

# Some lessons in stated choice survey design

**Stephane Hess and John M. Rose**

*Stephane Hess*

University of Leeds

[S.Hess@its.leeds.ac.uk](mailto:S.Hess@its.leeds.ac.uk)

*John M. Rose\**

The University of Sydney

[johnr@itls.usyd.edu.au](mailto:johnr@itls.usyd.edu.au)

## Abstract

A growing majority of discrete choice studies are now based on data collected through stated preference (SP) surveys, primarily in the form of stated choice (SC) questionnaires. The state-of-the-art in this area has evolved dramatically over recent years, as witnessed in a burgeoning literature. At the same time however, the state-of-practice has stagnated, especially in some countries. Additionally, the growing emphasis on theoretical developments, primarily to do with efficiency, has meant that a number of fundamental issues are often no longer talked about. In the present paper, we look in detail at the entire process going from initial survey planning to actual data collection, discuss, often with examples, a number of common but avoidable mistakes, and provide some guidance for good practice.

## 1. Introduction

Over recent years, there has been a hype of activity in the field of experimental design for stated preference (SP) surveys, leading to a move away from orthogonal design techniques to efficient design techniques. The advantage of these design techniques in a practical context is that more robust results can be obtained with smaller sample sizes, potentially leading to significant financial savings, especially with surveys involving face to face interviews.

Whilst these developments represent theoretical advancements that are gradually making their way into applied research, the literature as a whole appears to have largely focused on these advances to the neglect of more fundamental issues. The design and implementation of surveys for the collection of choice data is an in-depth process that adds to the already existing complexities of more traditional questionnaire construction and data collection. It is to these issues that the present paper returns, specifically dealing with the basic principles of good practice in the field of survey design.

The topics covered in this paper are survey technique, survey context, choice set design, experimental design, survey testing, and survey administration. For each topic, we discuss

the basic issues, highlight possible mistakes, often with the help of examples, and provide some guidance for good practice.

## 2. Survey technique

The majority of surveys looking at hypothetical scenarios are of the stated choice (SC) type, in which a respondent is faced with a choice between a finite number of mutually exclusive alternatives. Figure 1 shows an example of this type of response format for an unlabelled route choice experiment. It is this type of experiment that the majority of this paper focuses on.

	Route A	Route B	Route C	Route D	Route E
Petrol Costs	\$3.00	\$4.00	\$4.00	\$5.00	\$2.50
Toll Cost	\$2.00	\$3.00	\$2.00	\$2.00	\$3.00
Prob. of arriving late	0.5	0.4	0.3	0.1	0.4
Prob. of arriving early	0.1	0.2	0.2	0.1	0.2
Free Flow Time	10 mins	10 mins	10 mins	15 mins	20 mins
Congested Time	20 mins	15 mins	10 mins	5 mins	5 mins
Egress Time	10 mins	5 mins	10 mins	5 mins	10 mins
I would choose	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 1: Example of a typical choice response format

Nevertheless, it should be acknowledged that there exist a number of alternative (or in some cases complimentary approaches) that we will now touch on briefly. Indeed, the choice of SC should not be an automatic one, once the analyst has settled on SP methods rather than RP methods.

### 2.1 Rating and ranking

Rather than have respondents choose their single most preferred alternative, some researchers prefer to have respondents rate each alternative on some form of scale. Typically used rating scales involve respondents having to rate each alternative from 1 to 10 or 1 to 100 where higher values represent higher degrees of preference for that alternative. An example of this type of response method is shown in Figure 2. Some researchers prefer respondents to rank (with or without allowing for ties) each alternative in order of preference. An example of this is shown in Figure 3.

	Route A	Route B	Route C	Route D	Route E
Petrol Costs	\$3.00	\$4.00	\$4.00	\$5.00	\$2.50
Toll Cost	\$2.00	\$3.00	\$2.00	\$2.00	\$3.00
Prob. of arriving late	0.5	0.4	0.3	0.1	0.4
Prob. of arriving early	0.1	0.2	0.2	0.1	0.2
Free Flow Time	10 mins	10 mins	10 mins	15 mins	20 mins
Congested Time	20 mins	15 mins	10 mins	5 mins	5 mins
Egress Time	10 mins	5 mins	10 mins	5 mins	10 mins
From 1 to 10, rate each of the following routes (1 being the worst rating, 10 being the best)	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Figure 2: Example of a typical conjoint ratings response format

Together, ranking and rating type response data combine to form a single SP methodology referred to in some literature as traditional conjoint analysis or simply conjoint analysis. This differs to choice type responses (which for historical reasons is referred to as choice based conjoint in the same

literature) in a number of important ways. Firstly, the analysis used for such data typically relies on linear regression models as opposed to non-linear logit or probit type models often employed for choice data. Although ordered discrete choice models may also be used on such data, the literature dealing with traditional conjoint methods typically ignore such models in favour of linear regression models. This is because such literature usually seek to derive individual specific models as opposed to a model estimating the population ‘average’ parameters. This has proven somewhat controversial given that linear regression models assume interval or ratio scaled data for the dependent variable, with debate raging as to whether rating or rankings data meet this criterion. Secondly, the response metric (and in particular ratings scales) has also proven somewhat controversial from a psychological perspective with many researchers questioning whether different respondents assign the same psychological value to the values of the scale (i.e., does a rating of 4 on a 1 to 10 point scale have the same meaning to 2 different respondents). Discrete choices do not suffer from this issue. Nevertheless, ratings and rankings tasks offer two significant advantages over discrete choice tasks; firstly they provide full information on the relative preferences of all alternatives unlike choice which informs the analyst only what is the most preference option, and secondly, ranking of responses may allow the analyst to rank explode the data which may provide more observations per respondent from which to model with.

	Route A	Route B	Route C	Route D	Route E
<b>Petrol Costs</b>	\$3.00	\$4.00	\$4.00	\$5.00	\$2.50
<b>Toll Cost</b>	\$2.00	\$3.00	\$2.00	\$2.00	\$3.00
<b>Prob. of arriving late</b>	0.5	0.4	0.3	0.1	0.4
<b>Prob. of arriving early</b>	0.1	0.2	0.2	0.1	0.2
<b>Free Flow Time</b>	10 mins	10 mins	10 mins	15 mins	20 mins
<b>Congested Time</b>	20 mins	15 mins	10 mins	5 mins	5 mins
<b>Egress Time</b>	10 mins	5 mins	10 mins	5 mins	10 mins
<b>Please rank each route in order of preference (1 being most preferred)</b>	<input type="text"/> 1 2 3 4 5	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Figure 3: Example of a typical conjoint rankings response format

## 2.2 Best worst and best worst scaling

Over time, two different forms of best worst response formats have come into existence. The most recent format, as shown in Figures 4a and b, require that respondents indicate what they consider to be the best and worst alternatives from amongst those on offer. Whilst providing similar data to pure rankings data, those advocating this approach suggest that respondents are better equipped psychologically to distinguish between their most preferred and least preferred alternatives than they are to rank all alternatives, particularly when presented with large number of alternatives to consider. Where more than three alternatives are shown in any single choice task, requiring respondents to suggest only the best and worst alternatives will result in partial preference rankings of the alternatives (see Figure 4a). If the analyst wishes to obtain a full ranking of the alternatives, subsequent response tasks involving the remaining alternatives can be used (see Figure 4b). See Louviere et al. 2008 for a more detailed review of this response format.

	Route A	Route B	Route C	Route D	Route E
Petrol Costs	\$3.00	\$4.00	\$4.00	\$5.00	\$2.50
Toll Cost	\$2.00	\$3.00	\$2.00	\$2.00	\$3.00
Prob. of arriving late	0.5	0.4	0.3	0.1	0.4
Prob. of arriving early	0.1	0.2	0.2	0.1	0.2
Free Flow Time	10 mins	10 mins	10 mins	15 mins	20 mins
Congested Time	20 mins	15 mins	10 mins	5 mins	5 mins
Egress Time	10 mins	5 mins	10 mins	5 mins	10 mins
Please select the single best and worst routes	Best Worst	Best Worst	Best Worst	Best Worst	Best Worst

Figure 4a: Partial ranking best worst response format

	Route A	Route B	Route C	Route D	Route E
Petrol Costs	\$3.00	\$4.00	\$4.00	\$5.00	\$2.50
Toll Cost	\$2.00	\$3.00	\$2.00	\$2.00	\$3.00
Prob. of arriving late	0.5	0.4	0.3	0.1	0.4
Prob. of arriving early	0.1	0.2	0.2	0.1	0.2
Free Flow Time	10 mins	10 mins	10 mins	15 mins	20 mins
Congested Time	20 mins	15 mins	10 mins	5 mins	5 mins
Egress Time	10 mins	5 mins	10 mins	5 mins	10 mins
Please select the single best and worst routes	Best Worst	Best Worst	Best Worst	Best Worst	Best Worst
Please select the next best and worst routes	Best Worst	Best Worst	Best Worst	Best Worst	Best Worst

Figure 4b: Full ranking best worst response format

Originally proposed by Finn and Louviere (1992), best worst scaling offers yet another alternative response mechanism for collecting SP type data. Best worst scaling methods differ significantly to the other response methods in that the response mechanism is not active at the level of the alternatives, but rather at the level of the attributes (see Figure 5). Rather than present respondents with a number of alternatives to choose from amongst, the best worst scaling approach presents respondents with a single alternative and asks them to select the best and worst attribute for that alternative based on the attribute levels shown. The (log of the) frequency of times a particular pair of attributes is selected as the best and worst combination is then used as the dependent variable in a linear regression model to determine the desirability of each attribute and attribute level for different respondents.

	Best	Route	Worst
Petrol Costs	<input type="radio"/>	\$3.00	<input type="radio"/>
Toll Cost	<input type="radio"/>	\$2.00	<input type="radio"/>
Prob. of arriving late	<input type="radio"/>	0.5	<input type="radio"/>
Prob. of arriving early	<input type="radio"/>	0.1	<input type="radio"/>
Free Flow Time	<input type="radio"/>	10 mins	<input type="radio"/>
Congested Time	<input type="radio"/>	20 mins	<input type="radio"/>
Egress Time	<input type="radio"/>	10 mins	<input type="radio"/>

Figure 5: Best worst scaling response format

Proponents of best worst scaling point to two significant benefits in its use over more traditional SP response formats. Firstly, they argue that the traditional pick one choice responses, which represent the predominant data collection method used to date, are largely inefficient in terms of the amount of data obtained from the respondent. This criticism, whilst warranted, has been partially addressed

via the other response mechanisms outlined here. The second criticism relates mainly to the ability of discrete choice type models to untangle the base levels of categorical variables that are dummy coded from any estimated alternative specific constants (ASCs). Whilst effects or orthogonal coding overcomes this, there still remains a problem in interpreting the willingness to pay (WTP) values obtained for such categorical attributes. Such coding structures allow for a determination of the WTP values for the non-base levels however these values are estimated relative to the base level, the WTP value of which is not calculable. For example, consider a categorical attribute 'comfort' with levels, low, medium and high. Assuming this attribute is dummy coded with low as the base level, then further assuming the experiment contains a cost attribute, the WTP for the for the medium and high attribute levels can be determined. Unfortunately, these WTP values are relative to the base low level, the WTP for which is not known. Effects and orthogonal coding unconfound the base levels from the ASCs; however, interpretation of the WTP outputs remains equally problematic. The best worst scaling response format allow for the calculation of the WTP for all attribute levels, and hence is argued as being preferred if an experiment has non numeric attributes (see Marley and Louviere, 2005).

### 2.3 Frequency data

The final 'choice' based response mechanism that has been applied in the past involves respondents being assigned some value that they may then parcel out to the various alternatives. In transport studies, this typically involves respondents being asked to assign a number of trips to alternative routes as shown in Figure 6. Once collected, frequency data may then be converted to proportions which can then be estimated using the same statistical methods used for 'pick one' type choice data.

	Route A	Route B	Route C	Route D	Route E
<b>Petrol Costs</b>	\$3.00	\$4.00	\$4.00	\$5.00	\$2.50
<b>Toll Cost</b>	\$2.00	\$3.00	\$2.00	\$2.00	\$3.00
<b>Prob. of arriving late</b>	0.5	0.4	0.3	0.1	0.4
<b>Prob. of arriving early</b>	0.1	0.2	0.2	0.1	0.2
<b>Free Flow Time</b>	10 mins	10 mins	10 mins	15 mins	20 mins
<b>Congested Time</b>	20 mins	15 mins	10 mins	5 mins	5 mins
<b>Egress Time</b>	10 mins	5 mins	10 mins	5 mins	10 mins
<b>Assume you were to make the same trip 20 times. Please allocate the 20 trips to each of the 20 routes</b>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Figure 6: Frequency data

### 2.4 Transfer price / Contingent Valuation / WTP question

Another tool used especially in the context of applied work consists of so called *transfer price*, *contingent valuation* or *willingness to pay* questions. In the more blunt *willingness to pay* approach, a respondent is directly asked how much he/she would be willing to pay for a saving in travel time by a certain amount, while, in the slightly more refined *transfer pricing* approach, the respondent is asked whether he/she would be willing to pay a certain amount  $x$ , which is then gradually increased or reduced until the boundary willingness to pay (WTP) for that respondent is reached. Transfer price approaches are potentially affected by significant levels of strategic bias, and have come under considerable legal scrutiny in some countries, such as Australia<sup>1</sup>. In fact, it is the risk of strategic bias

<sup>1</sup> In a case dealing with the valuation of music, the Copyright Tribunal of Australia in 2007 ruled that transfer price methods were inappropriate and that discrete choice modelling was the preferred method for valuation studies (PPCA (the Nightclubs Matter) CT2/2004 [2007] ACopy1).

in such direct questioning approaches that is one of the motivations behind using multi-attribute hypothetical choice scenarios, masking at least up to a degree the true aim of the research. Even though, in some work, the results from the transfer price exercise seem to be largely consistent with those from the traditional SC work, the authors of the present paper are not convinced as to the motivation for retaining transfer price as a tool for future studies.

### 3. Survey context

In some cases, the topic of a study directly determines the context of the SC surveys, such as for example in the case of mode choice experiments. In some cases however, most notably in the context of work looking at valuation of travel time (VTT) measures, the analysis is more results than context driven, and various possible approaches arise, as discussed in the following subsection. Later on in this section, we also discuss issues with inappropriate and unrealistic contexts.

#### 3.1 Context for valuation of travel time studies

The VTT is the core WTP measure in a transport context, and a large share of SC surveys are commissioned precisely with the purpose of eliciting such VTT measures. Clearly, VTT measures can be obtained from surveys with very different contexts, and analysts have variously made use of route choice experiments, abstract choice experiments, and mode choice experiments.

Independently of the context of a survey, the main emphasis is on using the hypothetical choice scenarios to study the relative sensitivity of respondents to time and money components. The latter especially deserves some special attention. While in a public transport context, the cost component of the journey is relatively easily understood (i.e., journey fare), complications arise for car journeys. Indeed, the main cost component of a car journey is the running cost, which, although some weight should also be given to maintenance costs and depreciation, is essentially the fuel cost for the journey. Many studies do rely extensively on running costs when looking at the cost sensitivity of car drivers, but it is important to recognise that this is a difficult concept for respondents to comprehend, not least because fuel bills are not generally paid on a journey by journey basis.

Given the difficulties of relying solely on running costs, there is considerable interest in exploring the use of other cost components. With the absence of parking costs, which have little or no relationship to travel time, the main emphasis falls on road tolls. Route choice experiments are increasingly being framed as toll road experiments. In many ways, toll road studies offer one of the most realistic settings for studying valuations of travel time reductions, with a higher toll (or indeed a non-zero toll) applying to more rapid routes. This could be in the form of comparing several tolled routes, with negative correlation between the level of toll and the travel time (i.e., a more highly tolled route is faster), or a comparison between slower untolled routes and faster tolled routes.

Two main complications however also arise in toll road studies. The first of these is that of experience. Indeed, in some countries, toll roads are still relatively rare, and a large share of the driving population will have little or no experience of toll roads. This not only causes problems with sampling strategies, which we will return to below, but also potentially means that many respondents of the survey will have difficulties relating to the *toll* attribute.

The lack of experience with toll roads arguably also accentuates the second problem with such studies, namely that of a high aversion by respondents to choose tolled options. Road tolls are a contentious issue, and surveys including a toll attribute are often affected by a high level of political voting or lexicographic behaviour, with respondents refusing to choose a tolled option (if untolled options are available) or always choosing the option with the lowest toll. This is indeed especially the case in surveys making use of respondents with limited or no exposure to the *benefits* of toll roads, and respondents where the *current* route is untolled. Not surprisingly, results in such studies are often affected by misunderstanding as well as strategic bias, as recently observed by Chintakayala et al. (2009a).

Significant effort has also gone into using mode choice experiments in the study of VTT measures, notably in Switzerland (see e.g., Axhausen et al., 2008). Such experiments can be useful in producing VTT measures jointly for different modes, while they arguably also have an advantage in masking the aim of the work. However, mode choice studies often face major issues with mode allegiance, with many respondents being unwilling to switch mode even in return for large time savings. This is then however not necessarily a reflection of a high or low VTTS, but simply of high modal allegiance.

In an attempt to avoid problems with toll road and mode choice studies, VTT studies in many countries have made use of abstract choice scenarios, presenting respondents with explicit time money trade-offs. Not only are there potential issues with unrealistic time cost trade-offs, as addressed in the next section, but such abstract scenarios bear little resemblance to real world scenarios. This in itself can pose significant problems.

David Hensher, one of the leading advocates for realism in SP design, has put forward the notion of “*experientially meaningful configurations*”, i.e., ensuring that respondents are presented with choices that would be reflective, at least up to a degree, of real life scenarios so as to ensure an acceptable degree of realism and response quality. This is arguably not the case in such abstract choice scenarios, and it is not immediately clear whether getting respondents to make such a *leap of faith* in completing a SP scenario can be guaranteed to have no influence on results. Crucially, there is very little evidence on this issue to date, but abstract scenarios continue to be used quite widely in an applied context.

## 3.2 Unrealistic contexts

As already stressed above, it is of crucial importance to present respondents with realistic choice experiments. In this section, we look in particular at the realism of presented attribute level combinations before turning our attention to two examples of SC surveys affected by problems with realism.

### 3.2.1 Unrealistic attribute level combinations: the time/cost example

SC surveys present respondents with scenarios where alternatives are described by a range of attributes, and where respondents are expected to make trade-offs between different attributes. To this extent, surveys will often make use of designs in which attributes that we expect respondents to trade on being negative correlated with one another. This is clearly not the case with orthogonal designs, but problems with dominated choices may arise as a result, as discussed later.

The two attributes that we focus on in this discussion are time and cost. With the aim of encouraging trading between time and money, independently of the survey context, there is a natural temptation to allow for negative correlation between travel times and costs in the hypothetical choice sets. As such, a faster journey is more expensive than a slower journey. This generally makes good sense in a public transport context, but can create problems in a car context where running costs are included. Indeed, by definition, in real life scenarios, travel time and running costs are strongly correlated, while, with the above rationale, a SC choice set would tend to give respondents a choice between a cheaper but slower option and a faster but more expensive option. This however arguably fails the realism test in that the faster route now leads to greater running costs, which would not generally be the case in reality. Indeed, the faster route would arguably have to be significantly longer than the slower route for this to apply, leading to unrealistic speed assumption, coupled with the fact that fuel consumption per miles on faster roads is generally lower than on slower roads. If the cost attribute used in the surveys is a toll component, such negative correlation would be acceptable, with a faster route incurring a higher toll. However, toll road studies often make use of both toll costs and running costs, and once again the issue of correlations between travel time and running costs need to be addressed.

### 3.2.2 Example 1: Swissmetro

Swissmetro is a hypothetical underground railway system, using maglev technology and travelling at speeds of over 400 km/h under the whole of Switzerland, with extensions to other European cities (see Figure 7). The highly ambitious project is arguably not likely to ever be completed. Nevertheless, a SC survey was conducted, giving respondents a choice between car, rail and the Swissmetro (cf. Bierlaire et al., 2001). With a headline figure of Zurich to Berne in 12 minutes (where a conventional train takes 57 minutes), the advantages of the Swissmetro option are however so big that it should come as no surprise that Swissmetro was chosen in 58 percent of choice sets.

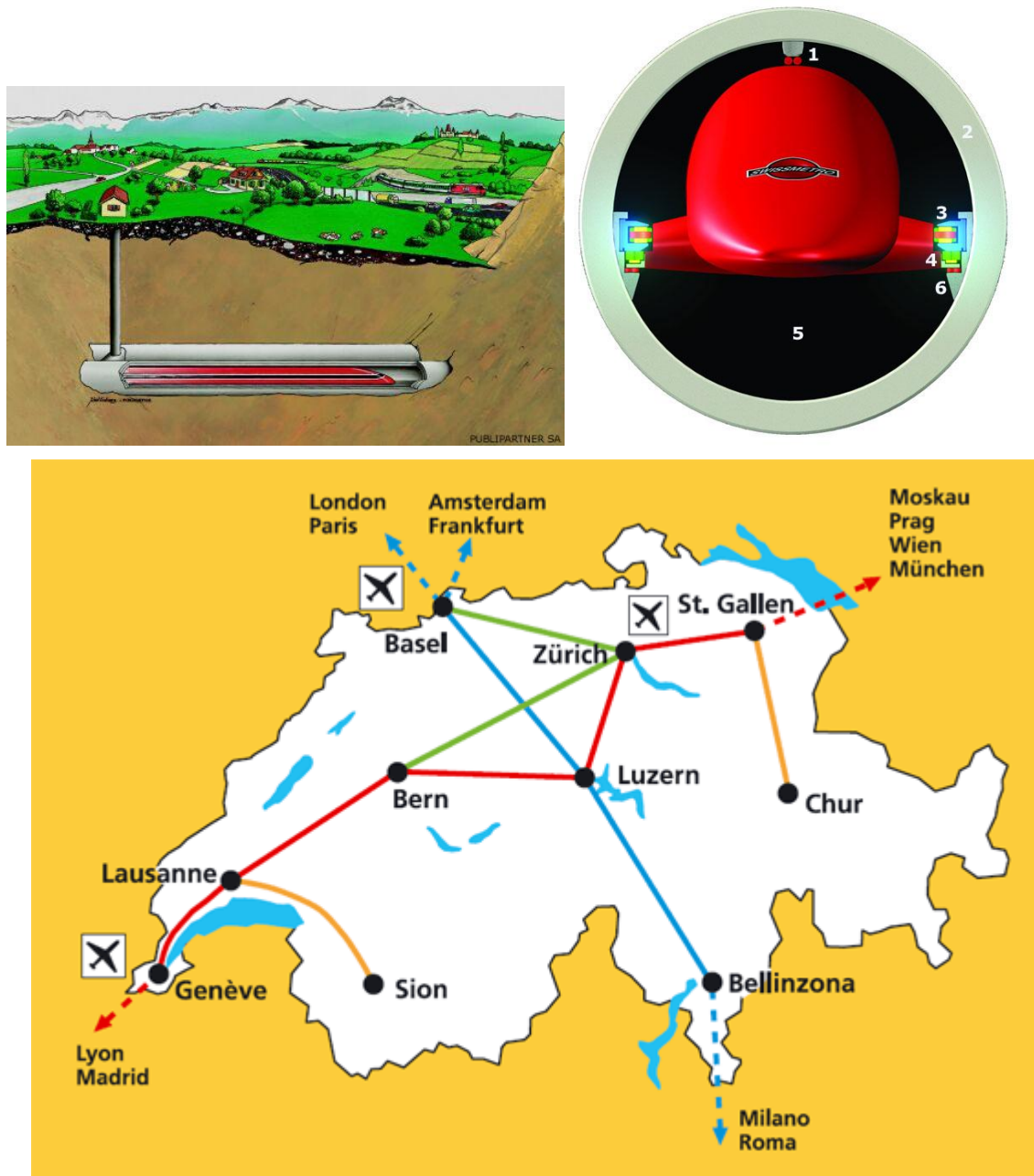


Figure 7: Swissmetro (figures copyright of Pro Swissmetro)

### 3.2.3 Example 2: The Sydney bush fire example

A large number of discrete choice projects, both consulting and research oriented, are typically constrained not only by budget, but by time considerations. This often means that such projects are rushed, with many aspects of the research design done in such a way as to cut corners. The easiest



and one of the most frequently left out part of many studies, is the use of qualitative research to refine the questionnaire design.

Consider a bush fire evacuation choice study conducted in Sydney, Australia in 2003. The project was designed to examine what factors would result in respondents evacuating their residence given an approaching bush fire. A preliminary examination of the literature resulted in a traditional grid like choice task design, where respondents were to be asked to select which bushfire they would be most likely evacuate from given two possible fires. An example of the proposed choice task is shown in Figure 8.

Description of Fire	Fire A	Fire B
<b>Characteristics of the Fire</b>		
Type of fire	Scrub fire	Fire storm
Length of fire front	2.1 – 5 km	5.1 – 10 km
Your distance from fire front	10 km	Greater than 10km
Speed of fire front	2 km per hr	5 km per hr
<b>Weather Characteristics</b>		
Wind direction	Away from property	Towards property
Temperature	25° C	35° C
Relative Humidity	70 %	90 %
Predicted precipitation	No chance of rain	Probable thunderstorm
<b>Evacuation Characteristics</b>		
Evacuation notice	No notice given	2 hours given
Road network	1 road available	3 roads available
<b>Other Characteristics</b>		
Recent hazard reduction	No	Yes
Neighbours activity	Not currently evacuating	Not currently evacuating
	<b>Fire A</b>	<b>Fire B</b>
Of the two, which scenario are you more likely to evacuate from	<input type="checkbox"/>	<input type="checkbox"/>
Would you evacuate from this scenario	Yes <input type="checkbox"/> No <input type="checkbox"/>	Yes <input type="checkbox"/> No <input type="checkbox"/>
I would evacuate		
Immediately	<input type="checkbox"/>	<input type="checkbox"/>
Wait 1 hour	<input type="checkbox"/>	<input type="checkbox"/>
Wait 2 hours	<input type="checkbox"/>	<input type="checkbox"/>
Wait 2 to 5 hours	<input type="checkbox"/>	<input type="checkbox"/>
Wait greater than 5 hours	<input type="checkbox"/>	<input type="checkbox"/>

Figure 8: Initial Choice task based on literature review

Fortunately, qualitative research was conducted whereby focus group participants were shown the above choice task and asked whether it made sense to them and whether they could answer the question accurately. The focus group participants were unable to understand the task, arguing that in reality, individuals were unlikely to be faced with having to choose between which of two different bushfires they would evacuate from. Furthermore, asking questions as to the likely timing of evacuation, as was proposed, was not realistic given that such decisions are based on quickly changing circumstances. Finally, focus group participants indicated that the decision to evacuate was not as simple as choosing to evacuate or not, with many indicating that they may only evacuate some of the household members, whilst others would remain behind. Given the above as well as discussions related to the specific attributes and the levels that they could assume, the final version of the survey used a completely different choice context, as well as relying on graphics and videos for presentation. An example of the final survey task is shown in Figure 9.

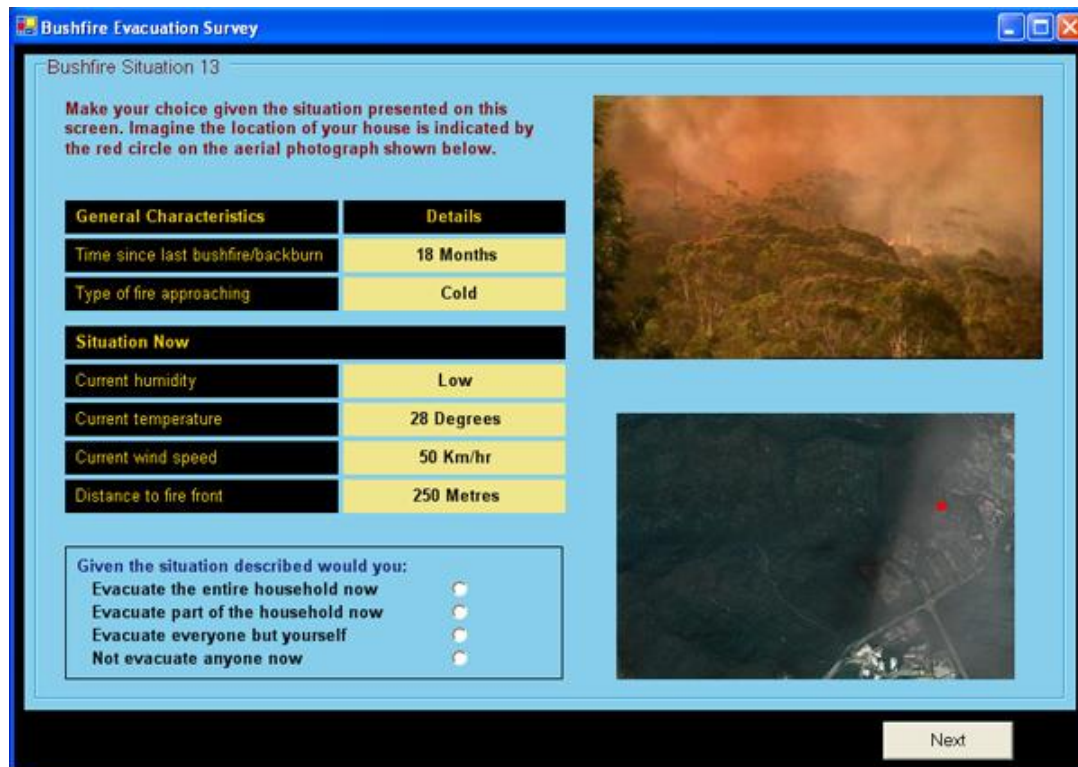


Figure 9: Final bushfire choice task based on qualitative research

## 4. Choice set design

The next step after choosing a survey context is to make decisions on the number of alternatives, attributes, and attribute levels and values. We will look at these different topics in turn, before also looking at the topic of referencing.

### 4.1

#### Alternatives and attributes

A choice situation in a SP survey presents a respondent with a fixed number of mutually exclusive alternatives, each described by a number of attributes. In generating a design for a survey, the analyst first needs to decide on the number of alternatives and attributes.

Many designs used in applied work still rely on binary choice experiments, i.e., involving only two alternatives in each choice situation<sup>2</sup>. Here, there is a major gap between theory and practice, with a large share of applied work relying on simplistic binary choice sets, while work of a more *academic* nature regularly presents respondents with choices involving three or more alternatives, sometimes up to five or six. The main argument in favour of using binary designs has been that of a reduction in respondent burden. However, work has not only shown that respondents can adequately deal with a larger number of alternatives, but that unnecessarily restricting the number of alternatives may in fact make the surveys too simplistic and transparent, while also bearing little resemblance to real life scenarios (see e.g., Caussade et al., 2005, who recommend four as the optimal number of alternatives). Additionally, a case can be made for increasing the number of alternatives on the grounds that this allows for greater variability in each choice set, increasing data richness while also reducing the overall sample size requirements. Finally, as discussed in the vehicle choice example

<sup>2</sup> In some applied work, the reliance on paper based surveys plays at least a partial role in the use of binary experiments.

below, it is not just the number of alternatives but also the type of alternatives that is of great importance.

Many surveys make use of only the most relevant attributes, typically time and cost. Such simplistic choice scenarios clearly avoid any risk of overburdening respondents, and this, in conjunction with the use of paper based surveys, was the main motivation for such an approach. However, simplistic scenarios can also be criticised, primarily on the grounds that they bear little resemblance to the more complex real life choices undertaken by travellers.

The incorporation of other attributes into the choice situations, such as departure time, reliability, or different travel time components, may be advantageous for three reasons. Firstly, they lead to a higher degree of realism, potentially improving response quality. Secondly, they mask the aim of the study, arguably reducing the risk of political voting. Finally, they obviously allow for the study of valuations in a broader context. As mentioned above, the main argument against increasing the complexity of stated choice scenarios is that of respondent burden. However, it has now been shown conclusively that not only are respondents able to cope with relatively complex scenarios (see e.g., Caussade et al., 2005; Chintakayala et al., 2009a), but that making choice sets relevant by including all important information may in fact improve response quality (see e.g., Hensher, 2006).

#### 4.1.1. Example 1: A vehicle choice example

A study currently being conducted in Sydney Australia dealing with Automobile choice involves respondents first having to provide information as to their most recently purchased new vehicle, the levels of which were then used as an alternative in the subsequent choice tasks they were asked to undertake. Aside from the most recent purchase, respondents were also shown three other hypothetical vehicles to choice from. The experimental design strategy involved selecting randomly selected vehicles of different sizes to make up each choice scenario. An example choice screen is shown in Figure 10.

**Choice Scenario 2**

Make your choice given the vehicles presented in this table.

**If an attribute is not relevant across all alternatives, then please click on the label of the attribute.**

In an attribute is not relevant for one or more specific alternatives, then please click on the box that the attribute is in.

		Current Vehicle	Medium Petrol	Small Luxury Diesel	Large Hybrid
<b>Initial Cost Price</b>	Purchase Price	\$23,000	\$50,000	\$41,250	\$58,000
<b>Fuel Cost</b>	Price of Fuel (dollars per litre)	\$1.23	\$1.23	\$1.40	\$1.12
<b>Annual Charges</b>	Registration (including CTP)	\$600	\$660	\$450	\$660
	Annual Emissions Surcharge (definition)	\$150.00	\$660.00	\$337.50	\$0.00
<b>Usage Charge</b>	Emissions Charge (per 10km) (definition)	\$0.30	\$0.33	\$0.31	\$0.28
<b>Vehicle Features</b>	Fuel Consumption (litres per 100km)	9.5	11	9	7
	Engine Capacity	6	4	4	6
	Seating Capacity	5	5	2	5
	Country of Manufacture	Japan	South Korea	Australia	Japan

Please rank the above choices in order of preference (1 = most preferred, 4 = least preferred)

<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
Current	Petrol	Diesel	Hybrid

Please indicate which vehicles are ones that you would find acceptable

<input type="radio"/> Yes	<input type="radio"/> Yes	<input type="radio"/> Yes
<input checked="" type="radio"/> No	<input type="radio"/> No	<input checked="" type="radio"/> No

Given that the vehicle you rated number one is your preferred choice, on the following scale, how certain are you that you would actually make this choice?


<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input checked="" type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7	<input type="radio"/> 8	<input type="radio"/> 9	<input type="radio"/> 10
Very Unsure									Very Sure

Next

Figure 10: Automobile choice example

In selecting randomly sized vehicles to construct each choice task, inevitably, many respondents who took part in a pilot of the survey instrument were confronted with situations of having to select vehicles that were priced twice as much as their most recent purchase. The result of this was that the current vehicle was selected almost 90% of the time as the most preferred vehicle. For the main field phase, the experimental design was changed so that at least one hypothetical vehicle would be the same size as the current vehicle, one would be either smaller, the same size or one size larger, and the other would final third alternative would be randomly selected from any size model. This resolved the issue, with respondents then trading off between the available alternatives.

Around the same time as the Australian study, a study on vehicle type and fuel type choice was also carried out in California, and similar problems were noted where respondents were being presented with the choice between say a small current vehicle, and a very large alternative vehicle (an example choice task is shown in Figure 11). Similarly, respondents were initially presented with choices between often very different fuel types. As a result, a weighting approach was used, ensuring that the more relevant options had a higher probability of being included while however still guaranteeing that all possible combinations had a non-zero probability.



If the following vehicle options were available to you, which would you choose?  
Please carefully examine all the attributes of each vehicle and then select the one you will most likely purchase by filling in the circle below your choice.

Vehicle Choice 1	Vehicle A	Vehicle B	Vehicle C	Vehicle D
<b>Vehicle type</b>	Midsized car	Compact SUV	Midsized car	Compact van
<b>Fuel type</b>	Gasoline	Natural Gas (NGV)	Plug-in Hybrid (PHEV)	Clean Diesel
<b>Age of vehicle</b>	New (2009)	New (2009)	New (2009)	New (2009)
<b>Purchase price</b>	\$29,400	\$36,600	\$31,100	\$20,900
<b>Incentive</b>	--	--	\$1,000 tax credit	--
<b>MPG or equivalent</b>	29 MPG	15 MPG	60 MPG	31 MPG
<b>Fuel cost per year</b>	\$1,090	\$1,950	\$780	\$1,170
<b>Fuel availability</b>		1 in 50 stations		
<b>Refueling time</b>		10 Minutes at station, 4 hours at home		
<b>Driving range</b>		300 Miles		
<b>Maintenance cost per year</b>	\$460	\$370	\$350	\$550
<b>Acceleration (0-60 mpg)</b>	10.2 seconds	11 seconds	8 seconds	11.8 seconds
<b>Select One:</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 11: 2<sup>nd</sup> automobile choice example

## 4.2. Presentation of *difficult* attributes: the case of variability

While attributes such as travel time or cost are generally easily understood by most respondents, complications may arise with more abstract attributes, such as for example comfort. In this section, we focus on one such attribute, namely travel time variability. While some studies are purely dedicated to the study of the WTP for improvements in reliability, and can hence make use of e.g. graphical representation, the majority of studies simply want to include travel time variability as one additional attribute. This however raises the important question of how to present it. An example of where the attribute might possibly have been better chosen or represented is shown in Figure 13. Figure 13 presents an example choice task from a toll road study conducted in Sydney Australia in 2004. In the study, travel time reliability was considered to be an important attribute influencing route choice. The attribute was shown as a  $\pm$  value around the current travel time, a representation which proved problematic for two reasons. Firstly, the experiment dealt with the travel times and costs for a specific trip, whereas travel time variability presented in this way represents an accumulation over many trips. Secondly, the presentation of the attribute as both a plus and a minus can be somewhat confusing to respondents as well as the analyst, as it is not certain whether they are reacting to the possibility of arriving earlier or later to the intended arrival time. In this way the attribute might be considered somewhat ambiguous, which may certainly explain why it often produces random parameter estimates with zero mean, but significant standard deviations (see e.g., Hess and Rose 2009a or Hensher et al. 2006).

**Sydney Road System**

Practice Game

Make your choice given the route features presented in this table, thank you.

	Details of Your Recent Trip	Road A	Road B
Time in free-flow traffic (mins)	50	25	40
Time slowed down by other traffic (mins)	10	12	12
Travel time variability (mins)	+/- 10	+/- 12	+/- 9
Running costs	\$ 3.00	\$ 4.20	\$ 1.50
Toll costs	\$ 0.00	\$ 4.80	\$ 5.60

If you make the same trip again, which road would you choose?  Current Road  Road A  Road B

If you could only choose between the 2 new roads, which road would you choose?  Road A  Road B

For the chosen A or B road, HOW MUCH EARLIER OR LATER WOULD YOU BEGIN YOUR TRIP to arrive at your destination at the same time as for the recent trip: (note 0 means leave at same time)  min(s)  earlier  later

How would you PRIMARILY spend the time that you have saved travelling?

Stay at home  Shopping  Social-recreational  Visiting friends/relatives  
 Got to work earlier  Education  Personal business  Other

Back Next

Figure 13: Example of possible poor attribute representation (travel time reliability)

Figure 14 represents a more recent stylisation of travel time reliability from an experiment conducted in Brisbane Australia in 2008. In it, the travel time reliability attribute is presented as probabilities or more accurately percentages (as respondents understood the concept of percentages much better than probabilities) of arriving earlier, on-time or later than expected. Qualitative research and pilot studies showed that this representation of the attribute was much more realistic for respondents and far less ambiguous as to its meaning (Li et al. 2009).

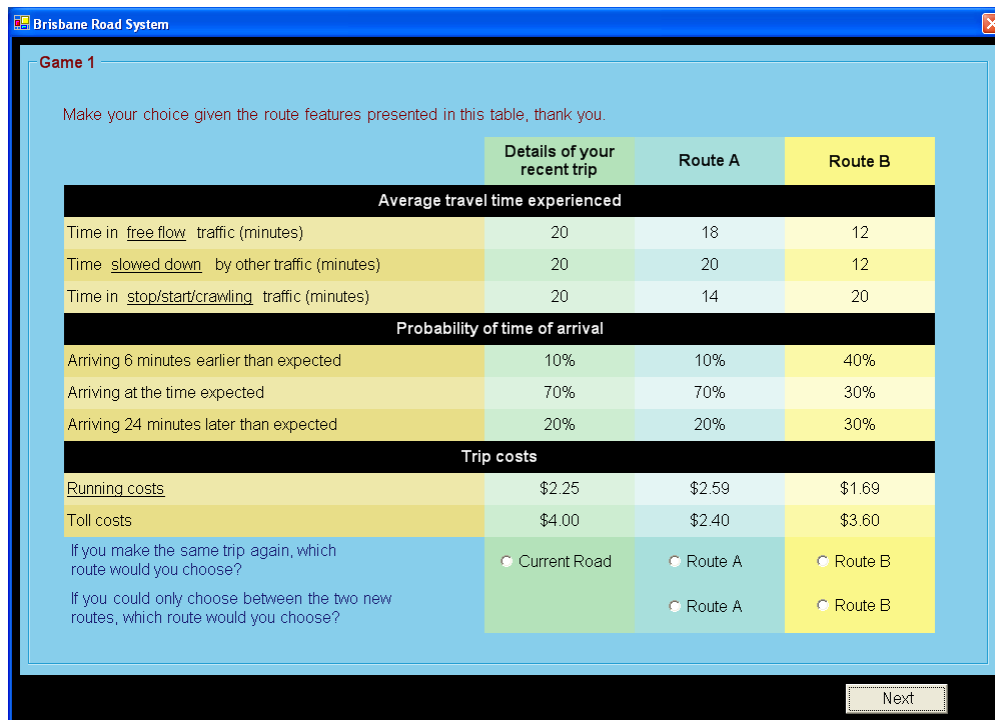


Figure 14: Example of a potentially better attribute representation (travel time reliability)

### 4.3 Realism and availability of alternatives

Realism in SC experiments arises from the fact that respondents are asked to undertake similar actions as they would in real markets (i.e., respondents are asked to make ‘choices’ just as they do in real markets, yet another reason why ‘pick one’ choice response formats have become the predominant data collection mechanism in SP studies). However, for any individual respondent, realism may be lost if the alternatives, attributes and/or attribute levels used to describe the alternatives do not realistically portray that respondent’s experiences or, in terms of ‘new’ or ‘innovative’ alternatives, are deemed not to be credible. As discussed above, concerns related to the attributes and attribute levels used within a SC experiment may be alleviated with significant prior preparation on behalf of the analyst (Hensher et al., 2005). Additionally, for quantitative variables, pivoting the attribute levels of the SC task from a respondent’s current or recent experience is likely to produce attribute levels within the experiment that are consistent with those experiences, and hence, produce a more credible or realistic survey task for the respondent (see for example, Rose et al., 2008).

Nevertheless, a significant proportion of products and services offered in real markets exhibit uneven degrees of distribution coverage (Lazari and Anderson, 1994). Such unevenness in availability may be either geographical, temporal or both. For example, train services may not be available to certain suburbs due to a lack of existing infrastructure or a train strike on a specific day might temporarily remove the train alternative from an individual’s choice set. Such constraints on availability are likely to be population wide (or at least impact upon a large proportion of the population), and as such, have an even impact over the entire study population. In SC experiments, such impacts may easily be handled through the removal of the affected alternatives from choice sets shown to respondents (removal may be from all choice sets within the experiment or subsets of choice sets to test availability effects in the presence or absence of an alternative). Other factors, however, may result in the non-availability of an alternative at the individual level. For example, for a specific journey, an individual may not have access to a car because a partner is using the vehicle, or alternatively, the household may not be able to afford a car in the first place.

Rose and Hensher (2006) discuss the generation of experiments that adapt to the reported availability to respondents of different alternatives. The purpose behind such experiments is to construct SC experiments in which the alternatives present within the choice tasks are respondent specific, and hence, reflect individual differences in the choice contexts that are likely to exist within real markets. In this way, alternatives that would never be available to specific individual respondents are not shown to them, and hence the preference structure that they reveal in undertaking the survey is much more likely to mirror that which they would exhibit in real markets. Figures 12a and b show one such adaptive survey where the alternatives shown to respondents was determined by whether the respondent had access to a car or not for a recent surveyed trip, and the origin and destination of that trip.

#### 4.4 Referencing and pivoting: pros and cons

With a strong interest in making surveys more relevant to respondents, choices are often framed around specific current levels. One approach in this context is to include a reference trip as one of the alternatives in the survey, typically alongside two further hypothetical alternatives. This is for example the standard approach in many Australian studies. While this has the advantage of putting a reference trip *in front* of the respondent, it also potentially leads to large levels of inertia (in the form of non-trading) and special care is required during the design and the analysis. This approach has also been shown to be of great use in analysing the differences between gains and losses (see e.g., Hess et al., 2008), where the question however arises as to how the SP presentation may in fact influence the results in terms of gains and losses, and whether this is an area where the incorporation of RP work may be desirable.

Independently of the nature of the design, another issue in this context is the actual definition of the reference point. Here, the question needs to be asked if the *current trip* is actually the most natural reference point for an individual when it could equally well be the *ideal trip*. This issue needs to be kept in mind at the design stage (to enable adequate pivoting) but is also of crucial importance at the modelling end, as discussed later on in this document.

		Light Rail connecting to Existing Rail Line	New Heavy Rail	Bus	Existing M2 Busway	Existing Train line	Car
Main Mode of Transport	Fare (one-way) / running cost (for car)	\$ 4.50	\$ 7.50	\$ 4.50	\$ 6.90	\$ 7.50	\$ 5.60
	Toll cost (one-way)	N/A	N/A	N/A	N/A	N/A	\$ 3.30
	Parking cost (one day)	N/A	N/A	N/A	N/A	N/A	\$ 10.00
	In-vehicle travel time	62 mins	56 mins	53 mins	45 mins	90 mins	60 mins
	Service frequency (per hour)	10	3	2	6	4	N/A
	Time spent transferring at a rail station	4 mins	2 mins	N/A	N/A	N/A	N/A
Getting to Main Mode	Walk time OR	6 mins	3 mins	11 mins	75 mins	10 mins	N/A
	Car time OR	1 mins	2 mins	2 mins	13 mins	3 mins	N/A
	Bus time	2 mins	2 mins	N/A	25 mins	9 mins	N/A
	Bus fare	\$ 4.00	\$ 2.00	N/A	\$ 2.25	\$ 3.10	N/A
Time Getting from Main Mode to Destination		10 mins	10 mins	19 mins	30 mins	15 mins	4 mins
Thinking about each transport mode separately, assuming you had taken that mode for the journey described, how would you get to each mode?		<input type="radio"/> Walk <input type="radio"/> Drive <input type="radio"/> Catch a bus	<input type="radio"/> Walk <input type="radio"/> Drive <input type="radio"/> Catch a bus	<input type="radio"/> Walk <input type="radio"/> Drive	<input type="radio"/> Walk <input type="radio"/> Drive <input type="radio"/> Catch a bus	<input type="radio"/> Walk <input type="radio"/> Drive <input type="radio"/> Catch a bus	
Which main mode would you choose?		<input type="radio"/> Light Rail	<input type="radio"/> New Heavy Rail	<input type="radio"/> Bus	<input type="radio"/> Existing Busway	<input type="radio"/> Existing Train	<input type="radio"/> Car

Figure 12a: Example mode choice experiment with 6 alternatives

		Light Rail connecting to Existing Rail Line	New Heavy Rail	Bus	Car
<b>Main Mode of Transport</b>	Fare (one-way) / running cost (for car)	\$ 2.20	\$ 3.30	\$ 3.75	\$ 1.35
	Toll cost (one-way)	N/A	N/A	N/A	\$ 4.00
	Parking cost (one day)	N/A	N/A	N/A	\$ 5.00
	In-vehicle travel time	10 mins	14 mins	23 mins	30 mins
	Service frequency (per hour)	13	4	2	N/A
	Time spent transferring at a rail station	8 mins	0 mins	N/A	N/A
<b>Getting to Main Mode</b>	Walk time OR	8 mins	10 mins	1 mins	N/A
	Car time OR	1 mins	1 mins	0 mins	N/A
	Bus time	5 mins	2 mins	N/A	N/A
	Bus fare	\$ 2.00	\$ 3.00	N/A	N/A
<b>Time Getting from Main Mode to Destination</b>		8 mins	6 mins	2 mins	2 mins
Thinking about each transport mode separately, assuming you had taken that mode for the journey described, how would you get to each mode?					
		<input type="radio"/> Walk	<input type="radio"/> Walk	<input type="radio"/> Walk	
		<input type="radio"/> Drive	<input type="radio"/> Drive	<input type="radio"/> Drive	
		<input type="radio"/> Catch a bus	<input type="radio"/> Catch a bus	<input type="radio"/> Catch a bus	
Which main mode would you choose?					
		<input type="radio"/> Light Rail	<input type="radio"/> New Heavy Rail	<input type="radio"/> Bus	<input type="radio"/> Car
<input type="button" value="Back"/>				<input type="button" value="Next"/>	

Figure 12b: Example mode choice experiment with 4 alternatives

A further issue with reference point designs which is only now starting to become apparent is that such designs appear to induce a significant proportion of respondents to exhibit inertia or non trading behaviour. Whilst such effects may indeed be the norm in many real markets (and hence suggest that the SP task is somewhat realistic or at the very least, produces realistic behaviour), inertia or non-trading does cause model estimation problems. If respondents always choose an alternative irrespective of the attribute levels of that and other alternatives, then unrealistic parameter estimates may result (e.g., the reference alternative may have a higher travel time, which if always chosen may produce a positive travel time parameter when modelled). Furthermore, respondent non-trading between alternatives may not provide any information as to the trade-offs that respondents are willing to make between the various attributes. If that is the case, then there might exist a possible trade-off for researchers between making choice experiments more realistic and making choice experiments that force trade-offs which may be useful in modelling respondents preferences. This is one of the arguments used by proponents of abstract experiments.

#### 4.5 Attribute levels

Another important decision relates to the number of levels used for each attribute in the design, and the actual values for these levels. Here, the main emphasis is generally on using a set of levels that is broad enough to allow for a diverse set of possible combinations and trade-off values while also not being so wide as to lead to unrealistic combinations.

Relatively little effort is however generally invested in the number of levels and the number of times each level is used in the survey. Raising the number of levels increases the number of possible combinations, improving the richness of the data, up to a point where the effects become detrimental (see e.g., Chintakayala et al., 2009a). Additionally, the more levels shown in an experiment, the more able the analyst is to detect non-linear marginal utility functions. However, from a statistical efficiency perspective, the more levels used, the less efficient the design will probably be. This is because the statistical efficiency of the design (which relates to the *t*-ratios one



will likely obtain from using that design) is a function of the choice probabilities. Contrary to what many might believe, the more attribute levels that are used, the more constrained the design will be in terms of the possible choice probabilities that it can achieve. For this reason, end point designs (designs with two levels at the extremes of the attribute levels) have often been found to produce the most statistically efficient results. However, this is also related to the range used for the attribute levels, as too wide a range may result in completely dominated alternatives. As such, there are several trade-offs that need to be considered in selecting what and how many attribute levels to use in a SC study.

#### **4.6 Counteracting effects of inertia: moving away from single choices**

As alluded to in Section 4.1.2, a significant issue with many experimental designs is inertia or non-trading between the alternatives. Whilst recent evidence is being gathered that suggests that this effect is a particular problem for experiments based on individual specific reference alternatives, similar effects may also be found in many experiments that involve status quo or no choice alternatives (for a discussion of this, see Hess and Rose, 2009b and Rose and Hess, 2009). A number of potential solutions to this issue exist, many of which we have already alluded to. The use of best worst, or rankings type response mechanisms, rather than the typical 'pick one' choice responses that are typically used, allow the analyst to capture information on additional preferences.

Even when 'pick one' type choices are used, the analyst may allow for partial rankings within the 'pick one' alternative approach. This solution involves the use of dual responses in SC experiments where respondents are first asked to select from amongst all non status quo alternatives (a forced choice) after which they are asked to make a second choice in which the status quo alternative is added (a non-forced choice). The use of dual responses in SC experiments, whilst potentially improving the statistical efficiency of estimated models as well as providing further information that can be used to refine the parameter estimates, may however lead to other potential modelling problems, in particular violations of the IID assumption if the data from the two choices are pooled into a single data set. IID violations may occur if the error variances between the two choice tasks are different. Brazell et al. (2006) acknowledge this potential problem, and found that in simulated data as well as in two empirical data sets, no such violations occur. Nevertheless, Dhar and Simpson (2003) who also explore issues related to the use of dual responses within SC choice tasks, did find limited evidence of such violations occurring. An alternative, though corresponding approach, is to first give respondents a choice from the full set of alternatives, and to force a choice between the purely hypothetical alternatives if the reference alternative is chosen in the first task.

This dual choice approach proved particularly useful in a recent toll road study conducted in the UK, where respondents have limited exposure and hence experience with such roads. In that data set, the reference alternative was selected in the majority of cases (75%), with resulting models being affected by major problems with retrieving significant estimates. Only once the second tier of choices was analysed where the reference alternative was not available could reliable models be estimated (see Chintakayala et al., 2009b). Of course, the discussions in this paper would point towards a need for adequate pilot work in such a context.

### **5. Experimental design issues**

One of the most critical components to any SP experiment is the underlying experimental design. Unfortunately, experimental design theory remains one of the least understood aspects of SP studies. To highlight this point, Bliemer et al. (2009) undertook a literature review of four top tier transport journals over the past decade and found that out of 61 SP studies in which the experimental design type and dimensions could be determined, 40 (66 percent) utilized an orthogonal design, 12 a D-efficient designs (20 percent), seven (11 percent) randomly assigned attribute levels shown to respondents and three (3 percent) used an adaptive design approach, alternating the levels shown to respondents based on the respondent's previous answers. Whilst

such a range of different design types would be understandable if the objectives of these papers were to explore the impact of different design methodologies, not a single one of these studies was specifically addressing experimental design issues. As such, despite decades of experience with SP studies, the disparity of design types employed suggests that the practical implications of using one design type over another is yet to be recognised within the literature.

## 5.1 What do we want from a good design?

As we have seen, one of the key aspects of SP design is the construction of realistic choice tasks. However, as we have also argued, realism may result in some undesirable choice behaviour, such as inertia or non-trading. Independent of how realistic the choice experiment is designed to be, there does exist within the literature a number of properties that are considered to be desirable in terms of the underlying experimental design used. Unfortunately, many of these properties may be considered to be myths that have developed over the years or be based on limited or untested experiences. For example, for many years, it was accepted wisdom that respondents could only complete three or four choice tasks at most with three or four attribute levels, thinking that many researchers today still adhere to. In their review of the literature, Bliemer et al. (2009) found choice experiments capturing between 1 and 25 choice tasks per respondents, with no evidence that respondents could not complete larger numbers of choices (although research does suggest more error variance occurs with larger numbers of choices captured per respondent; see Caussade et al., 2005). We now discuss properties considered desirable for experimental designs specifically for capturing discrete choice type responses.

### 5.1.1 Is attribute level balance a desirable property of designs?

Attribute level balance occurs when each level of an attribute occurs an equal number of times over the course of the experiment. The argument for such a constraint is that it will minimise behavioural bias in so far as respondents will not be exposed to situations with more or less 'better' or 'worse' attribute levels. For example, if low cost attribute levels are shown more than higher cost attribute levels, respondents may react much more strongly to the higher levels when viewed than would otherwise be the case if they saw high and low prices equally over the experiment. Evidence for this was found by Wittink et al. (1982, 1989, 1992) where experimental manipulations of the number of attribute levels and degree of balance resulted in systematic differences in estimated attribute sensitivity rankings. As such, evidence exists as to the behavioural impact that attribute level balance has upon derived model outputs.

Unfortunately, from a purely experimental design perspective, attribute level balance potentially produces some undesirable outcomes. Firstly, attribute level balance may result in larger than necessary experimental designs. This is particularly the case where odd and even attribute levels are combined in an experiment. For example, it might be possible to generate a design with 8 choice tasks if each attribute has either 2 or 4 attribute levels, but if one or more of the attributes has 3 levels, then the smallest design would require 12 choice tasks. Secondly, attribute level balance represents a constraint in generating a design with any constraint limiting the statistical efficiency of the design (see Kanninen 2002). This is not to suggest that statistical efficiency should be considered to be the most important aspect in generating the design, but simply to note that attribute level balance will likely produce less efficient designs. Finally, in the case of orthogonal designs, it might not be possible to find a design with the desired number of choice tasks that exhibits zero correlations between each of the attributes. This too may result in larger than necessary designs being generated.

### 5.1.2 Are larger designs better?

A common assumption in SP studies is that more choice tasks (in the design, not necessarily for each respondent) are required to produce greater levels of data variability which will aid in model estimation. Simply put, the assumption is that smaller designs will not provide enough coverage of

preference or utility space and hence generating designs with more choice tasks will produce better model outcomes. Bliemer et al. (2009) compared the results for the same choice problem using designs created with either 18 or 108 choice tasks. In that study, it was found that an efficient design with specifically chosen alternatives outperformed an orthogonal design with 108 choice tasks in terms of producing much smaller standard errors. This finding suggests that using more choice tasks is not necessarily better. What is important is how much information each choice task provides in terms of the trade-offs respondents are required to make. This also means that analysts should strive to produce designs that do not contain choice tasks that provide no additional information (e.g. dominated choices).

### 5.1.3 Are simpler designs better?

As we have seen, many researchers ascribe to the theory that simpler and smaller (per respondent) designs are better. Whilst this may or may not be the case in terms of the behavioural impact that more complex designs may produce, it does often result in one undesirable outcome. Where experimental design considerations are allowed to dominate the SP study, the experiments may become less realistic. Consider the experiments shown in Figure 12a and 12b. In that particular study, 453 respondents completed the choice experiment without being offered any form of incentive. This is partly explained by the fact that the sampled area is bereft of public transport options, and respondents when asked about completing the survey stated that they saw completing the survey as a means of expressing their frustrations in this regard to the government despite the survey taking between 45 minutes and an hour to complete. This suggests that the involvement of respondents may be just as important as the complexity of the task they are asked to complete when completing such surveys. As further evidence that using less complex experiments does not necessarily mean better modelling outcomes, contrast the experiments shown in Figures 13 and 14. The experiment shown in Figure 14 is far more complex in terms of the number of attributes shown to each respondent; however, the ambiguous nature of the travel time variability attribute used in the study represented in Figure 13 produces worse model outcomes. Indeed, the results produced from the second study, not only resulted in more intuitive WTP outputs, but also new insights into the VTT itself and how it might be confounded with the concept of reliability (see Li et al. 2009).

### 5.1.4 Is orthogonality important in choice experiments?

As suggested by Bliemer et al. (2009), the vast majority of experimental designs used within the literature are orthogonal in nature. This suggests that either researchers view the correlation structure of the design as being important, or that they are simply following processes and procedures used in the past blindly. This is not to suggest that reducing correlation is not important. Indeed, for several reasons, minimising correlations in the data may represent an ideal. Nevertheless, research into the design of experiments specifically for the use of SC studies has shown that the econometric models usually associated with such data do not require that the data be orthogonal. Indeed, the non-linear nature and the fact that these types of models are estimated as differences in utility, and hence differences in data, suggests that the correlation structure of the data is not what is important. This has resulted in the construction of so called efficient designs which generally trades orthogonality with the aim of reducing the standard errors of the estimated parameters. As such, efficient designs look not at the correlation structure, but rather at the expected asymptotic covariance matrix that will result from use of the design.

If one takes the view that the correlation structure of the data is not what is of primary importance for the econometric models that are estimated on SC data, provided that said data is not perfectly correlated, then questions arise as to whether the typical nomenclature commonly associated with experimental designs generated specifically with linear models, such as main effects only or main effects plus interaction effects, have any real meaning for SC experiments. Indeed, Rose and Bliemer (*in press*) demonstrate that despite the persistent use of such language in the SC literature, such designations do not actually match the reality of the models estimated. This is

because so called main effect and/or interaction effect designs are generated to produce the smallest possible standard errors for the parameter estimates as well as reduce to zero the parameter covariances (i.e., they are designed to produce independent parameter estimates of each effect). As Rose and Bliemer (*in press*) show, unlike linear models, once the parameters of a discrete choice model are no longer zero, then the parameter variances and covariances are no longer zero, with the values of the covariance matrix becoming larger for orthogonal designs as the parameters move further away from zero. In effect, this suggests that if an experiment has the desired outcome, that being to estimate non-zero parameter estimates (very few researchers upon suspecting an attribute will not play an important role in terms of observed choice behaviour will include that attribute in the experiment), then the more orthogonal the design, the worse the standard errors will be. As such, generating a design that have zero correlations for the main effects and/or selected interaction effects does not necessarily mean that such a design will deliver independent parameter estimates when discrete choice models are estimated upon data collected using the design.

Notwithstanding recent developments, which we return to below, the majority of SP questionnaires are still based on orthogonal designs. In an orthogonal design, the different columns in the design are uncorrelated. However, the use of orthogonal designs also poses a number of complications, primarily to do with dominance. Especially in simplistic designs, a potentially large number of choice sets will include dominated alternatives. Many studies largely ignore this issue, and retain such choice situations in the design, not realising that presenting respondents with such *no brainer* choices not only adds nothing to our understanding of the choice processes but potentially also has detrimental effects on response quality. Other studies take a more aggressive approach, simply removing these problematic choice situations. A problem with this approach is that it often leads to a loss of orthogonality, and almost invariably also leads to a loss of attribute level balance. Whilst not always perfect, efficient design techniques will mostly be able to be adapted to incorporate constraints and be set up to avoiding dominance.

Finally, it should be noted that manual designs, such as the so called Bradley design (see e.g. AHCG, 1996), and the design proposed by Hess & Adler (2009), also move away from orthogonality with a view to avoiding dominated choice scenarios and encouraging trading between relevant attributes.

### 5.1.5 What are the effects of blocking?

Often the total number of choice tasks generated for a design will exceed the number of choice tasks the analyst is willing to give any one respondent. When this occurs, the analyst needs to decide how to allocate subsets of the design to different respondents. In their review of the literature, Bliemer et al. (2009) also examined how such allocations were apportioned within SC surveys. In their review, they found that the majority of studies reported using a blocking column to allocate choice tasks to respondents (39, or 64 percent of designs), however, a number of studies were found to randomly assign choice tasks to respondents (8, or 13 percent). Three (5 percent) studies were found to use provide the full factorial to each respondent whilst it was not possible to determine how the choice tasks were allocated to respondents in 11 (18 percent) of the studies reviewed. The authors are also aware of studies in which the rows in the design are allocated sequentially. The aim behind including the block as an additional design column is that it avoids correlation between the blocks and the remaining attributes. In other words, we avoid a situation in which one block (and hence one group of respondents) is allocated say all the high cost scenarios.

Hess et al. (2008) compared the results of three different experimental designs including an orthogonal design with randomized choice set assignment, an orthogonal design with an orthogonal blocking column and an efficient design. In that study, they found that the efficient design performed only marginally better than the orthogonal design with blocking, but that the design with random assignment of choice tasks to respondents performed significantly worse than both the efficient design and the orthogonal design with blocking. As such, they concluded that the blocking

of the experiment was far more important than the underlying experimental design. In a similar vein, Bliemer et al. (2009) also empirically examined the impact of blocking, examining the effects of maintaining equal representation of blocks within a data set versus allowing an uneven sampling across each block. In that study, they found that statistical differences occur in terms of the standard errors of the parameter estimates that were found depending on the sampling over blocks that occur. As such, it is recommended that formal blocking columns be used in SC surveys and that random assignment of choice tasks be avoided where possible. Sequential blocking is of course worse still.

Finally, it should be noted that the use of inappropriate blocking approaches will jeopardise the characteristics of the data. Indeed, analysts sometimes forget that what matters are the qualities of the data, not of the base design. Even if the underlying design is perfectly orthogonal, using random blocking will mean that the final data is not, especially with small sample sizes.

## 6. Survey testing

### 6.1 Inclusion of consistency checks

It is becoming more popular of late to include so called *no brainer* choices in surveys, generally in the form of dominated choices, and to eliminate any respondents failing these tests. While we recognise the importance of such tests, especially in the case of a departure from current methods, we feel it is important to mention that such tests have to date often been performed in a potentially inappropriate manner. As an example, the recent Danish VTT study included a dominated choice as the sixth (out of nine) choice task. The problem with this approach is that, very much in the same way as retaining dominated choices in standard orthogonal designs, the presence of this choice scenario may lead to respondents not taking the remainder of the survey seriously. Work by Hess et al. (2009) shows some evidence of different behaviour before and after this choice scenario. For this reason, it is our recommendation that if such tests are to be included, this should be done at the end of the survey, that way avoiding any biasing influence on the remainder of the data.

### 6.2 Using simulation to test designs

Aside from what type of experimental design to use (i.e., orthogonal, efficient, etc.), many researchers fail to test how the design might perform in practice. One such example where such testing would have assisted in practice occurred as part of a research project conducted in South America in 2002 (reported in Efron et al., 2003). In that project, an orthogonal design was constructed, with orthogonal blocking columns. Each block was replicated an equal number of times over the sample so that orthogonality could be maintained through to the data set.

Once data was collected, the data was reformatted so that qualitative variables could be estimated using dummy codes. Unfortunately, despite each block being represented equally in the data, the correlation structure of the dummy codes was such that for some variables they were near perfectly correlated, meaning that these effects for the dummy codes could not be estimated. This was despite the design being orthogonal. Had the researchers tested (possibly via simulation) the design's performance in terms of how it was to be used in practice, this issue may have been identified earlier and the problem rectified prior to data being collected.

Nevertheless, simulations may not necessarily detect all possible issues that may arise once the experiment has gone to field. Indeed, a point that is often not recognised is that while simulation can show whether a design is able to capture specific effects, the impact of the design on actual behaviour cannot be tested in this manner. Indeed, most simulations assume that respondents act in a logit like fashion, when in reality, respondents often implement various behavioural rules of differing complexity (such as employing lexicographic choice rules). Unless such behavioural rules are incorporated in the simulation process, then the simulation is unlikely to mimic the real

outcomes associated with discrete choice data. For example, Hess et al., 2009 produced results that suggest that respondents lacked the ability to adequately distinguish between the alternatives in presented in the Danish VTT survey, where simulations had not revealed any problems.

### 6.3 Pilot and skirmish surveys

Independently of the decisions made in relation to survey design, significant pre-testing of the survey instrument is highly recommended, involving comparisons of various approaches prior to the main survey. Given the above discussion in relation to the shortcomings of simulation, such pre-testing and piloting should include actual data collection. Unfortunately, in academic settings involving transportation contexts, little evidence exists that such testing takes place (alternatively, one could equally argue that little evidence also exists that such testing does not take place). Unlike other discipline areas such as the environmental economics literature where pretesting and survey design can take 12 months or more, little credence is placed on discussing or even commenting upon whether and to what extent any pretesting and piloting has occurred within the transportation arena. Indeed, whilst it is hard to judge the extent of any pre-testing that occurs given what little comment is devoted to such an issue in published journal articles, it can only be assumed that in academia, as with consulting, very little if any pretesting and piloting occurs due to a lack of funding or available time. Nevertheless, when conducted, piloting and pretesting can provide invaluable insights (see Section 3.2.3).

In addition to the previously discussed Sydney bushfire survey, consider the process used to derive the choice survey shown in Figure 13. Prior to going to field, the survey instrument was first created in Microsoft Excel. In setting up the Excel version of the survey, the survey mock up was made to look as close to possible to the final survey form. Further, using the functionality available in Microsoft Excel, the survey was set up so that it could actually be used to collect data (shown in Figure 15; see Black et al. (2005) for a description of Microsoft Excel and its functionality in terms of being able to collect data). Once complete, the Excel version was sent to all stakeholders for comments allowing for changes to the survey instrument prior to it being formally programmed. Next, after the survey questions and layout were agreed upon, the survey was pre-piloted using a small number of respondents. This allowed not only for the survey instrument to be tested in terms of respondents answering each question, but also for a small scale test of the logistical requirements associated with administering the survey in field. This pre-test was followed by further discussions with additional changes to the survey then made (e.g., the inclusion of the travel time variability attribute shown in Figure 13). A more traditional pilot survey using the fully programmed computer aided programmed interview (CAPI) instrument was then held, further testing the logistical requirements of the survey, before the project was finally used in the main field phase.

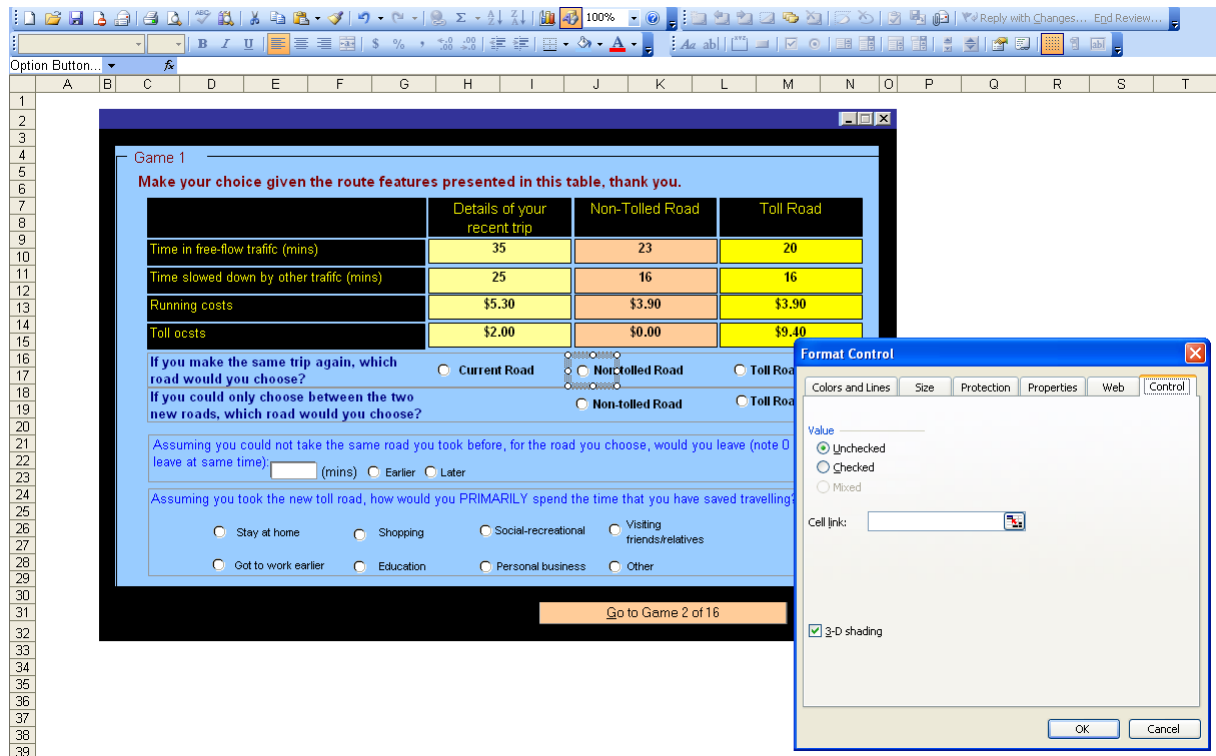


Figure 15: Example of Pre-test CAPI screen in Excel

## 7. Survey administration

### 7.1 What other information should be collected?

SC surveys invariably collect some form of additional information on top of the data relating to the respondents' preferences. This includes socio-demographic and attitudinal data, as well as data relating to the questions put to respondents to get them to explain the process that led to their choices, e.g. information processing strategies. Analysts should always attempt to collect whatever additional data they feel may be useful at the modelling stage, and should use as a warning the large number of studies that fail, for one reason or another, to collect a vital piece of information and then need to rely on arbitrary processes such as imputation. However, analysts should also be mindful of the fact that some of these additional questions can be of a personal nature, such as for example to do with attitudes and attributes such as income. For this reason, it is crucial that such data be collected after the actual choice experiments, so as to avoid influencing the actual behaviour. Analysts also need to make careful trade-offs between asking for enough information and not to *overdo it* and risk non-response, for example when using too high a level of disaggregation for income.

### 7.2 Sampling size requirements

Once the survey instrument is ready to go to field, important decisions need to be taken in relation to the overall sample size as well as the number of segment quotas, if any, that will be used. In terms of the overall sample size, many studies employ rather arbitrary approaches to calculate the sample size, e.g., using some arbitrarily chosen round number of respondents (such as 300) for each quota group. Typically, budget considerations underlie such strategies. Ad hoc formulae, constructed with little scientific or theoretical backing, also exist to determine the minimum number of respondents required for estimating model parameters (see e.g., Orme, 1988). More recently however, significant progress has been made in developing theoretically valid formulae that take additional prior information into account to provide more accurate guidance on necessary sample sizes (see e.g., Bliemer and Rose, 2009). In terms of placing quotas on various sub segments of the

population, Rose and Bliemer (2006) show how optimal sample size requirements can be determined for each segment. Whilst Bliemer and Rose (2009) suggest that sample sizes derived from such equations should represent a minimum bound in terms of the actual sample sizes required for data collection, Bliemer et al. (2009) found that the real sample sizes required were remarkably close to those suggested by the equations.

Nevertheless, in studies aimed at drawing wide and general conclusions, a certain degree of representativeness needs to be achieved in the sample. If only a small sample is collected, there does exist the potential that sampling bias will mean that the sample selected will not be representative of the overall population from which it is drawn. As such, even if smaller sample sizes are required than are generally collected, as suggested by Bliemer and Rose (2009), other external requirements may necessitate larger sample sizes be collected in practice.

### 7.3 Inappropriate contexts

The only way for data collected from SC surveys to be of use in the estimation of choice models is for respondents to make informed choices between the alternatives they are faced with. In some ways, this relates to the realism of the choice tasks presented to each respondent, where realism need not match reality, but rather match the sampled respondent's perceptions of reality. Unfortunately, in most contexts, different respondents will have different levels of experiences with the alternatives offered and given that experience is likely to be one of the key inputs into the formation of these perceptions, it is likely that different respondents will also have different perceptions of the offered alternatives. Furthermore, based on their previous experiences, some sampled respondents may be totally inappropriate for the study. Unfortunately, there exist many such examples where respondents have been faced with choice contexts in which they are unable to make informed decisions. For example, one mode choice survey that the authors are familiar with involved sampling respondents without a drivers license in a choice between car and rail.. Other such examples found within the literature include for example presenting a departure time choice to respondents with completely inflexible working times.

### 7.4 Data collection methods

There are still analysts who suggest that paper based surveys are the only acceptable survey delivery mechanism, given the greater flexibility for example for on-train surveys. Whilst paper based surveys offer many advantages, in particular they tend to allow a much wider reach in terms of sampling than computer based methods, as well as much lower costs, they also suffer from a number of disadvantages. As well as the obvious potential for coding mistakes (during the translation from paper to data files), paper based surveys may also be somewhat limited in terms of the degree of customisation possible. Typically, where customisation is required, separate surveys need to be prepared for a small number of segments. Furthermore, paper based surveys are generally faced with having to relate values of alternatives not in absolute terms relative to some reference option, but rather in such a way that respondents are required to carry out calculations to determine the precise value for themselves. As an example, an alternative may be described as:

*Current travel time + 10 minutes*

*Current travel cost – 50 pence*

Here, two main issues arise. Firstly, there is obvious scope for numerical mistakes. Secondly, any digression between what the respondent uses for the *current travel time* and *current travel cost* and the values used during modelling will potentially lead to significant problems. Here, even asking respondents for their current values at the start of the survey may not be sufficient. With both problems, the end result would be that what is modelled is not necessarily what was used by the respondent in evaluating the alternatives. Thirdly, given the generally accepted knowledge that respondents struggle with the notion of percentage changes, absolute time changes have to be



used, which may create problems if the number of separate surveys is low, i.e., the same savings are used for potentially quite different journey times.

All the above listed problems can to a large degree be avoided by making use of a computer based survey, either in the form of an interviewer assisted survey or an internet based survey. Such surveys allow for a high degree of customisation and also carry out calculations automatically, avoiding numerical issues while also guaranteeing a correspondence between those values used in the survey and those values used in the models. Furthermore, such surveys can rely on percentage variations, which may produce more realistic attribute levels, while also increasing data richness. Finally, a point that is rarely discussed is that in paper based surveys, respondents see all choice situations at the same time, potentially leading to cross-scenario comparisons of alternatives, an issue that does not arise in computer based surveys where one screen is used per scenario.

The cost of interviewer assisted surveys may be prohibitively high, paving the way for internet based surveys. Here, issues of sample representativeness may be avoided by still sampling respondents in the same way (e.g., roadside) and by handing out login details for the internet based survey, an approach that is again becoming more common place, for example in many river crossing and toll road surveys in the United States.

## 8. Using SP data inappropriately

Stated choice surveys have proven useful in examining many transportation related issues. For example, SC data has been used to examine the demand for cycle-way networks (e.g., Ortúzar et al., 2000), to examine the benefits derived from various calming measures on traffic (e.g., Garrod et al., 2002), to study the influences on parking choice (e.g., Shiflan and Bard-Eden, 2001; Hensher and King, 2001; van der Waerden et al., 2002) and to establish the VTT of commuters and non-commuters (e.g., Hensher, 2001a,b). Despite the flexibility of the method which allows it to be applied to almost any topic that one might wish to study, there do exist some limitations in terms of what outputs might be considered relevant to generate and report. In particular, very rarely should SP studies be used to generate measures of elasticities.

### 8.1 Elasticities and SP data

To understand why this is the case, consider the direct and cross elasticity equations for the MNL model. Both equations, shown as Equations (1) and (2), require the calculation of the choice probabilities. These choice probabilities are a function of the data as well as the parameter estimates, including any estimated ASCs. In discrete choice models, these ASCs represent the means of the random error terms, which reflect, after controlling for the modelled components of utility, the choice shares of the data. Given that SC are constructed as hypothetical markets, these choice shares are therefore reflective of the choice shares based on these hypothetical markets. Unfortunately, even if the choice survey is made as realistic as possible, the choice shares are unlikely to match those observed from equivalent real markets given that the experimental design imposed will not likely reflect the true market situation. As such, the ASCs from SC data, whilst required for estimation purposes, have limited meaning in terms of their interpretability. Unfortunately, unless the ASC values are calibrated so that the predicted market shares match those from real markets, the choice probabilities will also not reflect the true real market shares. Taking this argument to its logical conclusion, it follows that the elasticity values derived from such data will also not be correct.

$$E_{X_{ik}}^{P_i} = \frac{dP_i}{dX_{ik}} \cdot \frac{X_{ik}}{P_i} = \frac{dV_i}{dX_{ik}} X_{ik} (1 - P_i) \quad (1)$$

$$E_{X_{jk}}^{P_i} = \frac{dP_i}{dX_{jk}} \cdot \frac{X_{jk}}{P_i} = -\frac{dV_j}{dX_{jk}} X_{jk} P_j \quad (2)$$

This does not mean that the calculation of values such as elasticities (and likewise marginal effects) should never be contemplated. Indeed, the opposite is true. Where one wishes to compare the results across different models or data sets, then the generation of elasticities should be considered. However, unless the model constants have been calibrated, the actual values of the elasticities should be interpreted with care.

Additionally, there is a clear risk that the scale in SP data is different from that in RP data, i.e. that respondents' response to changes in attribute values is higher or lower than it would be in a real life scenario. Thus, it is not only the ASCs that need recalibrating, but also the scale of the marginal utility coefficients.

## 8.2 WTP and SP data

Whilst several studies have shown that SP experiments are able to reproduce the behavioural outputs, such as WTP measures, obtained from RP choice experiments (e.g., Carlsson and Martinsson, 2001; Lusk and Schroeder, 2004), contradictory evidence also exists that calls into question whether results obtained from SP experiments do in fact mirror those obtained from real markets. For example, Wardman (2001) and Brownstone and Small (2005) found significant differences between WTP values derived from RP and SP choice studies. In both these studies, VTT from SP experiments were found to be undervalued in comparison to the results from RP studies. Interestingly however, the opposite is typically observed in traditional contingent valuation studies where WTP values have been found to over value those observed in real markets (e.g., Harrison and Rutstrom, 2006; List and Gallet, 2001; Murphy et al., 2004; see Hensher (2008) for a detailed overview of differences obtained between WTP values from different survey methodologies).

## 9. Conclusions

In this paper, we have presented examples of previous SP studies to demonstrate some practical aspects of conducting such studies. In doing so, we have sought to provide practitioners and academic researchers alike with recommendations that might prove useful in order to avoid problems and mistakes that have been made in the past. We have also sought to provide discussion on other aspects of using SP surveys that hopefully will improve the reporting and use of SP survey results. Our discussions herein have led us to a number of conclusions, primary of which is that qualitative research is a must and that piloting and pretesting of SP surveys is a necessity.

We wish to conclude however by stating that such survey problems do not only afflict SP surveys. Many of the issues identified here may equally impact upon RP data collection as well. To this end, we offer the following example, taken from Hensher et al. (2005). "Some years ago a student undertook research into a household's choice of type of car. The student chose to seek information on the alternatives in the choice set by asking the household. Taking one household who owned one vehicle, their chosen vehicle was a Mazda 323. When asked for up to three alternatives that would have been purchased had they not bought the Mazda 323, the stated vehicles were Honda Civic, Toyota Corolla and Ford Escort. After collecting the data and undertaking the model estimation it was found that the vehicle price attribute had a positive sign (and was marginally significant). After much thought it became clear what the problem was. By limiting the choice set to vehicles listed by the respondent, we were limiting the analysis to the choice amongst similarly priced vehicles. Consequently more expensive vehicles (and much less expensive ones) were not being assessed in the data although some process of rejection had clearly taken place by the household. The price attribute at best was a proxy for quality differences amongst the vehicles (subject to whatever other vehicle attributes were included in the observed part of the utility

expression). Price would be better placed in explaining what alternatives were in or out of the choice set. If the student had simply listed all vehicles on the market (by make, model, vintage) and considered all eligible, then regardless of which grouping strategy was used (as discussed above) one would expect price to have a negative parameter; and indeed the model if well specified should have assigned a very low likelihood of the particular household purchasing a vehicle in a higher and a lower price range. An important lesson was learnt.”

## Acknowledgements

The first author acknowledges the financial support of the Leverhulme Trust in the form of a Leverhulme Early Career Fellowship.

## References

AHCG (1996), Value of Travel Time on UK Roads, report by Hague Consulting Group and Accent Marketing & Research for the UK Department for Environment, Transport and the Regions.

Axhausen, K.W., Hess, S., König, A., Abay, G., Bates, J.J. & Bierlaire, M. (2008), Income and distance elasticities of values of travel time savings: New Swiss results, *Transport Policy*, 15(3), pp. 173-185.

Batley, R., Grant-Muller, S., Nellthorp, J., de Jong, G., Watling, D., Bates, J., Hess, S. & Polak, J.W. (2008), Multimodal travel time variability, final report for the UK Department for Transport.

Bierlaire, M., Axhausen, K. and Abay, G. (2001). Acceptance of modal innovation: the case of the Swissmetro, Proceedings of the 1st Swiss Transportation Research Conference, Ascona, Switzerland

Bliemer, M.C. and Rose, J.M. (2009) Efficiency And Sample Size Requirements for Stated Choice Experiments, *Transportation Research Board Annual Meeting*, Washington DC January.

Bliemer, M.C., Rose, J.M. and Beelaerts van Blokland, R. Experimental Design Influences on Stated Choice Outputs, European Transport Conference, Leeuwenhorst, October 5-7.

Black, I., Efron, A., Anthony, C.I. and Rose, J.M. (2005) Designing and implementing internet questionnaires using Microsoft Excel, *Australasian Marketing Journal*, 13(2), 62-73.

Brazell, J.D., Diener, C.G., Karniouchina, E., Moore, W.L., Severin, V. and Uldry, P.F. (2006) The no-choice option and dual response choice designs, *Marketing Letters*, 17, 255-268.

Brownstone, D. and Small, K. (2005) Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A*, 39, 279-293.

Carlsson, F and Martinsson, P. (2001) Do hypothetical and actual marginal willingness to pay differ in choice experiments? *Journal of Environmental Economics and Management*, 41, 179-192.

Chintakayala, P.K., Hess, S., Rose, J.M. & Wardman, M.R. (2009a), Effects of stated choice design dimensions on estimates, paper presented at the inaugural International Choice Modelling Conference, Harrogate.

Chintakayala, P.K., Hess, S., & Rose, J.M. (2009b), Using second preference choices in pivot surveys as a means of dealing with inertia, paper presented at the European Transport Conference, Noordwijkerhout, The Netherlands.

- Dhar, R. and Simonson, I. (2003) The effect of forced choice on choice, *Journal of Marketing Research*, 40(2), 146-160.
- Efron A., Rose, J.M., and Roquero D. (2003) Truck or Train? A Stated Choice Study on Intermodalism in Argentina, presented at XVII Congresso de Pesquisa e Ensino em Transportes, Rio de Janeiro, Brazil, November 10<sup>th</sup> -14<sup>th</sup>.
- Garrod, G.D., Scarpa, R. and Willis, K.G. (2002). Estimating the Benefits of Traffic Calming on Through Routes: A Choice Experiment Approach, *Journal of Transport Economics and Policy*, 36(2), 211-232.
- Harrison, G.W. and Rutström, E.E. (2006) Experimental evidence on the existence of hypothetical bias in value elicitation methods, In: Handbook of Experimental Economics Results, C.R. Plott and V.L.Smith, Eds., Amsterdam: North-Holland.
- Hensher, D.A. (2008) Hypothetical bias and stated choice studies, submitted to *Transportation Research Part B*.
- Hensher, D.A. (2001a) The valuation of commuter travel time savings for car drivers: evaluating alternative model specifications, *Transportation*, 28(2), 101-118.
- Hensher, D.A. (2001b) Measurement of the Valuation of Travel Time Savings, *Journal of Transport Economics and Policy*, 35(1), 71-98.
- Hensher, D.A., Greene, W.H. and Rose, J.M. (2006) Deriving willingness to pay estimates of travel time savings from individual-based parameters, *Environment and Planning A*, 38, 2365-2376.
- Hensher, D.A. and King, J. (2001) Parking demand and responsiveness to supply, pricing and location in the Sydney central business district, *Transport Research Part A*, 35(3), 177-196.
- Hensher, D.A., Rose, J.M. and Greene, W.H. (2005) *Applied Choice Analysis: A Primer*, Cambridge University Press, Cambridge.
- Hess, S. & Adler, T. (2009), Experimental designs for the real world, ITS working paper, Institute for Transport Studies, University of Leeds.
- Hess, S. and Rose, J.M. (2009a) Allowing for intra-respondent variations in coefficients estimated on stated preference data, *Transportation Research Part B*, 43(6), 708-719.
- Hess, S. and Rose, J.M. (2009b) Should reference alternatives in pivot design SC surveys be treated differently?, *Environment and Planning A*, 42(3), 297-317.
- Hess, S., Smith, C., Falzarano, S. & Stubits, J. (2008) Measuring the effects of different experimental designs and survey administration methods using an Atlanta Managed Lanes Stated Preference survey, *Transportation Research Record*, 2049, 144-152.
- Hess, S., Rose, J.M. & Polak, J.W. (2009), Non-trading, lexicographic and inconsistent behaviour in stated choice data, *Transportation Research Part D*, accepted for publication, January 2009.
- Kanninen, B.J. (2002) Optimal Design for Multinomial Choice Experiments, *Journal of Marketing Research*, 39, 214-217.
- Li, Z., Hensher, D.A. and Rose, J.M. (2009) Willingness to Pay for Travel Time Reliability for Passenger Transport: A Review and some New Empirical Evidence, submitted to *Transportation Research Part E*.
- List, J. and Gallet, G. (2001) What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, 20, 241-254.
- Louviere, J.J., Street, D., Burgess, L., Wasi, N., Islam, T. and Marley A.A.J. (2008) Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information, *Journal of Choice Modelling*, 1(1), 128-163.

- Lusk, J. and Schroeder, T. (2004) Are choice experiments incentive compatible? A test with quality differentiated beef steaks, *American Journal of Agricultural Economics*, 86(2), 467-482.
- Marley, A.A.J. and Louviere, J.J. (2005) Some probabilistic models of best, worst, and best-worst choices, *Journal of Mathematical Psychology*, 49, 464-480.
- Murphy, J., Allen, P., Stevens, T. And Weatherhead, D. (2004) A meta-analysis of hypothetical bias in stated preference valuation, Department of Resource Economics, University of Massachusetts, Amherst, January.
- Orme, B. (1998) Sample Size Issues for Conjoint Analysis Studies, Sawtooth Software Technical Paper, <http://www.sawtoothsoftware.com/technicaldownloads.shtml#ssize>.
- Ortúzar, J. de D., Iacobelli, A. and Valeze, C. (2000) Estimating demand for a cycle-way Network, *Transport Research Part A*, 34(5), 353-373.
- Rose, J.M. and Bliemer, M.C.J. (*in press*) Constructing Efficient Stated Choice Experimental Designs, *Transport Reviews*.
- Rose, J.M. and Bliemer, M.C. (2006) Designing Efficient Data for Stated Choice Experiments, presented at 11<sup>th</sup> International Conference on Travel Behaviour Research, Kyoto, August 16-20, 2006, Japan.
- Rose, J.M., Bliemer, M.C., Hensher and Collins, A. T. (2008) Designing efficient stated choice experiments in the presence of reference alternatives, *Transportation Research Part B*, 42(4), 395-406.
- Rose, J.M. and Hensher, D.A. (2006) Handling individual specific non-availability of alternatives in respondent's choice sets in the construction of stated choice experiments, Stopher, P.R. and Stecher C. (eds.) *Survey Methods*, Elsevier Science, Oxford, pp347-371.
- Rose, J.M. and Hess, S. (2009) Dual Response Choices In Reference Alternative Related Stated Choice Experiments, Transportation Research Board Annual Meeting, Washington DC January.
- Shiflan, Y. and Bard-Eden, R. (2001) Modeling Response to Parking Policy, *Transport Research Record*, 1765, 27-34.
- van der Waerden, P., Timmermans, H. and Borgers, A. (2002) PAMELA: Parking Analysis Model for Predicting Effects in Local Areas, *Transport Research Record*, 1781, 10-18.
- Wardman, M. (2001) A review of British evidence on time and service quality Valuations, *Transportation Research Part E*, 37, 91-106.
- Wittink, D.R., Huber, J., Zandan, P. and Johnson, R.M. (1992) The Number of Levels Effect in Conjoint: Where Does It Come From and Can It Be Eliminated?, Sawtooth Software Conference Proceedings.
- Wittink, D.R., Krishnamurthi, L. and Nutter, J.B. (1982) Comparing Derived Importance Weights Across Attributes, *Journal of Consumer Research*, 8, 471 -4.
- Wittink, D.R., Krishnamurthi, L. and Reibstein, D.J. (1989) The Effects of Differences in the Number of Attribute Levels on Conjoint Results, *Marketing Letters*, (2), 113-23.