Davey, K.S. (1992). An empirical comparison of ratings-based and choice-based conjoint models, *Journal of Marketing Research,* XXIX, pp 368-377

Fenwick, I. (1978). A user's guide to conjoint measurement in marketing. *European Journal of Marketing,* 12, pp 203-211

Ford, I. (1976). (Title unknown) *PhD thesis, University of Glasgow*

Fowkes, A.S. (1985). *Checking equi-utility VOTs for SP design,* unpublished Value of Time Working Paper 25/2/85, UK DoT Value of Time Project

Fowkes, A.S. (1988). *Network South East quality of service study: Literature Review* Unpublished report for BR: NSE by the MVA Consultancy and Institute for Transport Studies, University of Leeds.

Fowkes, A.S. (1991a). Novel approaches to forecasting the demand for new local rail services, *Transportation Research* (25A), pp 209-218

Fowkes, A.S. (1991b). *Recent developments in stated preference techniques in transport research,* Transportation Planning Methods, p347, Proceeding of PTRC Summer Annual Meeting, Brighton, pp 251-263

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

Fowkes, A.S., Nash, C.A., and Tweddle, G. (1989). Investigating the market for inter-modal freight technologies, *Transportation Research* (25A), pp 161-172

Fowkes, A.S. and Preston, J.M. (1991). Novel approaches to forecasting the demand for new local rail services, *Transportation Research*, (25A), pp 209-218

Fowkes, A.S., and Tweddle, G. (1988). *A Computer guided stated preference experiment for freight mode choice* Proceeding of Seminar D, PTRC-SAM-1988, published as Transportation Planning Methods, P306, pp 295-305

Fowkes, A.S., and Wardman, M. (1987). The values of overcrowding and departure time valuations for InterCity rail travellers, *TN229 Institute for Transport Studies, University of Leeds,* (restricted distribution)

Fowkes, A.S., and Wardman, M. (1988). Design of SP travel choice experiments, with special reference to taste variations. *Journal of Transport Economics and Policy,* (22) pp 27-44

Fowkes, A.S., Wardman, M. and Holden, D.G.P. (1993). *Non-orthogonal Stated Preference design* in proceeding of PTRC-SAM-1993, UMIST, Transportation Planning Methods, P366, pp 91-97, PTR, London

Gaudry, M.J.I., Jara-Diaz, S.R., and Ortúzar, J. de D. (1989). Value of time sensitivity to model specification, *Transportation Reseach,* (23B), pp 151-158

Green, P.E. (1974). On the design of choice experiments involving multifactor alternatives. *Journal of Consumer Research* (1) pp 61-68

Green, P.E. (1984). Hybrid models for conjoint analysis: an expository review, *Journal of Marketing Research,* (21), pp 155-159

Green, P.E., Carroll, J.D., and Carmone, F. (1978). Some new types of practical factorial designs for marketing experiments, in *Research in marketing,* Vol 1 (ed J.N. Sheth). JAI Press

Green, P.E., Carroll, J.D., and Goldberg, S.M. (1981). A general approach to product design optimisation via conjoint analysis. *Journal of Marketing* (45): pp 17-37

Green, P.E., and Krieger, A.M. (1996). Individualized hybrid models for conjoint analysis, *Management Science,* (42), pp 850-867

Green, P.E., Krieger, A.M. and Bansal, P. (1988). Completely unacceptable levels in conjoint analysis: A Cautionary Note, *Journal of Marketing Research* (25), pp 293-300

Green, P.E., and Rao, V.R. (1971). Conjoint Measurement for quantifying judgmental data, *Journal of Marketing Research* (8), pp 355-363

Green, P.E., and Rao, V.R. and Desarbo, W. (1978). Incorporating group level similarity judgements in conjoint analysis. Journal of Consumer Research (5), pp 127-192

Green, P.E., and Srinivasan V (1978). Conjoint analysis in consumer research; issues and outlook. *Journal of Consumer Research:* (5) pp103-123.

Green, P.E., and Srinivasan, V. (1990). Conjoint analysis in marketing research: new developments and directions, *Journal of Marketing,* (54) pp 3-19

Green, P.E., and Tull, D.S., and Albaum, G. (1988). *Research for marketing decisions*, 5th ed, Prentice-Hall, Englewood Cliffs, NJ

Green, P.E., and Wind, Y. (1973). *Multiattribute decisions in marketing: A measurement approach.* Dryden Press Hinsdale, IL,

Green, P.E., and Wind, Y. (1975). New way to measure consumer judgements, *Harvard Business Review,* (53) pp 107-117

Gunn, H.F. (1981). Value of Time estimation, *Working Paper 157, Institute for Transport Studies, University of Leeds*

Gunn, H.F. (1984). An analysis of Transfer Price data. *Proceedings of 12th PTRC Summer Annual Meeting, Seminar H, University of Sussex, PTRC, London*

Gun, H.F. and Wardman, M.R. (1985), Practical possibilities for enriching value of time data sets, *UK DoT Value of Time Study Working Paper, unpublished*

Gunn, H.F. and Whittaker, J.C. (1981). Estimation errors for well-fitting gravity models, *Working Paper 149, Institute for Transport Studies, University of Leeds*

Gujurati, D.N. (1988). *Basic Econometrics.* McGraw Hill, New York

Hagerty, M.R. (1985). Improving the Predictive Power of Conjoint Analysis: The Use of Factor Analysis and Cluster Analysis, *Journal of Marketing Research,* (22), pp 164-184

Hagerty, M.R. (1986). The cost of simplifying preference models, *Marketing Science,* (5), pp 298-319

Hahn, G.J., and Shapiro S.S. (1966). *A catalogue and computer program for the design and analysis of orthogonal symmetric and asymmetric fractional factorial experiments.* Report 66-C-165, General Electric Research and Development Centre, Schenectady, New York

Hansen, D.E. and Wittink, D.R. (1995). Combining self-explicated priors with conjoint data using Bayesian regression, *Marketing Letters* (6), pp 63-71.

Harison, W.J., Pell, C., Jones, P.M., and Ashton, H. (1986). Some advances in model design developed for practical assessment of road pricing in Hong Kong, *Transportation Research,* (20A), pp 135-143

Heeler, R., Okechuku, C. and Reid, S. (1979). Attribute importance: contrasting measurements. *Journal of Marketing Research.* (16), pp 60-62

Hensher, D.A. (1982). Functional measurement, individual preference an discrete-choice modelling: theory and application, *Journal of Economic Psychology,* (2), pp 323-335

Hensher, D.A. (1990). The orthogonality issue in stated choice designs, in Fischer M.M., Nijkamp, P. and Papageorgiou, Y.Y. (eds) *Spatial Choices and Processes,* North-Holland, Amsterdam, 265-278

Hensher, D.A. (1991). Hierarchical stated response designs and estimation in the context of bus use preferences, *Logistics and Transportation Reviews* (26), pp 299-323

Hensher, D.A. (1994). Stated preference analysis of travel choices: The state of practice, *Transportation,* (21) pp 107-133

Hensher, D.A., and Barnard P.O. (1990). The orthogonality issue in stated choice designs, in Fischer, M. Wijkamp, P. and Papageorgiou, Y. (eds) *Spatial Choices and Processes,* North-Holland, Amsterdam, pp 265-278

Hensher, D.A., and Louviere, J.J. (1983). Identifying individual preferences for international air fares, *Journal of Transport Economics and Policy,* XVII, pp 225-245

Hoinville, G., and Berthond E. (1969). *Value of time : development project stage 2.* Social and Community Planning Research, London

Hoinville, G., and Countenay G. (1978). Measuring consumer priorities. in *Resource Management and environmental planning*. O'Riordan T. (ed), Wiley

Hoinville, G., and Johnson E. (1971). *The importance and values commuters attach to time savings,* Social and Community Planning Research, London

Holden, D.G.P. (1993). *Design procedure for stated preference experiments* PhD Thesis, Institute for Transport Studies, University of Leeds

Holden, D.G.P., Fowkes, A.S. and Wardman, M. (1992). *Automatic stated preference design algorithms* in proceeding of PTRC-SAM-1992, MMST Transportation Planning Methods, pp 153-166, PTRC, London

Horowitz, J.L. (1981). Sampling specification and data errors in probabilistic discrete-choice models, Appendix C of Hensher, D.A. and Johnson, L.W. (1981) *Applied Discrete-Choice Modelling,* Croom Helm (London) and Wiley (New York), pp 417-435.

Howe, S.M., Donnelly, E. and DesChamps, J. (1977). Trade-off analysis, theory and applications to transportation policy planning. *High Speed Ground Transportation* 11(1)

Howe, S.M., and Liou, P.S. (1975). Documentation of PROLO and MLOGIT, *Preliminary Research Report 98, Planning Research Unit, New York State Department of Transportation,* Albany, NY

Hu, C. (1995), *Conjoint analysis library*, University of Nevada at Las Vegas, NV 89154-6023

Huber, J. (1997). What have we learned from 20 years of conjoint research: when to use self-explicated, graded pairs, full profiles or choice experiments, *Proceedings of the Duke University Sawtooth Software Conference*

Huber, E., and Hansen D. (1986). Testing the impact of dimensional complexity and affective differences of paired concepts in adaptive conjoint analysis. *Duke University School of Business*

Huber, J., Wittink, D.R., Fiedler, J.A. and Miller, R. (1993). The effectiveness of alternative preference elicitation procedures in predicting choice. *Journal of Marketing Research,* (30), pp 105-114

Jackson, W.B., and Jucker, J.V. (1982). An empirical study of travel time variability and travel choice behaviour, *Transportation Science,* (16), pp 460 - 476

Jain, A.K., Acito, F., Malhotra, N.K., and Mahajan, V. (1979). A comparison of the internal validity of alternative parameter estimation methods in decompositional multiattribute preference models. *Journal of Marketing Research.* (16), pp 313-322

Johnson, E., Meyer, R.J., and Ghoose, S. (1989). When choice models fail: compensatory models in relatively correlated environmentals, *Journal of Marketing Research,* (26), pp 255-270

Johnson, R.M. (1973). Varieties of conjoint measurement, *Working Paper, Market Facts Inc.* Chicago

Johnson, R.M. (1974). Tradeoff analysis of consumer values. *Journal of Marketing Research,* (11), pp 121-127

Johnson, R.M. (1985). *Adaptive conjoint analysis*, Sawtooth Software Inc., Idaho.

Kagel, J.H. and Roth, A.E. (1995). *The handbook of experimental economics*, Princeton University Press, Princeton, USA

Kendall, M.G., and Stuart, A. (1969). *The advanced theory of statistics, Vol 1,* Griffin, London

Kocur, G., Adler, T., Hyman, W., and Aunet, B. (1982). *Guide to forecasting travel demand with direct utility assessments,* US DoT, Urban Mass Transportation Administration, Report UMTA-NH-11-0001-82-1, Washington, DC

Kocur, G., Hyman, W., and Aunet, B. (1982). Wisconsin work mode-choice models based on functional measurement and disaggregate behavioural data. *Transportation Research Record,* (895), pp 24-31

Koeppel, K.W.P. (1977). Functional forms of the utility integration rule in trade-off analysis. *Preliminary Research Report 122, Planning Research Unit, New York State Dept of Transportation, Albany, New York*

Koeppel, K.W.P., and Neveu, A.J. (1979). *An acid test of the trade-off method of attitude measurement.* Preliminary Research Report No 160, Planning Research Unit, NY State Department of Transportation, Albany NY

Krantz, D.H. (1964). Conjoint measurement: the Luce - Tukey axomatization and some extensions, *Journal of Mathematical Psychology,* (1), pp 284-297

Krantz, D.H. Luce, R.D., Suppes, P. and Tversky, A. (1971). *Foundations of Measurement,* Academic Press, New York

Krantz, D.H., and Tversky, A (1971). Conjoint Measurement Analysis of composition rules in psychology, *Psychological Review,* (78) pp 151-169, pp 457-458.

Kroes, E.P., and Sheldon R.J. (1988). Stated Preference methods: An introduction: *Journal of Transport Economics and Policy,* (XXII) pp 11-25

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data, *Journal of the Royal Statistical Society,* Series B, (27) pp 251-263

Lancaster, K.J. (1966). A new approach to consumer theory. *Journal of Political Economy,* (74) pp 132-157

Lehmann, D.R. (1989). *Market research and analysis,* 3rd ed. Richard D Irwin Inc, Homewood, IL

Leigh, T.W., McKay, D.B. and Summers, J.O. (1984). Reliability and validity of conjoint analysis and self explicated weight: a comparison, *Journal of Marketing Research,* (21), pp 456-462

Lerman, S.R., and Louviere, J.J. (1978). On the use of functional measurement to identify the functional form of the utility expression in travel demand models, *Transportation Research Record* (673), 78-86

Levin, I.P., and Louviere, J.J., Schepanski, A.A. and Norman, K.L. (1983). External validity tests of laboratory studies of information integration. *Organisation Behaviour and Human Performance,* (31), pp 173-193

Lieber, S.R. (1976). A comparison of metric and non-metric scaling models in preference research. In Golledge, R.G., and Rushton, G. (eds) *Spatial Choice and Spatial Behaviour.* Columbus, OH: Ohio State University Press

Louviere, J.J. (1974). Predicting the evaluation of real stimulus objects from an abstract evaluation of their attributes: the case of trout streams. *Journal of Applied Psychology* (59), pp 572-577

Louviere, J.J. (1979). Modelling individual residential preferences: a totally disaggregated approach. *Transportation Research,* (13A), pp 374-384

Louviere, J.J. (1983). Integrating conjoint and functional measurement with discrete choice theory: an experimental design approach. *Advances in Consumer Research* (10), pp 151-156

Louviere J.J. (1988a). Conjoint analysis modelling of stated preferences: 'A review of theory, methods, recent developments and external validity', *Journal of Transport Economics and Policy,* (XXII) pp 93-119

Louviere, J.J. (1988b). *Analysing decision making: Metric Conjoint Analysis.* Sage University Paper No 67, Newbury Park, Beverley Hills

Louviere, J.J. (1993). Conjoint analysis, in R. Bagozzi (ed.) *Handbook of Marketing Research,* Academic Press, New York

Louviere, J.J., Beavers, L.L., Norman, K.L., and Stetzer, F. (1973). *Theory, methodology and findings in mode choice behaviour, Working Paper No 11, Institute of Urban and Regional Research, University of Iowa*

Louviere, J.J., Henley, D.H., Woodworth, G., Meyer, R.J., Levin, I.P., Stoner, J.W., Curry, D., and Anderson D.A. (1981). Laboratory simulation versus revealed preference methods for estimating travel demand models: an empirical comparison. *Transportation Research Record,* (794), pp 42-51

Louviere, J.J., and Hensher, D.A. (1982). On the design and analysis of simulated choice or allocation experiments in travel choice modelling. *Transportation Research Record,* (890), pp 11-17

Louviere, J.J., and Hensher, D.A. (1983). Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event, *Journal of Consumer Research,* (10), pp 348-361

Louviere, J.J., and Kocur, G. (1983). The magnitude of individual level variation in demand coefficients: a Xenia, Ohio, case example, *Transportation Research,* (17A), pp 363-374

Louviere, J.J. and Timmermans, H. (1990), Stated preference and choice models applied to recreation research: a review, *Leisure Sciences,* (12), pp 9-32

Louviere, J.J., and Woodworth, G.G. (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data, *Journal of Marketing Research,* (20), pp 350-367

Louviere, J.J., and Woodworth, G.G. (1986). *You can have your cake and eat it too: correlated attributes choice designs that permit orthogonal parameter estimation.* Paper presented to the 1986 ORSA/TIMS Marketing Science Conference, Richardson, TX

Louviere, J.J., and Woodworth, G.G. (1987). *On the design and analysis of correlated conjoint experiments using difference designs.* Paper presented to the annual meeting of the Association for Consumer Research, Boston, MA

Luce, R.D. (1959). *Individual Choice Behaviour.* New York: John Wiley and Sons

Luce, R.D., and Suppes, P. (1965). Preference utility and subjective probability, in *Handbook of Mathematical Psychology,* (3) Wiley, New York

Luce, R.D., and Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement; *Journal of Mathematical Psychology,* (1), pp 1-27

Madden, G.G. (1992). Social experimentation in economics: an overview of the stated-preference experimental design method, paper presented to the *Fourth Conference of the Australian Centre for Experimental Economics,* University of Adelaide, 19 - 21 June

Malhotra, N.K. (1982). Structural reliability and stability of non-metric conjoint analysis, *Journal of marketing Research,* (19), pp 199-207

Marks, P., and Fowkes A.S. (1986). Stated Preference experiments concerning long distance business travel in Great Britain. Working Paper 219. *University of Leeds, Institute for Transport Studies*

Marks, P., and Fowkes, A.S. and Nash, C.A. (1986). Valuing long distance business travel time savings for evaluation: a methodological review and application, *Proceedings of Seminar M, PTRC-SAM, Transportation Planning Methods, Code P282,* PTRC, London pp 87-98

Mason, C.H., and Perreault, W.D. (1991). Collinearity, power, and the interpretation of multiple regression analysis, *Journal of Marketing Research* (28) pp 268-280

McCullough, J., and Best, B. (1979). Conjoint measurement: temporal stability and structural reliability. *Journal of Marketing Research,* pp 26-31

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour in *Frontiers in Econometrics*, ed. Zarembka P Academic Press, New York, pp 105-142

McFadden, D. (1981). Econometric models of probabilistic choice. In Manski, C. and McFadden, D. (eds). *Structural analysis of discrete data.* Cambridge, MA: MIT Press, pp 198-272

McFadden, D. (1986). The choice theory approach to marketing research. *Marketing Science,* (5), pp 275-297

Meyer, R.J., Levin, I.P., and Louviere J. (1978). Functional analysis of mode choice. *Transportation Research Record.* (673), pp 1-7

Miller, R.G. (1994). The Jacknife - A review, *Biometrica,* (61), pp 1-14

Ministry of Transport and Public Works, The Netherlands (1985). Workshop 5: Stated preference methods. In Ruhl, A. (ed) *Behavioural Research for Transport Policy.* Vol 2: Introductory Papers. Preprint of the proceedings of the 1985 International Conference on Travel Behaviour. Den Haag: The Ministry

Morikawa, T. (1994). Correcting state dependence and serial correlation in the RP/SP combined estimation method, *Transportation,* (21) pp 153-165

MVA Consultancy and ITS, University of Leeds (1989). *Provincial overcrowding study: Final report.* Unpublished report to BR Provincial sector.

MVA Consultancy; ITS, University of Leeds; TSU, University of Oxford (1987). *The value of travel time savings,* Policy Journals, Newbury, Berkshire

Nicolaidis, G., and Dobson, R. (1975). Disaggregated perceptions and preferences in transportation planning. *Transportation Research,*(9), 279-295

Oliphant, K., Eagle, T.C., Louviere, J.J. and Anderson, D. (1992). Cross-task comparisons of ratings-based and choice-based conjoint, in *Sawtooth Software Conference Proceedings,* ed Metergrando M., Sawtooth Software, Sun Valley, ID

Oppewal, H. (1995). A review of conjoint software, *Journal of Retailing and Consumer Services,* (2), pp 55-61 and pp 259-264

Ortúzar, J. de D., and Armstrong, P.M. (1995). Confidence Intervals and the social value of travel time savings, *Proceedings of Seminar F, Models and Applications, Code, P393, PTRC-SAM,* PTRC, London

Ortúzar, J. de D., and Garrido, R. (1991). Rank, rate or choice? An evaluation of SP methods in Santiago, in *Proceedings of Seminar G, Transportation Planning Methods,* P347, PTRC, London, pp 301-312.

Ortúzar, J. de D., and Garrido, R.A. (1992). An application of ordinal probit to SP rating data, *Transportation Planning Methods, Proceedings of Seminar E, P357, PTRC,* London.

Ortúzar, J. de D., and Garrido, R.A. (1994). On the semantic scale problem in stated preference rating experiments, *Transportation,* (21), pp 185-201.

Ouwersloot, H., and Rietveld, P. (1996). Stated choice experiments with repeated observations. *Journal of Transport Economics and Policy,* (30), pp 203-212.

Parker, B., and Srinivasan, V. (1976). A consumer preference approach to the planning of rural health facilities. *Operations Research,* (24), 991-1025

Payne, G., and Steer, J. (1981). Gauging passenger preference structures using a gaming approach. *Proc of Market Research Society Annual Meeting, March 1981*

Payne, J.W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis, *Organisational Behaviour and Human Performance,* (16), pp 366-387

Pearmain, D., Swanson, J., Kroes, E. and Bradley, M. (1991). *Stated Preference Techniques: A guide to practice,* Second Edition, Steer Davies Gleave and Hague Consulting Group

Pells, S. (1987). The evaluation of reductions in the variability of travel times to work, *Transportation Planning Methods, Proceedings of Seminar C, PTRC* pp 313-325

Roussens, S.W., and Hart, A.G. (1951). Experimental verification of a compositie indifference map, *Journal of Political Economy,* (59), pp 288-318.

Sawtooth Inc. (1995). *The CBC system for choice - based conjoint analysis*, Sawtooth Software Inc., Evanston

Sawtooth Inc. (1996). *ACA system - adoptive conjoint analysis - Version 4.0,* Sawtooth Software Inc., Evanston

Sheldon, R.J. and Steer, J.K. (1982). The use of Conjoint Analysis in transport research, *PTRC 10th Summer Annual Meeting, Seminar Q,* Warwick.

Silvey, S.D. (1980). *Optimal design,* Chapman and Hall, London

Slovic, P., and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organisational Behaviour and Human Performance.* (6), pp 649-744

Srinivasan, V., Jain, A.K., and Kalhotra N.K. (1983). Improving predictive power of conjoint analysis by constrained parameter estimation, *Journal of Marketing Research,* (20), pp 433 - 438

Srinivasan, V., and Shocker, A.D. (1973). (i) Linear programming techniques for multi-dimensional analysis of preferences (ii) Estimating the weights for multiple attributes in a composite criterion using pairwise judgements, *Psychometrika* (38)

Steckel, J.H., DeSarbo, W.S. and Mahajan, V. (1990). On the creation of feasible conjoint analysis experimental designs, *Decision Sciences.*

Steckel, J.H. and O'Shaughnessy, J. (1989). Towards a new way to measure power: Applying conjoint analysis to group discussions, *Marketing Letters,* (1), pp 37-46

Steer, Davies and Gleave Ltd (1981). *Final report on research into elasticity of demand in respect of service frequency and through trains.* Unpublished report for British Railways Board, London.

Steer, J., and Gleave G. (1977). Techniques for inter-urban investment studies. *Proc PTRC Summer Annual Meeting, Warwick University, July 1977.* London: PTRC Education and Research Services Ltd

Swait, J., and Admowicz, W. (1996). The effect of choice environment and task demands on consumer behaviour: discriminating between contribution and confusion. *Staff Paper 96-09, Department of Rural Economy, University of Alberta.*

Swanson, J. (1998). Factors affecting the validity of stated preference. *Proceedings of seminar D, transportation Planning Methods I,* (P423), PTRC, London, pp 1-16

Swanson, J., Pearman, D., and Holden, D.G.P., (1993). *New approaches to the design of stated preference experiments,* PTRC-SAM-1993, Seminar D, Published as 'Transportation Planning Methods', code P366, pp 99-109, PTRC London

Terzis, G. and Copley, G. (1992). *Seeking new business opportunities for trainload freight* PTRC-SAM-1992, seminar E (not in published proceedings)

Thurstone, L.L. (1931). The indifference function. *Journal of Social Psychology,* (2), pp 139-167

Timmermans, D. (1993). The impact of task complexity on information use in multi-attribute decision making. *Journal of Behavioural Decision Making* (6), pp 95-111

Timmermans, H.J.P. (1982). Consumer choice of shopping centre: an information integration approach. *Regional Studies,* (16), pp 171-182

Timmermans, H.J.P. (1984). Decompositional multiattribute preference models in spatial choice analysis: a review of some recent developments, *Proc Human Geog.* (8), pp 189-221

Tull, D.S., and Hawkins, D.I. (1990). *Marketing research,* 5th ed. Macmillan, New York

Tversky, A. (1967). A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology.* (4), pp 1-20

Tersky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review* (79), pp 281-299

Vriens, M., Wedel, M and Wilms, T. (1996). Metric conjoint segmentation methods: A Monte Carlo comparison, *Journal of Marketing Research,* (33), pp 73-85.

Wainer, H. (1976). Estimating coefficients in linear models: it don't make no nevermind. *Psychological Bulletin* (83), pp 213-217

Wallis, W.A., and Friedman, M (1942). The empirical derivation of indifference functions, *Studies in Mathematical Economics and Econometrics in memory of Henry Shultz*, ed. Lange, O., McIntyre, F., and Yntema, T.O., University of Chicago Press, Chicago

Wardman, M.R. (1987). *An evaluation of the use of Stated Preference and Transfer Price data in forecasting the demand for travel:* PhD Thesis

Wardman, M.R., (1998). The value of travel time, *Journal of Transport Economics and Policy,* (32), pp 285-316

Wardman, M.R., and Fowkes, A.S. (1987). The values of overcrowding and departure time variations, of InterCity rail travellers, *University of Leeds, Institute for Transport Studies, Technical Note 229.* Unpublished, commercial in confidence

Watson, S.M., Toner, J.P., Fowkes, A.S., and Wardman, M.R., (1996). Efficiency properties of orthogonal stated preference designs, *Proceedings of Seminar D, Transportation Planning Methods,* (P404), pp 1-10, PTRC, London

Widlert, S. (1994). *Stated Preference studies: The design affects of results.* Paper presented to the 7th International Conference on Travel Behaviour, Valle Nevado, Santiago, Chile

Wiley, J.B. (1977). Selecting Pareto Optimal subsets from multiattribute Alternatives, in *Advances in Consumer Research,* (4), Perreault, W.D. Jr. ed. Atlanta, Association for Consumer Research, pp 174-174

Winer, B. (1962). *Statistical principles in experimental design,* McGraw Hill

Wittink, D.R., and Cattin, P. (1981). Alternative estimation of methods for Conjoint Analysis: A Monte-Carlo study, *Journal of Marketing Research,* (18), pp 101-106

Wittink, D.R., and Cattin, P. (1989). Commercial use of conjoint analysis: an update, *Journal of Marketing,* (53), pp 91-96

Wittink, D.R., and Montgomery, D. (1979). Predictive validity of trade-off analysis for alternate segmentation schemes. In *Proceedings, American Marketing Association Educators Conference,* Chicago, American Marketing Association

Wittink, D.R., Vrienns, M., and Burhenne, W., (1994). Commercial use of conjoint analysis in Europe: results and critical reflections, *International Journal of Research in Marketing,* (11), pp 41-52

Wright, P.L., and Kriewall, M.A. (1980). State-of-mind effects on the accuracy with which utility functions predict marketplace choice. *Journal of Marketing Research,* (17), pp 217 - 293

Zanna, M.P. (1994). *Experimental social psychology.* Vol 26, Academic Press, London