I T L S

WORKING PAPER

ITLS-WP-10-13

Interactive stated choice surveys: A study of air travel behaviour

By

Andrew T Collins, John M Rose and Stephane Hess[1]

[1] *Institute for Transport Studies*
University of Leeds

July 2010

ISSN 1832-570X

INSTITUTE of TRANSPORT and LOGISTICS STUDIES

The Australian Key Centre in Transport and Logistics Management

The University of Sydney

Established under the Australian Research Council's Key Centre Program.

NUMBER: Working Paper ITLS-WP-10-13

TITLE: **Interactive stated choice surveys: A study of air travel behaviour**

ABSTRACT:

Stated preference (SP) experiments are becoming an increasingly popular survey methodology for investigating air travel choice behaviour. Nevertheless, some evidence suggests that SP experiments do not mirror decisions in real markets. In this paper we introduce a novel survey methodology that aims to make air travel surveys more consistent with real world settings, with the aim of obtaining more realistic results. The survey is modelled on the interface and functionality of an online travel agent (OTA). As with a real OTA, many ticket options are presented. Sort tools allow the options to be reordered, search tools allow options to be removed from consideration, and a further tool allows attributes to be hidden and shown. Extensive use of these tools is made by the 462 respondents, with the captured data revealing some attribute preferences at the individual level, and significant heterogeneity of preference across individuals. A traditional SP component was also completed by the respondents. Our exploratory analysis as well as random utility model estimation results confirm not only that respondents seem to *engage* more actively with the interactive survey, but also that the resulting data allows for better performance in model estimation. These results have implications for the study of other complex travel choices where interactive surveys may similarly be preferable to standard approaches.

KEY WORDS: ***Airline choice, stated choice experiments, information search, survey realism, online travel agent***

AUTHORS: Andrew T Collins, John M Rose and Stephane Hess

CONTACT:

Institute of Transport and Logistics Studies (C37)
The Australian Key Centre in Transport Management
The University of Sydney NSW 2006 Australia

Telephone: +61 9351 0071
Facsimile: +61 9351 0088
E-mail: itls@sydney.edu.au
Internet: http://www.sydney.edu.au/business/itls

DATE: July 2010

1. Introduction

Stated preference (SP) experiments have grown to become the predominant data paradigm in the elicitation of behavioural responses of individuals, households and organisations over diverse choice situations and contexts. One partial explanation for this is research evidence suggesting that SP experiments are capable of replicating decisions made in real markets (see e.g., Burke et al., 1992; Carson et al., 1994). Several studies have shown that SP experiments are able to reproduce the behavioural outputs, such as willingness to pay (WTP) measures, obtained from revealed preference (RP) data (e.g., Carlsson and Martinsson, 2001; Lusk and Schroeder, 2004). Nevertheless, 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 Brownstown and Small (2005) found significant differences between WTP values derived from RP and SP choice studies. In both these studies, values of travel time savings (VTTS) 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 (see e.g., Hensher, forthcoming, for a detailed overview of differences obtained between WTP values from different survey methodologies). At stake is the external validity of the data collected via SP experiments and hence confidence in the findings emanating from these data.

Given such a divergence of evidence, of particular research interest is to determine firstly, to what extent SP experiments are able to replicate real market decisions, and secondly, if a difference between SP and RP results does exist, what factors can bridge the gap. A number of researchers such as Rose and Hensher (2006), Lanscar and Louviere (2008) and Hess and Rose (2009) argue that one such factor is the degree of realism used in SP surveys. Rose and Hensher (2006) suggest that the realism of SP experiments is bolstered by aligning the alternatives, attributes, and attribute levels with the respondent's experiences. However, for the analyst, the decision about what and how many alternatives, attributes, and attribute levels to include in the SP task is often a challenging one. The decision may be influenced by what attributes and alternatives the analyst believes will systematically alter choice. Yet the literature cautions against the inclusion of too much information. Research into what constitutes appropriate choice task dimensions has tended to centre on identifying sources of cognitive burden placed upon respondents undertaking SP tasks, (e.g., Arentze et al., 2003; DeShazo and Fermo, 2002) as well as reducing the cognitive load placed on those same respondents (e.g., Louviere and Timmermans, 1990; Wang et al., 2001). In the market, the amount of information in a choice task varies from one type of choice to the next, as well as from one individual to the next, as tastes and motivation levels vary. Lanscar and Louviere (2008) argue that the inclusion of 'constant' alternatives such as no choice alternatives as well as showing respondents only realistic attribute level combinations increases realism. Hess and Rose (2009) discuss the use of different response formats, unrealistic contexts and unrealistic attribute level combinations in SP studies, as well as the impact of using more realistic alternatives including the use of reference or pivot alternatives.

Despite lingering concerns within the literature about the necessity of often having to trade-off realism with the minimisation of task complexity, a number of authors are questioning what precisely task complexity means in terms of SP studies and how important it really is. Caussade et al. (2005) suggest that some respondents may prefer more complex choice environments, while others may prefer simpler ones. Hensher (2006) argues that the complexity of a choice task should be equated with the relevancy, not the quantity, of the information that must be processed. In taking this argument, Hensher reasons that more information does not necessarily make a choice more difficult and that in some circumstances, a lack of information on relevant attributes and alternatives may actually make the choice more difficult. We suggest that, under the reasonable assumption of between-individual variation in what information is relevant, a specific choice task of fixed dimensions will exhibit varying levels of complexity across a

sample or population. Several studies have sought to account for this heterogeneity of information relevance by explicitly asking the respondent what attributes were relevant to them, and then accounting for this within the choice model either deterministically (Rose et al., 2005; Puckett et al., 2008), or stochastically (Hensher et al., 2007). However, in these studies, the dimensions of the choice tasks, and hence the task complexity, remained fixed by the analyst. Furthermore, the wisdom of using stated information on processing strategies has recently been questioned by Hess and Hensher (forthcoming).

Additionally, however, it may not just be the makeup of the choice set (in terms of alternatives, attributes, and attribute levels) that allows a respondent to better relate to the scenarios he or she is faced with, but also the setting in which these choices are to be made, in other words the presentation. In this paper, we attempt to address both of the above issues in the context of a study looking at air travel behaviour. Specifically, we present a survey with a choice environment that allows the respondent to control the amount of information in the choice task in a way that is meaningful to them. The motivation is to boost the realism of the survey by creating a choice mechanism that functions more like those found in real markets, where the individual has more control over the amount of information that they are presented with and hence need to process.

The experience of an individual when searching for and deciding between choice alternatives may vary greatly from one type of transport context to another. These experiences in turn may be very different from the experience of completing an SP experiment conducted in the same choice context. Yet despite possible discrepancies in experiences between the two, the increasing use of the internet in transport choice contexts has resulted in many real market decisions being made in choice environments that are much more like SP experiments. These online choices range across numerous travel modes including air, rail, coach, ferry and cruise ship. Other choices that can be facilitated online or influenced by online information include car hire, route choice via online map tools, and various public transport choices through websites that create personalised timetables and routes. The way in which information is presented across the websites varies, however many common characteristics exist that liken them somewhat to SP experiments. For example, alternatives are frequently placed on the same page and there is often some consistency across these alternatives in what attributes are presented, which together invite ready comparison of the alternatives.

Unlike an SP experiment however, such websites present respondents with alternatives generated from real market data. Such alternatives are often prone to exhibit a number of limitations such as attribute level invariance, attribute correlation, and alternatives that exist only within the technological frontier of the market place as it then exists. This is often regarded as having a potential detrimental effect on the ability to retrieve sensitivities (see Louviere et al., 2000 and Hensher et al., 2005). This difference between typical SP experiments and their real world counterparts (e.g. artificially inflated scope for trade-offs) may however also be seen to unduly reduce survey realism and affect response quality. Further, whereas SP tasks typically present respondents with a small number of alternatives to choose from, users of these websites may be confronted with a very large number of alternatives. For example, a search for a long haul flight on a busy route may return dozens of flights in a one day period, spanning a dozen airlines, each described by a dozen attributes. To assist the customer, such websites typically provide search tools that allow the user to customise their search by viewing only relevant alternatives that meet some minimum or maximum desired level on one or more criteria. While the amount of information presented at the time of choice is influenced by the market offerings and decisions by the website architects, individual users typically have ultimate control over what they view.

Given the increase in the number of market choices being made online, and the natural congruence of SP and online RP choice environments, it is possible that SP experiments that are made to look and react in a fashion similar to real market RP contexts may improve the results of SP studies. In particular, this could involve mimicking the look and feel of RP choice environments, presenting more alternatives (not less), and including navigation mechanisms such as search and sort tools that allow the quantity and composition of the information to be controlled by the user. This paper introduces such a survey in the

context of airline choice. Two choice environments are presented to a sample of respondents; one that mimics the results of a search with an online travel agent (OTA), and one that follows a traditional SP grid like format with a limited number of alternatives shown. Section 2 examines existing studies of air travel choice behaviour, and introduces the OTA, before section 3 describes our survey in detail. Section 4 outlines the methodology that will be applied to test for differences between the two datasets. Section 5 presents results, and section 6 offers a discussion and conclusions.

2. Air travel behaviour modelling and online travel agents

A wide range of studies have investigated air travel choice behaviour using both SP and RP methods. Kanafani and Sadoulet (1977) modelled the choice among fare types for long haul journeys. Proussaloglou and Koppelman (1995) examined the choice of airline for recent trips using mail-in RP data. In recent years, a majority of studies have used the SP methodology. Bradley (1998) used SP data to examine the choice of departure airport and route from Schiphol, Brussels and Eindhoven airports. Hensher et al. (2001) used SP data for airline choice between New Zealand and Australia. Hess et al. (2007) and Hess (2007) also made use of SP data collected via the internet and retrieved effects of a number of attributes which often cause problems in RP data (fares, frequent flier benefits). Bliemer et al. (2009) examined different types of experimental designs, keeping the design dimensions shown to respondents fixed, whilst using airline choice as the decision context.

Whereas the SP studies above utilised a conventional SP task, Proussaloglou and Koppelman (1999) conducted a novel SP air travel survey that markedly departed from the conventional format. The study incorporated one way that travellers may search for information when talking to a travel agent on the phone. Presented with a travel scenario, the respondents had to elicit from the interviewer the available flights as described by schedule and fare. Flights could be revealed in any order the respondent wished – according to schedule or fare, and a choice could be made at any stage. The interviewer had a record of what flights had been revealed when the choice was made. This study allowed the respondent to drive the information search process prior to making a decision. However, any difference in results between this innovative survey mechanism and a traditional SP task could not be determined, as no traditional tasks were presented.

Of all the types of online travel choices discussed in Section 1, air travel is probably the most prominent example, where online travel agents have emerged as viable competitors to traditional travel agents, and account for a significant percentage of market share. In 2007, more travel was purchased online (through both OTAs and airline websites) than offline in the United States (PhoCusWright, 2007). Yet despite this, no SP choice study to our knowledge has presented a choice environment that resembles that of an OTA. Academic research into OTAs has examined price dispersion (Clemons et al., 2002) and the threat to conventional travel agents (Law et al., 2004). Smith et al. (2007) outlined how extensive RP data from a real OTA was used to generate choice models that formed part of a wider suite of models, which were applied to help meet performance targets and maximise profit. While the broad framework was outlined, and increases in profit detailed, no empirical choice model results were presented. Additionally, the use of data from a real OTA carries with it the usual limitations in terms of access to socio-demographic information due to data protection issues.

In this study we compare the results obtained from traditional SP and OTA-inspired choice environments. The reader is invited to examine real world OTAs, however given that they will change over time, it is worth documenting what their defining characteristics were at the time of this study. OTAs present highly detailed information on a large number of options that a traveller may choose from. To help customers make sense of so much information, a range of tools are typically provided. Searches can be refined on a range of criteria, and the alternatives can be sorted on many of the attributes. The level of control over the search process varies across OTAs, as does the mix of attributes used in the description of the options, where for example information on seat pitch and entertainment options are only gradually being included.

3. Survey description

The SP scenarios in the current study ask respondents to choose a ticket for travel from Sydney, Australia to either London or Paris, with the destination selected by the respondent at the start of the survey. A long haul route was used as it was believed that travellers are more discerning of attributes such as in-flight entertainment and seat pitch on such routes. The choice was framed as a leisure trip, hence avoiding any issues with business travellers having their tickets paid for by their employer. In the interests of survey simplicity, respondents were only presented with economy ticket options.

The survey presented to respondents contained two choice components; a traditional SP component consisting of a practice task followed by four simple choice tasks; and an interactive component modelled on OTAs, also with a practice task followed by four actual interactive choice tasks. The order of the two components was randomised as was the order of the tasks within each component.

For the traditional SP task, three unlabelled alternatives were included (although an attribute indicated the airline) alongside a "no choice" option. Two choices were captured: one between the three alternatives only, and one that allowed respondents to also select the no choice option. For the OTA task, the number of alternatives varied across tasks and respondents, ranging from 12 to 22. The same attributes were used for both presentation formats, and are listed in Table 1. The descriptions of the attributes provided to the respondents for both datasets can be found in the top half of Figure 2. All prices were displayed in Australian dollars. The average exchange rates for February 2008 (the time of the survey) were AUD1 = $US0.91 and AUD1 = €0.62.

Table 1: Attributes in SP and OTA tasks

Attribute	**SP levels**	**OTA levels or range**	**OTA: From real flight?**	**Typical online travel agent attribute?**
Price	AUD1600, AUD1900, AUD2200, AUD2500	AUD1809 – AUD6036	Yes	Yes
Carbon tax	AUD0, AUD120, AUD240, AUD360	AUD0 – AUD460.76	No	No
Airline	9 possible	13 possible	Yes	Yes
Departure time	6am, 10am, 5pm, 10pm	Continuous	Yes	Yes
Arrival time	Based on departure time and flight duration	Continuous	Yes	Yes
Total duration	20hrs, 22hrs, 24hrs, 26hrs	22hrs 20mins – 38hrs 40mins	Yes	Yes
Flight duration	Based on total and stopover duration	21hrs 20mins – 26hrs	Yes	Yes
Stopover duration	1hr, 2hrs, 3hrs, 4hrs	40mins – 14hrs 50mins	Yes	Yes
Number of stops	1, 2	1, 2, 3	Yes	Yes
Plane type	747, 777, A330, A340		No	Yes
Seat pitch	31”, 32”, 34”		No	No
Seat allocation available?	Yes/No		No	Yes
Entertainment system	Overhead televisions (shared), Personal screens with limited movie selection, Personal screens with video on demand		No	No
Itinerary change cost	AUD0, AUD100, AUD200, AUD300		No	Often hidden

While an experimental design was used for the SP tasks, the OTA tasks primarily made use of information from real world flights where available, in an effort to boost the realism of the survey. An experimental design was applied to select values for any remaining attributes. Two price components were shown: a carbon tax, and the ticket price excluding the carbon tax. Real airline names were displayed, always with their logo visible. Some of the comfort related attributes are not typically shown on real world ticket booking websites, as highlighted in Table 1. Here, our survey presents respondents with more detailed information while still allowing them to eliminate these attributes to simplify the tasks performed. In real decision environments, a decision about both the departing and return flights must also be made. In the interest of simplicity, for both presentation formats we only required a choice for the departing flight and asked the respondent to assume that the return flight would have similar service levels.

Along with cost, travel duration and timing, and equipment and quality of service indicators, the survey also looked at the effects of frequent flyer membership, which is widely recognised to have a significant influence on airline choice (Chin, 2002; Proussaloglou and Koppelman, 1999; Hensher et al., 2001; Proussaloglou and Koppelman, 1995). Increasingly, airlines are recognising this and are attempting to encourage passengers to choose more expensive fare classes in return for bigger benefits (i.e., fewer miles with discount tickets). To simplify the survey task, respondents were asked what, if any, frequent flier programs they were currently members of.

Finally, unlike some previous studies, airport and access mode choice were ignored, where the effect of this is possibly mitigated by the long haul nature of the flights presented. Furthermore, Sydney is only served by a single international airport.

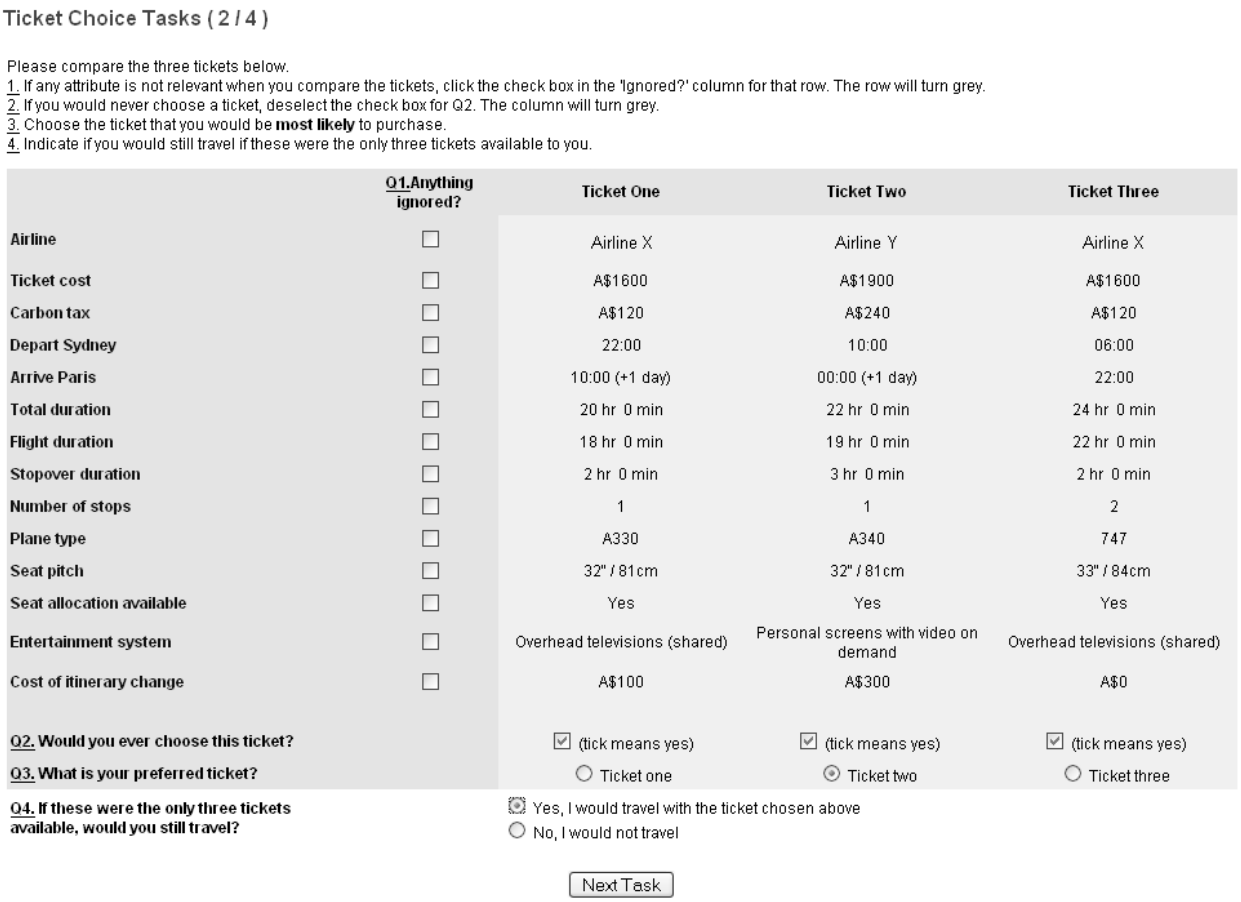
3.1 Traditional SP tasks

The SP component consisted of four choice tasks, each with three alternatives described by all of the attributes listed in Table 1. Respondents were asked to indicate their preferred flight, both with and without the 'no choice' option available. Furthermore, for each task, respondents were directed to indicate if any attributes were ignored, and were asked if some of the alternatives would never be chosen. An example of the choice screen is shown in Figure 1 (with airline names masked). A D-efficient design (see e.g., Rose and Bliemer, 2008) was used, with 18 blocks of four choice tasks each.

3.2 Interactive OTA tasks

The flights for the OTA tasks were based on real flights that were obtained from a popular real-world OTA. To prevent extensive correlations between airlines and service attributes, the plane type, seat pitch, seat allocation, entertainment system and cost of itinerary change attribute levels were not drawn from the real flights. Instead, for each attribute, each level was allocated an equal number of times. The levels were then swapped between flights such that the correlations between attributes were minimised.

Four OTA tasks were presented to the respondents, in addition to a practice task which contained four flights only. Real flight prices vary over time for the same flight due to yield management systems. Also, travel at certain times of the year will be more expensive due to high demand. Consequently, each of the four tasks represented flights at different times in the future allowing for a good coverage of flight prices over the sample. Flights were selected for departure in two weeks' time, in a month's time, in five months' time, and during the Christmas holiday season. These timeframes were randomised in presentation order across respondents and explicitly mentioned to the respondents to help them understand why the average prices varied from task to task. Figure 2 shows how the tickets appeared in the OTA tasks, with all attributes shown in this example (with the airline names masked in this screenshot).



Ticket Choice Tasks (2 / 4)

Please compare the three tickets below.
1. If any attribute is not relevant when you compare the tickets, click the check box in the 'Ignored?' column for that row. The row will turn grey.
2. If you would never choose a ticket, deselect the check box for Q2. The column will turn grey.
3. Choose the ticket that you would be **most likely** to purchase.
4. Indicate if you would still travel if these were the only three tickets available to you.

	Q1. Anything ignored?	Ticket One	Ticket Two	Ticket Three
Airline	☐	Airline X	Airline Y	Airline X
Ticket cost	☐	A$1600	A$1900	A$1600
Carbon tax	☐	A$120	A$240	A$120
Depart Sydney	☐	22:00	10:00	06:00
Arrive Paris	☐	10:00 (+1 day)	00:00 (+1 day)	22:00
Total duration	☐	20 hr 0 min	22 hr 0 min	24 hr 0 min
Flight duration	☐	18 hr 0 min	19 hr 0 min	22 hr 0 min
Stopover duration	☐	2 hr 0 min	3 hr 0 min	2 hr 0 min
Number of stops	☐	1	1	2
Plane type	☐	A330	A340	747
Seat pitch	☐	32" / 81cm	32" / 81cm	33" / 84cm
Seat allocation available	☐	Yes	Yes	Yes
Entertainment system	☐	Overhead televisions (shared)	Personal screens with video on demand	Overhead televisions (shared)
Cost of itinerary change	☐	A$100	A$300	A$0
Q2. Would you ever choose this ticket?		☑ (tick means yes)	☑ (tick means yes)	☑ (tick means yes)
Q3. What is your preferred ticket?		○ Ticket one	◉ Ticket two	○ Ticket three

Q4. If these were the only three tickets available, would you still travel?
◉ Yes, I would travel with the ticket chosen above
○ No, I would not travel

Next Task

Figure 1: Stated preference task

The top of the OTA task screens contained a set of tools that allowed respondents to sort alternatives by a given attribute, search for alternatives that satisfy certain attribute requirements, and to hide attributes as well as a description of each of the attributes. All attributes could be sorted on, with the best quality attribute shown first: lowest price, shortest duration, best entertainment system and so on. By default, the flights were sorted on price for the first choice task. Subsequent sort selections were preserved from one task to the next. Figure 2 shows an example of this part of the screen.

All attributes except for departure and arrival time could be used to refine the search. All costs and most duration times could be searched on with a respondent specified maximum. Other attributes could be searched on by choosing a category. Searches on stopover duration were limited to distinct categories that did not overlap. This was done both for simplicity and to test whether some respondents wanted a minimal stopover time while others wanted some longer time period. Any number of searches could be performed. By default, no search criteria were applied, although the final search criteria in each task were preserved for the next task.

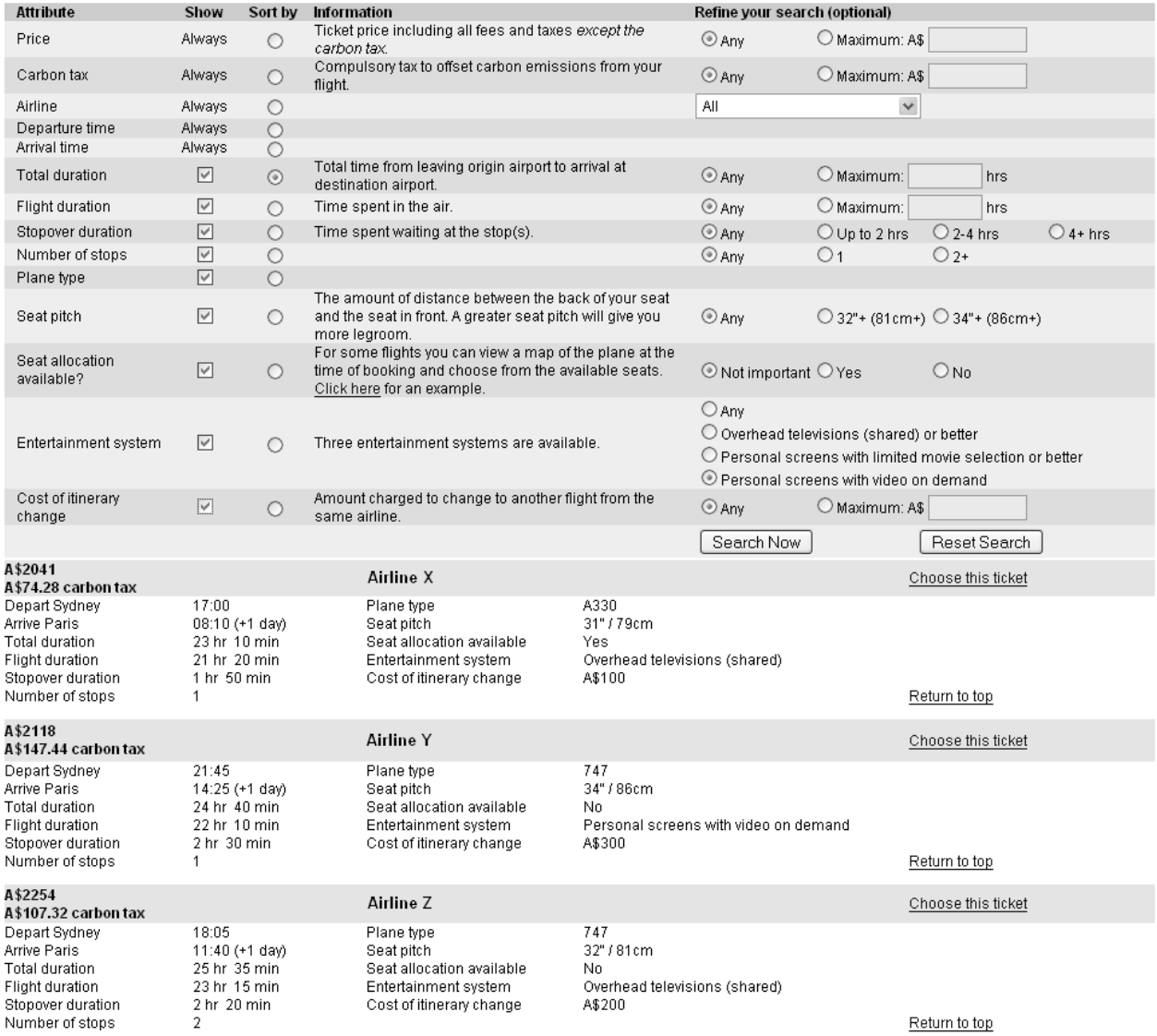


Figure 2: Search task

Price, carbon tax, airline name, departure time and arrival time were always shown. All other attributes could be hidden and shown as desired via the set of tools. This option was provided as a mechanism for respondents to remove irrelevant information from the screen so as to help facilitate easier and faster decision making on attributes that matter to the respondent. Attributes that could be hidden were not initially shown to respondents so as to force them to identify the attributes that were relevant to them in the decision making process.

In order to find out how respondents use the sort, search and show/hide tools (which we will collectively refer to as *OTA tools*), large amounts of data were captured by the survey instrument. In addition to the state of the OTA tools at the time of choice, all actions performed using the tools were captured, as was the resulting choice scenario. This information allows the analyst to examine the numerous strategies that people employ to refine their search. It is worth noting one significant difference between the OTA survey as presented to respondents and real OTA choice environments. In the latter, decision makers are required to enter preferred travel dates as part of the initial search criteria. The user can change the day of travel if the alternatives presented are not satisfactory, or if they want to compare available flights across multiple

days. Although not done here, a more complex extension of the survey instrument could include searches across days and so capture more complex search processes.

3.3 Collection of other information

In addition to the responses in the two types of tasks (SP & OTA), information was collected on how many times the respondent had travelled domestically, internationally, and to Europe over the last three years, broken down by whether the ticket had been paid for by themselves or others. The number of unique airlines flown with over the previous three years was obtained, as was information on frequent flyer membership and the usual class of ticket purchased for international flights. Finally, data was also collected on a range of socio-demographic indicators, including age, type of employment, pre-tax income, and gender.

3.4 Survey recruitment

Survey participants were recruited from an online sample of Sydney residents. To be eligible for the study, respondents were required to have travelled to Europe in the last three years, hence ensuring some degree of relevance for the experiments. Screening on the likelihood of travel in a future time period might be more suitable for future studies, especially as it is plausible (and testable) that travellers lacking recent experience will search more than experienced travellers. After screening for eligible respondents and some further data cleaning, a final sample of 462 respondents was obtained. Table 2 details the socio-demographic characteristics of the sample. Good coverage can be observed over age, work type, income and gender.

Table 2: Socio-demographics of respondents

Age		**Work type**		**Personal pre-tax income**		**Gender**	
18 to 24	77	Full time	313	Under AUD10,000	16	Male	209
25 to 34	160	Part time (< 30 hours/week)	78	AUD10,000 - AUD19,999	17	Female	253
35 to 44	105	Causal	29	AUD20,000 - AUD29,999	26		
45 to 54	62	Does not work	34	AUD30,000 - AUD39,999	39		
55 and over	58	Undisclosed	8	AUD40,000 - AUD49,999	45		
Undisclosed	0			AUD50,000 - AUD59,999	59		
				AUD60,000 - AUD79,999	69		
				AUD80,000 - AUD99,999	58		
				AUD100,000 - AUD119,999	37		
				AUD120,000 - AUD149,999	14		
				Over AUD150,000	21		
				Undisclosed	61		

4. Methodological framework

The performance of the OTA survey mechanism can be measured in three key ways. First, a model from the OTA data can be judged on its own merits, including the ability to identify systematic influences on choice that are of plausible sign, magnitude and significance. Second, the OTA model can be compared on a range of dimensions to a model estimated on the traditional SP data that is attempting to identify the same systematic influences. Finally, the OTA model can be compared to a model estimated on RP data. In the absence of RP data, we are only able to perform the first two tests in the current paper.

Our analysis first tests to see if a plausible model can be estimated from the OTA data. Analysis of each of the two survey mechanisms is limited to the multinomial logit (MNL) model. The use of more advanced model structures on this data will be the topic of future work, where the aim of the present paper is to provide an overview of the new survey approach as well as some initial empirical evidence. The effects of the Independence of Irrelevant Alternatives (IIA) assumption inherent to the MNL error structure are assumed to be minimal in the absence of any obvious correlation structure, and the impacts of the cross-sectional estimation approach were addressed by correcting the standard errors of the models using a Jackknife procedure (see Cirillo et al., 2000) to account for the panel nature of the data.

The SP and OTA MNL models can be tested independently with criteria such as overall model fit, plausible parameter sign and magnitude, and parameter significance. However, comparisons between these two models are problematic. As they are based on different data sets, direct comparison of the model outputs is not possible given possible differences in scale. Likewise, simple comparisons of the log-likelihood functions and other model fit statistics are not possible given the non-nested nature of the two data sets. However, WTP measures, taken as the ratio of two parameters, represent scale free measures that can be directly compared between the two different data sets. As such, examination of WTP outputs is a central theme of this paper.

Aside from the WTP analysis, there is however also some interest in looking at the scale differences between the two datasets, giving an indication of the relative sensitivity to changes in explanatory variables in the two datasets, with greater sensitivity leading to a more deterministic choice process. In order to test for scale differences between the two datasets, we make use of pooled estimation that allows for differences in the absolute sensitivities between the two datasets while potentially maintaining equality in the relative sensitivities (i.e. the WTP measures).

Here, we make use of an approach first proposed by Bradley & Daly (1991), and later also discussed by Hensher and Bradley (1993). This approach makes use of Nested Logit (NL) structure, and has been referred to as the 'Nested Logit trick'. It works by grouping alternatives into dataset specific nests, where normalisation of one of the inclusive value (IV) parameters allows the estimation of the remaining IV parameter to explain the extent to which the scale, and hence error variance, varies between the two datasets. Consequently, in our model, the SP alternatives were assigned to one branch, and the OTA alternatives to a second branch. For each choice observation, all alternatives viewed in the corresponding choice task were assigned to the appropriate SP or OTA branch, with no alternatives assigned to the other branch. The scale (i.e., IV) parameter of the OTA branch was normalised to one at the lowest level (i.e., RU1 normalisation) with the remaining freely estimated scale parameter associated with the SP alternatives.

Using this approach, differences in scale are not only accounted for, but are also used as a way of comparing the error variances between the two survey mechanisms. If the non-normalised scale is equal to or not statistically different from one, then the unobserved effects do not differ. If the scale is greater than one, then the SP tasks have a lower error variance than the OTA tasks. Such a finding might suggest that people cannot handle the extra complexity of the OTA tasks. If the scale is less than one, then the SP tasks have a higher error variance than the OTA tasks. This might support the claim that more information is not

in itself problematic if the presented information is deemed relevant. Section 6 will continue this discussion, armed with the findings from the study.

In addition to any differences in scale, it is also of interest to test whether, after accounting for potential scale differences, there remain any significant differences in the relative sensitivities. The validity of the assumption of taste homogeneity (after accounting for scale differences) can be assessed using a hypothesis test put forward by Swait and Louviere (1993). A χ^2 statistic for the hypothesis of preference homogeneity with β^*- 1 degrees of freedom is calculated as

$$\chi^2_{\beta^*-1} \sim -2\left[\left(LL^{DS_1} + LL^{DS_2}\right) - LL^{Joint}\right]$$

where LL^{DS_1} and LL^{DS_2} are the log likelihoods of the separate MNL models for each dataset, LL^{Joint} is the log likelihood of the combined model, and $\beta^*$is the number of generic parameters across the two data sets. This test is applied to the combined dataset, to determine whether parameters can be treated as being homogenous after accounting for scale differences.

In its most basic form, the above test is applied to the case where we have a model in which all coefficients are generic, after accounting for scale differences, and where this is compared to two dataset-specific models. However, there is clearly also a possibility that some of the coefficients are generic, after accounting for scale differences, while the homogeneity assumption is not justified for others. With this in mind, an iterative approach was used in which we individually tested the validity of the homogeneity assumption for each parameter in the pooled model. The results of this process are discussed later on in the paper.

5. Analysis and results

Before examining the results of the choice models estimated from the two datasets, it is worth taking a close look at the ways in which the sort, search, and hide/show tools were utilised by the respondents in the OTA choice tasks. Such an examination informs us about the extent to which the default presentation was customised by the respondents.

5.1 Sort behaviour

In the OTA search tasks, the flights could be sorted on any attribute, with the initial default being a sort by price. Table 3 indicates which attributes were sorted on at the time a choice was made and how many times an actual sort was explicitly performed. Since sort information is preserved between tasks, for any given attribute there may be fewer sort actions than tasks that were sorted on that attribute at the time of choice. Furthermore, since many sorts can be performed before a choice is made, there may be more sort actions than tasks that were sorted on that attribute at the time of choice. Table 3 includes both the practice search task and the four main search tasks. Of the 1,380 sort actions, 862 were performed in the practice task, which suggests that many of the sorts were performed experimentally or to establish a preferred sort preference.

Table 3: Sorting strategies

	Tasks with this sort at time of choice		Individuals with this sort at choice for all tasks		Sort actions performed	
Price	1019	44%	159	34%	539	39%
Price (by default)	793	34%	147	32%	-	
Carbon tax	63	3%	7	2%	134	10%
Airline	129	6%	17	4%	188	14%
Departure time	39	2%	5	1%	88	6%
Arrival time	43	2%	5	1%	60	4%
Total duration	45	2%	4	1%	88	6%
Flying duration	25	1%	1	0%	50	4%
Stopover duration	10	0%	0	0%	45	3%
Number of stops	8	0%	0	0%	27	2%
Plane type	7	0%	1	0%	25	2%
Seat pitch	37	2%	5	1%	33	2%
Seat reservation	24	1%	3	1%	37	3%
Entertainment system	48	2%	6	1%	39	3%
Ticket change charge	20	1%	2	0%	27	2%
Combination	-	-	100	22%	-	-
Total	2310	100%	462		1380	100%

Clearly price is the dominant sort attribute, with flights being sorted on price explicitly and by default for 78 percent of choice tasks. Cumulatively the remaining attributes account for 22 percent of sorts at choice, which is a non-trivial minority. Sort preference for these remaining attributes is roughly equal, which indicates an overall heterogeneity of sort preference. There are more sorts on airline than any other individual non-price attribute, which suggests that some respondents may have strong airline specific preferences. At the individual level, Table 3 shows that most respondents are consistent with their sort preference at time of choice. Indeed, only 22 percent varied their sort at choice over the five tasks (i.e., four 'real' and one practice).

5.2 Search behaviour

Table 4 shows, at the attribute level, the number of tasks for which a search criterion was applied at the time of choice. Whereas price was the dominant attribute for sorting, relatively few tasks included a search on price or carbon tax. Instead, searches were performed in greater numbers on the comfort attributes, including entertainment system (for 21 percent of all tasks), seat reservation (11 percent) and seat pitch (nine percent). Many searches were also performed on attributes concerned with stopovers, namely numbers of stops (eight percent) and stopover duration (seven percent).

The manner in which each attribute was searched is interesting. Some attributes have a clear preference sign, including price and entertainment system. Price limits were typically low but reasonable, ranging from AUD1800 to AUD3000 with an average of AUD2482, and entertainment system searches were evenly split between restriction to video on demand and personal screens or better. Other attributes are likely to be considered in different ways across the population. The stopover duration levels were mutually exclusive, and searches on this attribute were split between a desire to minimise time spent at a stopover (up to two hours) for 75 percent of cases, and a desire to have a more leisurely stop (2-4 hours) for 25 percent of cases. Either search strategy is plausible. The former would minimise total travel time, while the latter would provide a lengthy break from a confined environment, or perhaps provide an opportunity for shopping.

Table 4: Number of tasks with search criteria applied for each attribute at time of choice

	Number of tasks with search criteria applied at time of choice	percent
Price	96	4%
Carbon tax	36	2%
Airline	76	3%
Total duration	49	2%
Flying duration	27	1%
Stopover duration	167	7%
Number of stops	187	8%
Seat pitch	198	9%
Seat reservation	258	11%
Entertainment system	476	21%
Ticket change charge	40	2%
All tasks	2310	100%

Unlike sort selections, search criteria can be applied across multiple attributes concurrently. An analysis of the data showed that 18.3 percent of all tasks were completed with multiple search criteria applied. It is with these complex searches that the search tool is most useful. If only one search criterion is applied, it might be quicker to just perform a sort. However, the sort tool is cumbersome and ineffective if more than one attribute is deemed to be of importance.

Whereas sort actions only reorder the flights on screen, search actions actually add or remove flights from view. This makes a search a stronger form of filter, as any flight that fails to meet the search criteria cannot be chosen. These reductions are quite large in absolute terms when some search tasks contain 22 potential flights. On average, the choice set size after applying search criteria was reduced to seventy-three percent of its original size, where for a quarter of respondents, it was reduced to under forty percent of its original size.

5.3 Showing and hiding of attributes

The price, carbon tax, airline, departure time and arrival time attributes were always visible and could not be hidden. All other attributes were not shown by default and had to be actively chosen for display. As evidenced by Table 5, none of these attributes were shown for more than half of the tasks, with the least shown attribute being ticket change charges. At the individual level, 37 percent of respondents did not show any of the additional attributes for any of their tasks at the time of making their choice. This may have been due to satisfaction with the default attributes as the sole means of ticket differentiation, for example with highly price sensitive respondents. It also may have been due in part to a lack of engagement with or understanding of the survey task.

Table 5: Number of tasks with attributes shown

	Number of tasks with attribute shown	Percentage of all tasks
Total duration	1034	45%
Flying duration	914	40%
Stopover duration	945	41%
Number of stops	1023	44%
Seat pitch	744	32%
Seat reservation	824	36%
Entertainment system	951	41%
Ticket change charge	698	30%

5.4 Model results for the individual datasets

As a starting point, separate MNL models were estimated on each dataset. The final sample consisted of 462 respondents, each with one practice and four real choice tasks for each of the two survey interfaces. The observations from the practice tasks were not used in the analysis. Additionally, seven SP observations were removed due to data corruption, and six OTA observations were removed due to realism issues with the presented fares. Only those OTA flights visible at the time of choice were included, so that the flights removed by the search tool did not enter the utility expressions. Similarly, only those OTA attributes that were visible at the time of choice were included, so that the attributes that were not chosen to be shown did not enter the utility expressions.

The model results are listed in Table 6. Although care must be taken when comparing ρ^2 values for different datasets, the OTA model can be seen to have a much higher ρ^2 value than the SP model. Most parameter estimates in the OTA model are of equal or higher statistical significance than their SP equivalent, including carbon tax, which is highly significant in the former, but only marginally significant in the latter. Additionally, the charge for a flight change is strongly significant in the OTA model, but not significant in the SP model. One key difference between the OTA and traditional SP choice tasks lies in the ability of the respondent to sort the alternatives in the former. To account for this, additional dummy variables were created representing the order that an alternative appears on the final screen used when the respondent made their choice. An option appearing as one of the first eight alternatives shown has a higher likelihood of being chosen than those shown after eight, *ceteris paribus*, with diminishing impacts within the first eight as one moves from the first shown to the eighth shown. Only the first eight order dummies are reported in Table 6, as the remaining 13 dummies are not statistically significant. The inclusion of these constants may be criticised on endogeneity or self-selection grounds. Indeed, a respondent who ranks the options by travel time is likely to be more travel time sensitive and will as a result also be more likely to choose the higher ranked options. However, our analysis showed not only the expected substantial improvements in fit when including these constants, but also produced more reliable underlying sensitivities. Here, it can be argued that the inclusion of the constants also captures lexicographic or apparent lexicographic behaviour, and the absence of a treatment of this would have had an undue influence on model estimates (cf. Hess et al., forthcoming).

Table 6: Model results

	SP data		OTA data		Combined data						
					$\beta_{SP}=\beta_{OTA}$	Homogeneous		SP		OTA	
	Par.	**(*t*-ratio)**	**Par.**	**(*t*-ratio)**	**(*t*-ratio)**	**Par.**	**(*t*-ratio)**	**Par.**	**(*t*-ratio)**	**Par.**	**(*t*-ratio)**
Price	-0.0020	(-21.22)	-0.0034	(-9.42)	(0.46)	-0.0033	(-13.91)	-	-	-	-
Carbon tax	-0.0004	(-1.96)	-0.0033	(-10.47)	(4.64)	-	-	-0.0008	(-1.97)	-0.0033	(-10.20)
Charge for flight change	-	-	-0.0020	(-4.41)	-	-	-	-	-	-0.0020	(-4.37)
Travel time	-0.001	(-3.72)	-0.001	(-4.02)	(0.15)	-0.001	(-6.56)	-	-	-	-
Number of stops	-0.244	(-4.56)	-0.446	(-3.21)	(0.02)	-0.411	(-5.05)	-	-	-	-
Seat pitch	0.055	(2.51)	0.335	(6.97)	(4.35)	-	-	0.090	(2.34)	0.333	(8.14)
Seat assignment	0.080[1]	(2.13)	0.252[6]	(4.55)	(1.88)	0.183[10]	(5.18)	-	-	-	-
Entertainment (shared)	-0.227[2]	(-6.36)	-0.371[7]	(-5.55)	(0.16)	-0.371[11]	(-8.09)	-	-	-	-
Airline constant 1	-0.137[3]	(-3.14)	-0.212[8]	(-7.02)	(1.00)	-0.215[12]	(-7.44)	-	-	-	-
Airline constant 2	-0.223[3]	(-3.56)	-0.430[8]	(-7.17)	(0.56)	-0.413[12]	(-7.75)	-	-	-	-
Airline constant 3	-	-	-0.845[8]	(-5.44)	-	-	-	-	-	-0.842[12]	(-6.59)
FF constant 1	0.577[4]	(4.47)	0.913[9]	(6.82)	(0.00)	0.919[13]	(10.46)	-	-	-	-
FF constant 2	0.298[4]	(6.37)	0.325[9]	(6.61)	(1.79)	0.349[13]	(8.62)	-	-	-	-
Arrive (9pm – midnight)	-0.109[5]	(-2.44)	-	-	-	-	-	-0.171[14]	(-2.59)	-	-
Arrive (1am)	-0.310[5]	(-3.84)	-	-	-	-	-	-0.480[14]	(-3.82)	-	-
1st alt. shown	-	-	3.363	(4.29)	-	-	-	-	-	3.402	(4.74)
2nd alt. shown	-	-	2.771	(3.49)	-	-	-	-	-	2.800	(3.90)
3rd alt. shown	-	-	2.362	(2.93)	-	-	-	-	-	2.377	(3.31)
4th alt. shown	-	-	2.230	(2.75)	-	-	-	-	-	2.245	(3.12)
5th alt. shown	-	-	1.684	(1.97)	-	-	-	-	-	1.705	(2.35)
6th alt. shown	-	-	1.741	(2.14)	-	-	-	-	-	1.764	(2.43)
7th alt. shown	-	-	1.690	(2.14)	-	-	-	-	-	1.702	(2.34)
8th alt. shown	-	-	1.661	(2.07)	-	-	-	-	-	1.668	(2.30)
SP alternative 1	0.148	(2.67)	-	-	-	-	-	0.254	(2.36)	-	-
SP alternative 2	0.209	(3.53)	-	-	-	-	-	0.350	(3.23)	-	-
Scale (SP)	-	-	-	-	-	-	-	0.610	0:(12.22) 1:(-7.81)	1	Fixed
Model fits											
LL(0)	-2022.545		-5693.70		-7716.25						
LL(β)	-1706.689		-3263.08		-4973.21						
Number of parameters	15		35		41						
ρ^2	0.156		0.427		0.355						
Adjusted ρ^2	0.149		0.416		0.348						
Observations	1841		1842		3683						
Respondents	462		462		462						

Base levels of effects codes: [1]No seat assignment (-0.080), [2]VOD/personal screen (0.227), [3]Airline constant 4(0.360), [4]No membership(-0.875), [5]Other times (0.420), [6]No seat assignment (-0.252) , [7]VOD/personal screen (0.371), [8]Airline constant 4(1.486)', [9]No membership(-1.237), [10]No seat assignment (-0.183), [11]VOD/personal screen (0.371), [12]Airline constant 4 (1.470), [13]No membership(-1.268), [14]Other times (0.651)

A number of qualitative attributes were effects coded in estimation, including seat assignment, entertainment, arrival time, and airline. Effects coding was chosen over dummy coding to prevent the base level from being confounded with the alternative specific constants. Care must be taken when interpreting the effects coded parameters, as the base level of utility is not zero, but rather $-\sum_i \beta_i$. No aircraft effects were retrieved for either of the datasets, suggesting that the respondents were indifferent between flying on a 747, 777, A330 or A340. Respondents were given the choice of three different entertainment system levels; shared screens (shared), personal screens (personal) or personal screens with video on-demand (VOD). In estimating the model, VOD was treated as the base attribute level. As would be expected, in both datasets, shared entertainment was less preferred than personal entertainment or VOD. However, personal entertainment was not statistically significant in either dataset and so personal and VOD effectively collapsed to form a single base level. The impact of arrival times was only significant in the SP data[1]. After extensive testing of alternate groupings of arrival times, effects coded levels of 9pm-midnight and 1am were generated, with all other times forming the base level. The results show that arrival at these times was viewed unfavourably.

All airlines were initially effects coded, however the parameters for five airlines were not statistically different from each other. As such, for reasons of parsimony these airlines were combined, with a single associated parameter, 'Airline constant 1', estimated. 'Airline constant 2' is a Middle East carrier with little market presence. 'Airline constant 3' is a Chinese carrier that was only presented in the OTA choice tasks. The base level, 'Airline constant 4', is comprised primarily of four airlines, all of which could be considered premium carriers, and three of which have a strong presence on the routes used in the study.

The survey contained questions to capture which, if any, relevant airlines the respondent had frequent flyer (FF) membership for, either through airline or alliance programs. Effects coded interaction dummies were created for each FF program and airline of interest. For every flight alternative that a respondent viewed, the corresponding FF program and airline interaction dummy was set to one if that respondent was a member of the airline's FF program, or a FF program of an associated alliance. Thus, each FF program interaction dummy parameter represents the mean sample utility associated with a flight for which the respondent has FF membership. As with the airlines, the FF parameters were combined when not statistically different. 'FF constant 1' represents two prominent Asian carriers, 'FF constant 2' comprises eight airlines, and the base level ('FF base'), can be considered as having no membership for the FF program of the airline in question. That airlines represented in FF constants 1 and 2 differ in the effect of their FF program is curious. Several interpretations are possible, if not identifiable. Airlines under FF constant 1 may have more rewarding FF programs, where more or higher value points may be earned on a return flight from Sydney to Europe than with the other airlines. Alternatively, utility that would otherwise be captured in the airline parameters is captured in the FF interaction (i.e. for those with FF membership) to a greater extent with FF constant 1. The difference here is between the impact of the FF program per se, and the preferences of those who for whatever reason are FF members.

5.5 Model results for the combined dataset

Table 6 also presents the results from a combined model estimated using the NL trick detailed in Section 4. Before settling on a final combined data model to analyse, it was necessary to determine which parameters could be treated homogeneously across the two data sets. To this extent, an initial model was estimated in which all those parameters that were applicable to both datasets were treated as homogeneous. This model was then compared to the two individual models, leading to a χ^2 test statistic of 38.66 with 13 degrees of freedom, where, with a critical

1 Arrival times were not adequately handled in the experimental design of the OTA tasks, and we believe that this was why arrival times were only significant in the SP data. Section 6 provides further discussion of problems associated with using real market data in the survey.

level of 22.36 at the 95 percent confidence level, the hypothesis of complete preference homogeneity was rejected. We then gradually relaxed the homogeneity assumption, finally leading to a model where we obtain a $\chi 2$ test statistic of 6.88 with 8 degrees of freedom, where the corresponding 95 confidence critical level is 15.51, meaning that the assumption of homogeneity (for the remaining affected coefficients) is no longer rejected.

This final model is documented in Table 6, where its composition can be gleaned from the allocation of each parameter to the last three columns of the table. Where appropriate, tests of statistical difference are presented in the first column related to this model, where a separate model was estimated for each homogeneous parameter, treating it as heterogeneous to calculate the appropriate t-ratio. Homogenous parameters and their t-ratios are listed in the homogenous column only, while heterogeneous parameters are listed in one or both of the SP and OTA columns. While many comparable parameters are homogeneous, several are not, namely 'Carbon tax', 'Seat pitch', 'Charge for flight change', 'Arrive 9pm-midnight' and 'Arrive 1am'. That the latter three parameters are not homogeneous is not surprising, since none of them are significant in both MNL models. By contrast, 'Carbon tax' and 'Seat pitch' are significant in both MNL models, and their heterogeneity across the datasets cannot be readily explained. The significance of the homogeneous parameters typically represents an improvement on the significance of the respective parameters in the dataset specific MNL models. This is not surprising, as these parameters are estimated with more observations. Using the same reasoning, we could anticipate the observed similarity in significance of the heterogeneous parameters with their counterparts in the corresponding MNL model.

The scale parameter for the SP data set is significantly different to both zero and one. Since the scale parameter is normalised to one for the OTA data set, the latter test is the crucial one. The scale parameter of 0.61 for the SP alternatives implies that the SP choice tasks exhibit greater error variance than the OTA choice tasks. Section 6 contains a full discussion of the differences in error variance, as well as the homogeneity of most parameters.

5.6 Monetary valuations

Table 7 shows the monetary valuation values derived from the combined model, together with the associated *t*-ratios and 95 percent confidence intervals (CIs). The latter two measures were calculated using the Delta method. For those attributes for which a homogenous parameter was estimated, a single measure is presented, while, for the remaining attributes, a valuation is presented for each dataset in which the attribute's parameter (i.e., the numerator) is significant, where the price parameter (i.e., the denominator) is homogenous. The use of effects coding for some attributes has consequences for their associated monetary valuations. Effects coding of attributes prevents the base level from being confounded with the alternative specific constant, and the base levels of other effects coded attributes. Consequently, the effects coded attributes have monetary valuation values for *all* attribute levels, including the base level. For interpretation reasons, the WTP to move between any two levels of an attribute is thus the difference in the corresponding valuations.

Some of the levels used in Table 7 are desirable levels, while others are undesirable. For the former, we thus present a willingness to pay for improvement, while for the latter, we present the required monetary incentive, i.e. the drop in fare that is necessary for this level or change to become acceptable.

All monetary valuations are significant at the 95 percent confidence level, and the estimates appear plausible in the context of a return economy airfare, where the journey in each direction typically takes about 24 hours, and involves one or two stops. For example, respondents were prepared to pay $126.10 to avoid a stop, $27.24 to avoid an hour of travel time, $112.30 to be able to select a seat before the flight, and $227.70 to have personal screens or video on demand available instead of shared screens. While the parameters underlying the effects coded valuations are statistically different, a lack of overlap in the CIs gives us further confidence that the values themselves are statistically different. There is no overlap in the CIs of the four airline

groups or the three frequent flyer groups. By contrast, there is some overlap in the CIs of arrival at 1am and 9pm to midnight.

Table 7: Monetary valuation results from the combined model

Attribute	Data set	Willingness to pay (improvement)	Willingness to pay (avoiding change)	(*t*-ratio)	C.I. lower	C.I. upper
Travel time (hour)	Homogeneous		$27.24	(5.61)	$17.72	$36.76
Number of stops	Homogeneous		$126.1	(5.39)	$80.21	$171.99
Seat pitch (inch)	SP	$27.67		(2.28)	$3.92	$51.41
Seat pitch (inch)	OTA	$102.29		(4.72)	$59.78	$144.79
Seat assignment (yes)	Homogeneous	$56.15		(4.26)	$30.32	$81.98
Seat assignment (no)	Homogeneous		$56.15	(4.26)	$30.32	$81.98
Entertainment (shared)	Homogeneous		$113.85	(6.90)	$81.51	$146.19
Entertainment (personal & VOD)	Homogeneous	$113.85		(6.90)	$81.51	$146.19
Airline constant 1	Homogeneous		$66.08	(6.39)	$45.82	$86.34
Airline constant 2	Homogeneous		$126.72	(7.07)	$91.61	$161.83
Airline constant 3	Homogeneous		$258.20	(5.80)	$171.00	$345.41
Airline constant 4	Homogeneous	$451.00		(8.19)	$343.08	$558.92
FF constant 1	Homogeneous	$281.97		(6.57)	$197.91	$366.04
FF constant 2	Homogeneous	$106.99		(6.15)	$72.92	$141.07
FF base	Homogeneous		$388.97	(7.89)	$292.29	$485.65
Arrive (9pm – midnight)	Homogeneous		$52.53	(2.57)	$12.44	$92.62
Arrive (1am)	Homogeneous		$147.17	(3.61)	$67.24	$227.10
Arrive (any other time)	Homogeneous	$199.70		(11.46)	$165.55	$233.85

6. Discussion and conclusions

This paper has discussed the findings of a study making use of an innovative survey environment for investigating air travel choice behaviour. By mimicking the interface of an OTA, we are able to boost realism and capture additional information on how people handle choice environments that contain large amounts of information. The findings of this work have implications not just for the study of air travel behaviour but also other areas where extensive use is made of SP surveys to study complex travel choices.

The initial parts of the analysis showed that extensive use was made by respondents of the sort, search and hide/show tools, with the resulting choice task dimensions varying greatly across respondents. The large reduction in choice set size that resulted from some of the searches demonstrates the extent to which respondents are prepared to definitively eliminate alternatives in a non-compensatory fashion, when a large number of alternatives are available. Additionally, the mix of attributes chosen to be visible varied greatly over the respondents. For some respondents, the level of interaction was clearly more minimal, with no additional attributes being shown, no search criteria applied, and the default sort on price being retained. While this behaviour may signify a lack of engagement with the survey, it is also a plausible decision strategy. A respondent who is not prepared to pay much more than the cheapest fare would only need to examine the first few alternatives on the screen, as these flights would represent the cheapest available. The order of the flights could of course have been randomised by default, as this would have helped us distinguish between disengaged and price conscious respondents, but this would have moved us away from the standard approach of real world OTAs.

In the actual choice modelling analysis, dataset specific models were estimated alongside a pooled model that accounts for scale heterogeneity between the two datasets, while also imposing a homogeneity assumption (after scale differences) only for some of the parameters, following extensive testing. Here, most parameter estimates and WTP measures were found to be homogenous across the two choice environments, suggesting that a move away from a realistic choice environment to a traditional SP environment may not change the behavioural outputs of the model. Nonetheless, the OTA data had significantly lower error variance,

suggesting that a complex choice environment is not necessarily problematic, as has generally been suggested. We argue that the reduction in error variation is the product of the ability to adjust the choice environment and make it more relevant. This in turn leads to more consistent choices that can be more readily estimated, resulting in lower error variance. The survey methodology thus shows promise as a viable alternative to traditional SP surveys for capturing preference in a variety of choice scenarios, e.g. choice of train, hotel, car hire and consumer durables. Conversely however, the overall homogeneity in WTP estimates between the two treatments makes a contribution to the extensive debate about the external validity of SP choice experiments by suggesting that the limited realism of the SP grid format does not necessarily bias the behavioural results. Nevertheless, the lower error variance of the OTA tasks is appealing, and would suggest that, *ceteris paribus*, the sample size required to determine the underlying preferences would be lower for the OTA survey than for the traditional survey. However, no definitive conclusions can be drawn from a single study, and further research is needed in both the OTA and other contexts. In particular, the large number of alternatives, and the search, sort and show tools, which make sense and are readily applicable in the OTA context, may be less appropriate in other contexts, such as mode choice.

Several avenues exist for future research. The availability of each of the OTA tools could be varied, and the impact on preference homogeneity and scale differences observed. In this paper, the dataset was the only systematic influence on error variance. A more nuanced understanding of the interaction between the use of the OTA tools and the structure of the error variance would be valuable. Finally, as mentioned earlier, the methodology could be readily applied to other settings, including choice of train, hotel, car hire and consumer durables.

Acknowledgements

The authors would like to thank Jun Zhang for his work in coding the internet survey. The third author acknowledges the support of the Leverhulme Trust in the form of a *Leverhulme Early Career Fellowship*.

References

Arentze, T., Borgers, A., Timmermans, H. and DelMistro, R. (2003). Transport stated choice responses: effects of task complexity, presentation format and literacy. *Transportation Research Part E*, 39(3), 229–244.

Bliemer, M.C., Rose, J.M. and Beelaerts van Blokland, R. (2009) Experimental Design Influences on Stated Choice Outputs, *European Transport Conference*, Leeuwenhorst, October 5-7.

Bradley, M. A. and Daly, A. J. (1991) Estimation of Logit Choice Models using Mixed Stated Preference and Revealed Preference Information, presented to 6th. International Conference on Travel Behaviour, Québec

Bradley, M.A. (1998). Behavioural models of airport choice and air route choice. In: Ortuzar, J. de D., Hensher, D.A. and Jara-Diaz, S.R. (eds.) *Travel behaviour research: updating the state of play (IATBR 94)*, Elsevier, Oxford, 141–159.

Brownstone, D. and Small, K. (2005). Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transportation Research Part A*, 39(4), 79-293.

Burke, R.R., Harlam, B.A., Kahn, B.E. and Lodish, L.M. (1992). Comparing Dynamic Consumer Choice in Real and Computer-Simulated Environments. *Journal of Consumer Research*, 19(1), 71–82.

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(2), 179-192.

Carson, R., Louviere, J.J., Anderson, D., Arabie, P., Bunch, D., Hensher, D.A, Johnson, R., Kuhfeld, W., Steinberg, D., Swait, J., Timmermans, H., and Wiley, J. (1994). Experimental Analysis of Choice, *Marketing Letters*, 5(4), 351-367

Caussade, S., Ortuzar, J. de D., Rizzi, L.I. and Hensher, D.A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B*, 39(7), 621-640.

Chin, A.T.H. (2002). Impact of frequent flyer programs on the demand for air travel. *Journal of Air Transportation*, 7(2), 53–86.

Cirillo, C., Daly, A. and Lindveld, K. (2000) Eliminating bias due to the repeated measurements problem in SP data. Ortuzar, J. de D. (ed.) *Stated Preference Modelling Techniques: PTRC Perspectives 4*, PTRC Education and Research Services Ltd., London.

Clemons, E. K., Hann, I. H. and Hitt, L.M. (2002). Price dispersion and differentiation in online travel: An empirical investigation. *Management Science,* 48(4), 534-549.

DeShazo, J.R. and Fermo, G. (2002). Designing choice sets for stated preference methods: the effects of complexity on choice consistency. *Journal of Environmental Economics and Management*, 44(1), 123-143.

Hensher, D.A. (2006). How do respondents process stated choice experiments? Attribute consideration under varying information load. *Journal of Applied Econometrics* 21(6), 861-878.

Hensher, D.A. (*forthcoming*). Hypothetical bias, choice experiments and willingness to pay, in press, *Transportation Research Part B.*

Hensher, D.A. and Bradley, M. (1993). Using Stated Response Choice Data to Enrich Revealed Preference Discrete Choice Models. *Marketing Letters*, 4(2), 139-151.

Hensher, D.A., Rose, J.M. and Greene, W.H. (2005) *Applied Choice Analysis: A Primer*, Cambridge University Press, Cambridge.

Hensher, D.A., Rose, J. and Bertoia, T. (2007). The implications on willingness to pay of a stochastic treatment of attribute processing in stated choice studies. *Transportation Research Part E*, 43, 73-89.

Hensher, D.A., Stopher, P.R. and Louviere, J.J. (2001). An exploratory analysis of the effect of numbers of choice sets in designed choice experiments: an airline choice application. *Journal of Air Transport Management*, 7(6), 373–379.

Hess, S., Adler, T. and Polak, J.W. (2007). Modelling airport and airline choice behaviour with stated-preference survey data. *Transportation Research Part E*, 43(3), 221–233.

Hess, S. (2007). Posterior analysis of random taste coefficients in air travel choice behaviour modelling. *Journal of Air Transport Management*, paper accepted for publication, 13(4), 203-212.

Hess, S. and Hensher, D.A. (*forthcoming*). Using conditioning on observed choices to retrieve individual-specific attribute processing strategies, *Transportation Research Part B.*

Hess, S. and Rose, J.M. (2009) Lessons in stated choice survey design *European Transport Conference*, Leeuwenhorst, October 5-7.

Hess, S., Rose, J.M. and Polak, J.W. (*forthcoming*). Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transportation Research Part D.*

Kanafani, A. and Sadoulet, E. (1977). The partitioning of long haul air traffic – a study in multinomial choice. *Transportation Research*, 11(1), 1–8.

Louviere, J.J. and Timmermans, H.J.P. (1990). Hierarchical information integration applied to residential choice behaviour. Geographical Analysis, 22(2), 127–145.

Louviere, J.J., Hensher, D.A. and Swait, J.D. (2000) Stated Choice Methods: Analysis and Application, Cambridge University Press, Cambridge.

Lanscar, E. and Louviere, J.J. (2008). Conducting discrete choice experiments to inform healthcare decision making: A user's guide. Pharmacoeconomics, 26, 661-667.

Law, R., Leung, K. and Wong, R.J. (2004). The impact of the Internet on travel agencies. International Journal of Contemporary Hospitality Management, 16(2), 100-107.

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.

PhoCusWright (2007). PhoCusWright's U.S. Online Travel Overview Seventh Edition, November 2007.

Proussaloglou, K. and Koppelman, F.S. (1995). Air carrier demand: an analysis of market share determinants. Transportation, 22(4), 371–388.

Proussaloglou, K. and Koppelman, F.S. (1999). The choice of air carrier, flight, and fare class. Journal of Air Transport Management, 5(4), 193–201.

Puckett, S.M. and Hensher, D.A. (2008). The role of attribute processing strategies in estimating the preferences of road freight stakeholders. Transportation Research Part E, 44, 379-395.

Rose J.M. and Bleimer, M.C.J. (2008). Stated Preference Experimental Design Strategies. In Hensher, D.A. and Button, K.J. (eds) Handbook of Transport Modelling, Pergamon Press, Oxford.

Rose, J.M., Hensher, D.A. and Greene, W.H. (2005). Recovering costs through price and service differentiation: Accounting for exogenous information on attribute processing strategies in airline choice. Journal of Air Transport Management, 11, 400-407.

Rose, J.M. and Hensher, D.A. (2006). Accounting for 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.

Smith, B. C., Darrow, R., Elieson, J., Guenther, D., Rao, B. V. and Zouaoui, F. (2007). Travelocity becomes a travel retailer. Interfaces, 37(1), 68-81.

Swait, J. and Louviere, J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of Marketing Research, 30(3), 305-314.

Wang, D., Jiuqun, L. and Timmermans, H.J.P. (2001). Reducing respondent burden, information processing and incomprehensibility in stated preference surveys: principles and properties of paired conjoint analysis. Transportation Research Record 1768, 71–78.

Wardman, M. (2001). A review of British evidence on time and service quality Valuations. Transportation Research Part E, 37(2-3), 91-106.