THE DESIGN OF STATED PREFERENCE
TRAVEL CHOICE EXPERIMENTS

With Special Reference to Interpersonal Taste Variations

By Tony Fowkes and Mark Wardman*

Stated Preference (SP) experiments present individuals with hypothetical travel scenarios and seek their preferences. SP techniques have advantages over Revealed Preference methods which are based on actual choices, because the individual can be asked to make more than one travel choice and can be presented with trade-offs rather than dominated choices. These benefits can, however, be squandered if the SP experiment is poorly designed. To produce a good SP design, one will probably have to produce and discard many bad ones. In this paper we shall suggest ways of testing the design so that faults can be detected and remedied. We shall set out guidelines for good SP design and show how the design can be tested by simulated data sets. It is a particular concern of ours that the SP design and the simulated data sets should allow for inter-personal taste variations. That tastes vary across individuals is taken as axiomatic. The paper therefore starts by setting out the theoretical background and discussing the implications of variations in taste. Although much of the discussion focuses on the value of travel time, the issues considered apply equally to relative valuations in general.

THE IMPACT OF TASTE VARIATIONS

Theoretical background

Suppose that the individual maker of a travel decision is confronted (either in practice, or in a hypothetical SP experiment) with a choice set of travel alternatives, which, for example, might be various travel modes by which some desired destination can be reached. The choice of mode is assumed to depend on the

* Institute for Transport Studies, University of Leeds. While writing this paper the authors were funded by British Rail and by the Science and Engineering Research Council. Many colleagues have contributed to the understanding of the issues reported, principally John Bates, Mark Bradley, Hugh Gunn, Phillipa Marks and Juan de Dios Ortuzar.
relative utilities of the various travel options. We shall follow the conventional simplification of using an additive utility function:

\[ U_{im} = \sum_j \beta_{jm} X_{jm} \]  

where

- \( U_{im} \) is the modelled value for the utility perceived by individual \( i \) for alternative \( m \),
- \( X_{jm} \) is the value of the \( j \)th relevant attribute (explanatory variable) which is hypothesised to influence travel behaviour, and
- \( \beta_{jm} \) are parameters "known" by the individual decision maker, reflecting utility weights, but unknown to the observer, and to be in some way estimated.

The \( X \)'s may be transforms of measurable quantities; for instance, logarithms will give a multiplicative form.

The individual will choose that travel alternative from all those on offer which is perceived as having the highest utility. However, there will be influences on an individual's choice which we cannot measure or recognise. Thus a stochastic element, \( \epsilon_{im} \), is added to the deterministic expression of equation (1) to represent the net effect of omitted factors, and to form "Random Utilities", \( RU_{im} \), as:

\[ RU_{im} = U_{im} + \epsilon_{im} \]  

The probability that alternative 1 will be chosen rather than alternative 2 can be expressed as:

\[ P_{11} = \text{Prob}[(U_{i1} + \epsilon_{i1}) > (U_{i2} + \epsilon_{i2})] \]  

The assumptions made about the error terms determine the form of model that is developed. The assumption that the errors are independently and identically distributed with a Weibull distribution yields the most commonly used form of random utility model, the multinomial logit model of:

\[ P_{11} = \frac{\exp(U_{i1})}{\sum_m \exp(U_{im})} \]  

which is fully derived in McFadden (1974). Calibration will provide estimates of scale transformations of the marginal utilities, \( \beta_{jm} \) of equation (1).

Conventional logit models do not allow the coefficients to vary over individuals; that is, they assume that:

\[ \beta_{ijm} = \beta_{jm} \text{ for all } i \]  

If this is not true, we say that there is inter-personal taste variation. The taste variation may be purely random, or it may have a systematic nature; that is, the variation in tastes may be explained by reference to some additional variable. Suppose, for simplicity of illustration, that the coefficients \( \beta_{jm} \) have the following structure:

\[ \beta_{jm} = \begin{cases} \alpha_{om} + \alpha_{1m} Z_{im} & \text{for } j = 1 \\ \beta_{jm} & \text{otherwise} \end{cases} \]
DESIGN OF SP EXPERIMENTS AND TASTE VARIATIONS

T. Fowkes and M. Wardman

Then, substituting, we have:

\[ RU_{im} = (\alpha_{0m} + \alpha_{1m} Z_{im}) X_{i1m} + \sum_{j=2}^{\infty} \beta_{ijm} X_{ijm} + \epsilon_{im} \]  \hspace{1cm} (7)

so that, if the composite variable ZX is not included in the model, there will be specification error. An example might arise where the coefficient for a “cost” or “income” variable, representing the marginal utility of money to the individual, varied over individuals inversely with their income level. Here \( Z_{im} \) would be income (in this case a “generic” rather than “mode specific” variable, so that the \( m \) subscript is not required) and \( \alpha_1 \) would be negative. At the estimation stage, attempts should be made to uncover such systematic taste variation, either by inserting into the model dummy variable attribute coefficients for separate income groups, or by inserting cross-product terms ZX. The former method is equivalent to segmenting the data by income groups. During the design simulation stage it will often be useful to test the adequacy of the design in the presence of strong systematic variation in taste.

An alternative form of taste variation is Random taste variation, where the \( \beta_{ijm} \) are randomly distributed with means \( \beta_{ijm} \) so that:

\[ \beta_{ijm} = \beta_{ijm} + \epsilon_{ijm} \]  \hspace{1cm} (8)

so that, from (1) and (2):

\[ RU_{im} = \sum_j \beta_{ijm} X_{ijm} + \sum_j \epsilon_{ijm} X_{ijm} + \epsilon_{im} \]  \hspace{1cm} (9)

If tastes vary randomly in this way, the assumptions of the standard logit calibration are not satisfied, and the error terms are no longer distributed independently of the explanatory variables. Thus the coefficient estimates from standard logit calibration will be biased.

The multinomial probit model can also be derived from equation (3) by assuming the errors to have a multivariate Normal distribution (Hausman and Wise, 1978). This model allows for inter-personal taste variations – with the proviso, which applies also when the logit model is used in this way (Daly and Zachary, 1975), that some distribution of tastes must be specified. These random coefficient models are less restrictive than the standard fixed-coefficient logit model, and empirical studies have found statistical support for using random coefficient rather than fixed coefficient models (Daly and Zachary, 1975; Fischer and Nagin, 1981; Hausman and Wise, 1978).

To go one step further: in principle, separate models can be calibrated for each individual when SP data are available, given a sufficient number of choices per individual. This has the advantage, over both random coefficient and fixed coefficient models, that tastes are allowed to vary freely across individuals and no a priori restrictions are placed on the distribution of tastes.

The effect of variations in taste on estimation of relative valuations

Using the special case of linear utility functions, Horowitz (1980) concluded on the basis of simulations that, in the presence of taste variation, “Multinomial logit can give highly erroneous estimates of the choice probabilities of multinomial probit models. However, logit models appear to give asymptotically
accurate estimates of the ratios of the coefficients of the systematic components of probit utility functions, even when the logit choice probabilities differ greatly from the probit ones”.

We have previously investigated taste variations (Fowkes and Wardman, 1985). We found that a limitation of Horowitz (1980) was that he assumed zero correlation between the parameter weights which varied over individuals. The question of correlation was very important to us, since we were concerned to estimate values of time, which are derived as the ratio of the estimates of time and cost parameters. It was our view that individuals’ parameters for time and cost, reflecting the marginal utility of each, were very likely to be correlated. In particular, we felt that the marginal utility of time would tend to increase with income, but the marginal utility of money would tend to fall as income increased, and so there would be negative correlation between the time and cost parameters. We undertook some simulations to test the effect of this and some other assumptions concerning taste variation. The simulation results are briefly reported below.

The “ratio of means” problem

Another problem due to taste variation arises because the estimation produces estimates of time and cost parameters that are averages over the sample, and they are then used in ratio form to give the value of time. The true value of time would be the average over the sample of individuals’ values of time, these values being the ratios of their individual time and cost coefficients. It is not difficult to demonstrate mathematically that the appropriate mean of the ratios estimate is not necessarily equal to the ratio of the means estimate.

\[
\frac{\frac{1}{n} \sum_{i=1}^{n} a_i}{\frac{1}{n} \sum_{i=1}^{n} b_i} \neq \frac{1}{n} \left( \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \right)
\]

(e.g. \( n = 2, a_1 = 5, a_2 = 9, b_1 = 1, b_2 = 3 \) gives \( 4 \neq 3.5 \))

unless \( b_i = b \) for all \( i \)

or \( b_i = ka_i \) for all \( i \)

Our simulations showed that this “ratio of the means” problem could be very serious where individuals’ time and cost parameters were negatively correlated. A probit model would not overcome this problem.

Simulation results

Table 1 presents some of the simulation results. These were restricted to the straightforward binary choice context, using simulated data sets of 1,600 observations of individuals’ discrete travel choices. Only time and cost attributes were considered together with an alternative-specific constant. The objective of each of the four tests was to return the prespecified population (true) value of time shown in column 1. Because of the “ratio of means” problem, the best that could be hoped for was that the figure in column 2 might be returned, so we referred to this as our “Target”. For all four tests the average of the population
### TABLE 1

**Value of Time Estimation with Taste Variation**

<table>
<thead>
<tr>
<th>Test Number</th>
<th>1 (Mean of Ratios)</th>
<th>2 (Ratio of Means)</th>
<th>3 (Estimated VOT)</th>
<th>4 Standard Error of ( \hat{V} )</th>
<th>5 Average Rho-Bar Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>( V^T )</td>
<td>( \hat{V} )</td>
<td>( SE(\hat{V}) )</td>
<td>( \hat{p}^2 )</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10.00</td>
<td>10.00</td>
<td>10.02</td>
<td>0.038</td>
<td>0.63</td>
</tr>
<tr>
<td>2</td>
<td>10.00</td>
<td>10.00</td>
<td>9.63</td>
<td>0.038</td>
<td>0.45</td>
</tr>
<tr>
<td>3</td>
<td>10.07</td>
<td>10.00</td>
<td>9.95</td>
<td>0.048</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>12.22</td>
<td>10.00</td>
<td>9.86</td>
<td>0.048</td>
<td>0.23</td>
</tr>
</tbody>
</table>

\[
V = \frac{1}{1600} \sum_{i=1}^{1600} (\beta_{ti}/\beta_{ci})
\]

where
- \( i \) is the \( i \)th person in the 1600
- \( t \) denotes time
- \( c \) denotes cost

so that
- \( \beta_{ti} \) is the time coefficient for person \( i \)
- \( \beta_{ci} \) is the cost coefficient for person \( i \)

\[
V^T = (\sum_{i=1}^{50} (\beta_{ti}/\beta_{ci})) / (\sum_{i=1}^{50} \beta_{ci}), \text{ i.e. the ratio of the means of individuals' coefficients, the } 1/1600's \text{ having cancelled}
\]

\[
\hat{V} = \frac{1}{50} \sum_{j=1}^{50} (\hat{\beta}_{tj}/\hat{\beta}_{cj}), \text{ where } j \text{ denotes the } j \text{th run from the 50 replications done}
\]

\[
SE(\hat{V}) = \sqrt{\frac{\sum_{j=1}^{50} (\hat{V}_j - \hat{V})^2}{49 \times 50}}, \text{ i.e. the standard error of the mean value of the set of the 50 estimated values of time}
\]

where
- \( \hat{V}_j = (\hat{\beta}_{tj}/\hat{\beta}_{cj}) \), the ratio of the time and cost coefficients on run \( j \).

Time coefficients was ten times the average of the population cost coefficients. Thus our Target value of time is always 10.

Coefficient values were varied over individuals, and a random number generator was used to specify errors from a Weibull distribution. Individuals were assumed to possess linear compensatory utility functions, and were assigned to that travel alternative which had for them the greatest random utility. Coefficient estimates
were obtained from the BLOGIT program (Crittle and Johnson, 1980). Each test took the average of 50 repeat runs.

Test 1 assumed there was no taste variation, and so all assumptions were in line with those of the conventional logit model. Table 1 shows that for Test 1 the mean of the value of time estimates derived from the 50 repeat runs was 10.02, which is close to the Target value of 10. The sampling distribution of the sample mean from a sample of size 50 can be taken to be approximately Normal with a standard error approximately equal to the standard deviation of the 50 values, divided by the square root of 50. These are the values reported in column 4. A 95 per cent Confidence Interval for $\hat{V}$ can be formed by going two standard errors either side of $\hat{V}$ — that is, $10.02 \pm 0.076$. This can be seen to incorporate $V_T$, so we can conclude that the logit estimation is returning unbiased estimates.

Test 2 introduced taste variation for the time coefficient only. This would represent the (somewhat unlikely) case where the marginal utility of money was broadly constant over our population of interest, but an individual’s marginal utility of time varied because of differing time constraints depending on, say, whether or not the person worked full time, worked part time, had to look after children, was unemployed, etc. Since cost, the denominator of the “value of time” ratio, is constant, the ratio of means does equal the mean of the ratios. Thus, for Test 2, both the True VOT and the Target VOT are 10. The average value from the logit estimation (over the 50 runs) was 9.63, again with a standard error of 0.038. This time the Target and estimate are about 10 standard errors apart, which is much more than could be accounted for by chance. Hence we conclude that under these circumstances the logit estimates are biased.

Test 3 varied both the time and cost coefficients, so that they were positively correlated. This represented a situation where the people hardest pressed for money were also the hardest pressed for time. This might be thought to arise because richer people were able to buy themselves extra “time” by purchasing time-saving goods and services. We have not kept the coefficients in exact ratio, which would have given everybody a constant value of time, so the ratio of means problem does come into play. However, the positive correlation is such that the True value of time for Test 3 only differed slightly from our Target (10.07 as against 10.00). The logit estimation returned an average estimate of 9.95. This had a higher standard error of 0.048, giving an indication that all was not well. However, using logit we have made a good estimate of the target; this is within the 95 per cent confidence interval. Note, though, that the True VOT is not inside the 95 per cent confidence limit ($9.95 \pm 0.096$).

Test 4 again varied both coefficients, but this time they were negatively correlated. This corresponds to what we consider the most likely form of taste variation, namely, that richer people tend to be more pressed for time while poorer people tend to have more “time on their hands”, as poorer people are less able to afford the costs of leisure time activities and tend not to have full-time employment. The “ratio of the means” problem now has its most serious effect. As in all the tests, the sum (and hence the average) of the time coefficients is ten times the sum of the cost coefficients, giving a Target VOT of 10. However, the mean of the ratios for our particular set of negatively correlated coefficients was 12.22. The logit estimation returned an estimate very close to the Target, but not
containing it in the 95 per cent confidence interval. Our conclusion (Fowkes and Wardman, 1985) was that the extra computational difficulty of using a probit estimation was not worth the gain, given the potentially much larger error due to the “ratio of the means” problem in the context of “value of time” estimation. We recommended that, as far as practicable, data sets should be segmented so that separate coefficients were calibrated for groups of individuals with reasonably homogeneous values of time, even if these values arose as the ratios of markedly different time and cost coefficients.

EXPERIMENTAL DESIGN

This section will present recommendations for the SP design. It is our view that one of the most important and most neglected requirements of SP experimental design is that the trade-offs presented cover a sufficiently wide variation in tastes across individuals. This entails that the implied trade-offs presented must be carefully checked in the light of the foregoing discussion of taste variation.

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.

Attribute levels and values

SP experiments are typically based on attributes entering only at a few levels. A practical constraint is that the exercise must be kept within manageable proportions; but the number of evaluations to be made by the individual generally increases as the number of levels in the factorial design increases. The problem worsens as the number of attributes increases, or if it is desired to consider interaction effects, or if main effects are to be independent of interaction effects.

The attribute values must be chosen to be reasonably realistic. As the values increasingly diverge from individuals' experiences or from what appears plausible, the SP responses can be expected to become less reliable. Where the individuals to be surveyed have widely different experiences, attempts can be made to "customise" the design, for example, by using designs with different values for subsets of the sample. Such customising can also be achieved by basing the SP experiment on changes from individuals' current circumstances (Andersen et al., 1986) or by using computer interactive methods in the fieldwork (Ampt et al., 1986; Jones et al., 1986).1 Computer interactive methods have been found to increase the reliability of responses (MacBride and Johnson, 1980).

1 See also the paper by Bradley in this issue.

33
Combinations of attribute levels

Less consideration appears to have been given to the combination of attribute values. When factorial designs are used, there is the chance that attribute values can be combined in an unrealistic manner, and this may influence the quality of the responses supplied. Hensher and Truong (1983) have questioned the use of designs which remove individuals from what they term "experientially meaningful configurations". Where there are unavoidable correlations between attributes, the use of orthogonal designs may not be appropriate.\(^2\)

On one occasion, we had to design an SP experiment for motorists' route choices in which each route was characterised by its cruising speed, its distance and the cost of petrol used. It soon became apparent that an orthogonal design would be inappropriate because of the inherent physical relationships between the three variables. Even though the scenarios to be evaluated are hypothetical, perceptual problems will be encountered if individuals are required to abstract from reality and consider attributes which are combined without due account of the relationships which must exist.

The SP responses may also become less reliable if attribute levels are combined in a manner which is unrealistic to individuals, even though there is no inherent physical relationship which must be observed. Individuals may not be able to perceive train as being cheaper than coach for long-distance journeys, or in some instances a reduction in bus fares may be treated as highly implausible. In such cases, responses may be adjusted to discount options which are considered to be infeasible.

In an actual survey of motorists' route choices (Wardman, 1986), the SP experiment described two routes which motorists could use in terms of travel time, toll charge and petrol costs. At the pilot stage, respondents were contacted to establish their opinions on the questionnaire and on the SP experiment. Some respondents commented on what appeared to them to be unrealistic situations, for example, where the shorter route had a somewhat longer travel time and higher petrol costs. The design was amended to avoid these unrealistic combinations as far as possible, while also not creating unduly high correlations between attributes.

Small variations in attributes

In some choice contexts, the variations in certain attributes across alternatives may be small. If the SP design presents such small attribute differences, they may be ignored by some respondents, if other attributes do vary between the options to be compared. Thus the attributes with the small variations will not have their true influence upon choice, and this can be a seriously distorting influence on the relative valuations obtained. This appeared to be a problem for walk and wait time in an analysis of the SP responses of commuters for the choice between train and coach in North Kent (see the paper by Wardman in this issue). Careful design can ensure that the trade-offs are clearly offered. In a pairwise comparison, it may

\(^2\) This problem is further discussed in section 3.1.2 of the paper by Louviere in this issue.
be necessary for the levels of some attributes to be the same for both options. This focuses individuals' attentions on the remaining attributes which vary across options. Even where variations in attribute values are not particularly small, there are still attractions in restricting the number of attribute variations in this way, as the comparisons to be undertaken by the individual are simpler than if all attributes vary between options.

A similarity may be detected between the argument advanced here, that small attribute differences may be ignored by respondents, and the argument that "small time savings" should be valued by planners at a lower rate (possibly zero), as they are of less use than larger ones and quite possibly may not be perceived at all. Our position is that we support the conventional practice of valuing all time savings at an equal rate, irrespective of their size. We imagine travellers having a uniform distribution of unusable small amounts of time before receiving their travel time saving. This means that some people get no benefit, while others are pushed over their "threshold" and get larger benefits. Mathematically, the result is the same as if the travel time savings had all been valued at a common rate, as is shown in the Appendix. However, when we present people with hypothetical SP choices involving very small time savings, we surely cannot expect them to assume they have amounts of unused time waiting to be made up to usable amounts. This must be true particularly if we accept the argument that these small amounts of time are not even perceived. If we ask our respondents to put themselves in a situation as close as possible to that of an actual travel choice, we may get some people who do take full account of any unused time they may have; but we suspect that these people are a minority and that the resulting estimated values are corrupted. To avoid this, it is necessary to ensure that attribute differences are not trivially small.

**Boundary or equi-utility values**

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:

\[
U_m = \beta_c \text{COST}_m + \beta_x X_m
\]  

(11)

where \text{COST} is the monetary cost, say in pence, and \( X \) is some attribute of mode
Let us now define a boundary relative valuation of $X$ in terms of money as

$$B(X:COST) = (COST_1 - COST_2)/(X_2 - X_1)$$

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. If an orthogonal design is used together with realistic attribute values, possibly from a pilot survey, it may well be that the informational content of the responses is much less than it could be. This is because the design does not necessarily contain a satisfactory range of boundary values.

Often, we will have a third attribute varying, say $Y$, that is,

$$U_m = \beta_x COST_m + \beta_x X_m + \beta_y Y_m$$

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

$$\beta_y = k \beta_x$$

so that we feel able to choose for test purposes low, medium and high values for the unknown $k$. Generalising (12), we have

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

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
hypothesized 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 alternatives. 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:

\[
\begin{align*}
U_1 &= ASC + \beta_c COST_1 + \beta_x X_1 + \beta_y Y_1 \\
U_2 &= \beta_c COST_2 + \beta_x X_2 + \beta_y Y_2
\end{align*}
\]

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:

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

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

In (17) the term \((ASC/\beta_c)\) is merely the ASC expressed in monetary units, while in (18) the term \((ASC/\beta_x)\) 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.

The difficulty of incorporating a satisfactory range of equi-utility relative valuations varies across choice contexts. Whereas values of time can be expected to lie in a relatively narrow range, the range of values can be reasonably expected to be much larger for the analysis of individuals’ option values for local public transport services. It would be beneficial in such cases to try to “customise” the SP experiment. Screening questions, perhaps based on willingness to pay, could be used to identify a likely range of relative valuations to which the more sophisticated SP technique can be applied. Computer interactive surveying methods are potentially of considerable benefit in tailoring the boundary values of the design.
In-built tests of SP responses

The design of the SP experiment can contain some means by which the quality of the responses supplied can be assessed. In-built tests examine the consistency and rationality of the responses supplied. The most straightforward test, commonplace when individuals are required to rank options in order of preference, consists of having alternatives which are logically superior or inferior to all others; that is, they contain attributes which are all at their “best” or “worst” levels. Those who have not understood the exercise, not taken it seriously, or found it too difficult, are unlikely to place these options at the appropriate extreme of their ranking, and so the proportion of unsuitable respondents can be greatly reduced by removing those who fail this test.

A more rigorous test of rationality can be conducted by considering the relative valuations implied by the choices each individual makes. For example, the design could incorporate trade-offs to check whether an individual who valued time savings at more than five pence per minute on one choice also valued them at more than three pence per minute on a second choice. Those who have intransitive preferences can be omitted from the analysis. Some “errors” in responses are permissible because of the stochastic element within random utility models, but it may well be possible to improve the models developed by omitting respondents whose SP responses are considered to contain serious irrationalities.

A further test is to consider whether an individual’s responses follow some systematic pattern, such as may stem from non-compensatory decision making or from always preferring the same option in the SP experiment as is preferred in practice through habit or justification bias. Such behaviour is inconsistent with the assumptions of conventional random utility models. Policy response bias will give rise to misleading responses, which may not be detected by rationality tests: for example, if the quicker option is always preferred in an attempt to influence policy regarding journey time. Though it is possible to omit respondents erroneously on the basis of these considerations (for example, always choosing the cheapest or quickest option may be the result of a very low or high value of time), the design can reduce the chance of this if some choices are designed with implausibly high or low boundary values of time. Identifying and omitting responses which contain serious error or which appear inconsistent with the models to be used can lead to worthwhile improvements in the models developed.

In an analysis of motorists’ route choices (Wardman, 1986), respondents were omitted on this reasoning from the data sets for the final models on commuting, leisure travel and employers’ business trips. The average effect of this on estimated values of time was only about 1 per cent; yet the t ratios of the value of time estimates were increased by around 10 per cent on average, even though the samples were 13 per cent smaller. For SP models of commuters’ mode choices (see paper by Wardman in this issue) there was a 10 per cent change on average in the value of time estimates when the sample was reduced by 25 per cent, but the t ratios of the values of time increased by 25 per cent on average. We should therefore attempt to ensure that our SP design, supported by other questions about current actual behaviour, will permit some assessment of the quality and the nature of the SP responses supplied.
TESTING THE EXPERIMENTAL DESIGN

Some of the different aspects considered above are complementary. Restricting the number of attributes which vary between alternatives makes it easier for tests of the rationality of the responses to be built into the design, and also simplifies both the task required of the respondent and the incorporation of a satisfactory range of equi-utility relative valuations. However, even if the SP experiment is designed according to the principles outlined, it is still necessary to test the experimental design by simulating the choice process, using synthetic data and then applying the chosen technique for estimation. Choices are based on the random utilities of options, which in turn are a function of the attribute values of the design, relative valuations and an error component. The simulations test whether the design can accurately recover a reasonable range of relative valuations; in this way any shortcomings in the design can be identified and corrected.

Segmentation analysis

Unless we have some way of identifying segments before administering the SP experiment, we must ensure that the SP design can recover relative values from all sections of the population to be sampled. Thus we should allow for taste variation within our synthetic data, though the design will be acceptable provided only that parameter values can be accurately recovered from post-sample segmented subsets. To see this point, suppose that the population can be assumed to be divided into two groups, having constant parameters within groups but differing between groups. Provided we assume that we can allocate respondents to each group from their questionnaire responses, the design will merely have to be able to return the two parameter sets separately in two segmented estimations. All respondents will, however, face the same SP design, so this must have boundary values that can cope with the parameters from both groups.

The "ratio of the means" problem arises in the application of conventional random utility models where individuals' responses are combined into a single model. One way of overcoming the problem is to calibrate models for each individual separately and estimate an average relative valuation as the mean of each individual’s ratio of coefficients. This can be compared with the potentially erroneous estimate derived as the ratio of the mean estimates for all individuals combined in one model. Such a comparison (Wardman, 1986) found the differences between values derived by the two methods to be small and non-significant, though the potential for a discrepancy remains. However, a large number of evaluations will need to be undertaken by individuals if robust estimates for all coefficients are to be obtained for each individual, and some respondents may be unwilling to make the effort; this imparts a possible bias.

If individuals' SP responses are to be combined into a single model, the degree of error due to the "ratio of the means" problem can be reduced by allowing different marginal utilities across segments of the sample obtained. Variations in the marginal utilities which are purely random cannot be captured by a segmentation process, but those variations in tastes which are systematically related to some variable, such as variations in the marginal utility of money due to variations in income, can be allowed for by using a segmentation approach.
This may remove the worst excesses of the problem.
If the marginal utility of money varies across individuals according to their income, the segmentation of the cost variable can be specified as:

\[
\sum_{y=1}^{Y} \beta_{ey} d_y \text{COST}
\]

(19)

where \(d_y\) is a dummy variable for each of the \(Y\) income categories to be considered. A separate cost coefficient is then estimated for each of the income groups. The cost attribute can also be segmented by additional factors. Other attributes can also be segmented by appropriate socio-economic variables, though the "ratio of the means" problem arises because of variations in the denominator (cost) term, so the segmentation of cost is of greater importance.

**Random error**

At the present stage of availability of computer software, we are not proposing that researchers do other than use logit analysis for estimation. However, researchers should include in the synthetic data sets the form of error they believe to be most reasonable. In most cases Normal errors will be preferable, but even simpler errors will suffice. The size of the error in relation to representative utility will determine the strength of the effect of attribute variations on choice. Consequently, in the presence of large errors the parameter estimates will be very poorly recovered. It may well be worth while to test the design to see at what size of error the estimated parameters become unacceptable, and to consider whether this is clearly greater than can reasonably be expected to occur in real data. For given parameters, this will be more important when small attribute differences are presented. If the errors used are too small in relation to the deterministic part, the estimation of discrete choice models will fail. Larger errors should then be used.

**Computer simulation**

In order to show clearly how our recommendations can be implemented on a computer, we present the following very simple exercise. We first decide on the number of synthetic respondents on which to test the design; it should preferably be similar to the sample size expected. We set up a loop for that number of repetitions. Let us suppose we wish to model taste variations very simply with high and low values for each of the two coefficients, combined with negative correlation. Then we might have something like:

\[
R = \text{Random Number in the range 0 to 1}
\]

If \(0.0 \leq R < 0.2\) Param 1 = High  Param 2 = High
If \(0.2 \leq R < 0.5\) Param 1 = High  Param 2 = Low
If \(0.5 \leq R < 0.8\) Param 1 = Low  Param 2 = High
If \(0.8 \leq R \leq 1.0\) Param 1 = Low  Param 2 = Low

For example, parameters could be \(\beta_x\) and \(\beta_y\), having low values \(-2\) and \(-3.2\) respectively, and high values \(-4\) and \(-9.6\) respectively.
Table 2

Decisions for the Computer Simulation Example, for Persons having Differing Relative Valuations of Y in terms of X

<table>
<thead>
<tr>
<th>Relative Value of Y in terms of X</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>Option 1</td>
</tr>
<tr>
<td>2.5</td>
<td>Option 2</td>
</tr>
<tr>
<td>3.5</td>
<td>Option 2</td>
</tr>
</tbody>
</table>

Hypothetical choices are:

Choice (1) Option 1: X = 4, Y = 7
Choice (2) Option 2: X = 6, Y = 6
Choice (2) Option 1: X = 9, Y = 3
Choice (2) Option 2: X = 3, Y = 5

The boundary values for Y in terms of units of X are:

Choice (1) \( B(Y: X) = \frac{(X_1 - X_2)}{(Y_2 - Y_1)} \)
\( = \frac{-2}{-1} = 2 \)

Choice (2) \( B(Y: X) = \frac{6}{2} = 3 \)

Table 2 shows the decisions, in the absence of error terms, for persons having relative valuations of Y in terms of X of 1.5, 2.5 and 3.5. We can see that the boundary values tell us when people will switch between options. Table 3 shows inferences arising from boundary values at 2 and 3.

For each of our synthetic respondents with randomly generated parameters, \( \beta \), we can calculate the utility of each option, \( i \), as:

\[ U_i = \beta_x X_i + \beta_y Y_i + \epsilon_i \]  \hspace{1cm} (20)

Table 4 generates six synthetic responses, showing the choices made, for clarity, in the absence of an error term. We can see that error terms of the magnitude of about 10 would completely dominate the deterministic part, whereas error terms of less than 0.4 would have no effect on choices. This suggests that we might take error terms, \( \epsilon_i \), normally distributed with mean zero and standard deviation 2. The synthetic data set can then be analysed to see whether the known parameter values can be adequately recovered. If not, either the SP design can be modified or the sample size can be increased.
TABLE 3

Inferences of Relative Valuations of Y in terms of X, drawn from Choices in the Computer Simulation Example

<table>
<thead>
<tr>
<th>Choice (1)</th>
<th>Choice (2)</th>
<th>Relative Valuation of Y in terms of X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>Less than 2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Between 2 and 3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Greater than 3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Irrational</td>
</tr>
</tbody>
</table>

TABLE 4

Synthetic Responses generated from the Computer Simulation Example

<table>
<thead>
<tr>
<th>Random No.</th>
<th>Parameters</th>
<th>Choice (1)</th>
<th>Choice (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_x$</td>
<td>$\beta_y$</td>
<td>$U_1$</td>
</tr>
<tr>
<td>0.47</td>
<td>-4</td>
<td>-3.2</td>
<td>-38.4</td>
</tr>
<tr>
<td>0.09</td>
<td>-4</td>
<td>-9.6</td>
<td>-83.2</td>
</tr>
<tr>
<td>0.56</td>
<td>-2</td>
<td>-9.6</td>
<td>-75.2</td>
</tr>
<tr>
<td>0.78</td>
<td>-2</td>
<td>-9.6</td>
<td>-75.2</td>
</tr>
<tr>
<td>0.14</td>
<td>-4</td>
<td>-9.6</td>
<td>-83.2</td>
</tr>
<tr>
<td>0.97</td>
<td>-2</td>
<td>-3.2</td>
<td>-30.4</td>
</tr>
</tbody>
</table>

CONCLUSIONS

There are several aspects of Stated Preference experimental design, and some of the issues involved appear to us to have been neglected. This paper has offered advice on how to evaluate proposed designs. We have taken it as axiomatic that in this evaluation the presence of inter-personal taste variation should be allowed for. We have discussed the implications of taste variation for conventional logit modelling, as well as for experimental design. On the first point we have reported simulation work that suggests that, while it is not worth using a multinomial probit model (because of the “ratio of means” problem), one should perform logit analysis only on segments of the sample that can be taken to have reasonably uniform values of the relative valuation of interest. On the second point, it was seen that the experimental design should contain a much wider variation of what
we have termed "boundary relative valuations" than would be required merely to allow for uncertainty in the final overall value expected. Designs should have adequate spreads of these boundary values for all plausible valuations of the attributes concerned. It is important to test this both by inspecting the boundary values for a wide range of eventualities, and by simulation. Amendments should be made to improve the SP design as appropriate, even if this entails a movement away from an orthogonal design. Care should, however, be taken to ensure that the hypothetical choices presented are meaningful, and that variations in attribute levels between alternatives are not so small as to invite being ignored.

APPENDIX

Valuation of Small Time Savings

Let the threshold be $T$ and the initial amount of unusable time be $r$, distributed uniformly as:

$$f(r) = \begin{cases} 1/T & \text{for } 0 < r < T \\ 0 & \text{otherwise} \end{cases}$$

Let the value of "usable" time be $v$.

The effect of giving everyone an additional travel time saving, $t$, is to give no benefit to those with $r + t < T$, but to give a usable travel time saving $T$ to those with $r + t \geq T$, that is,

$$\text{Value per person} = 0 \int_0^{T-t} f(r)dr + v T \int_{T-t}^T f(r)dr$$

$$= vT \left[ T \right]_T^{T-t}$$

$$= vt$$

which is just valuing the time savings, $t$, all at rate $v$.

REFERENCES


