Article:
Collins, A, Rose, JM and Hess, S (2012) Interactive stated choice surveys: a study of air travel behaviour. Transportation, 39 (1). 55 - 79 . ISSN 0049-4488

https://doi.org/10.1007/s11116-011-9327-z

Interactive stated choice surveys: a study of air travel behaviour

ANDREW T. COLLINS[1,*], JOHN M. ROSE[1] & STEPHANE HESS[2]
[1]*Institute of Transport and Logistics Studies, Faculty of Economics and Business, The University of Sydney, NSW 2006, Australia;* [2] *Institute for Transport Studies, University of Leeds*
(Author for correspondence, E-mail: andrew.collins@sydney.edu.au)*

Key words: Airline choice, stated choice experiments, information search, survey realism, online travel agent

Abstract

Stated preference (SP) experiments are becoming an increasingly popular survey methodology for investigating travel behaviour. Nevertheless, some evidence suggests that SP experiments do not mirror decisions in real markets. With an increasing number of real world decisions made using the internet, an opportunity exists to improve the realism of the SP counterparts of such choices by aligning the choice environment with such online portals. In this paper, we illustrate the benefits of such an approach in the context of air travel surveys. Our 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. 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 compared to a more conventional SP experiment. These results have implications for the study of other complex travel choices where interactive surveys may similarly be preferable to standard approaches.

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 Brownstone 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 often been found to exceed those observed in real markets (see e.g., Hensher, 2010, for a detailed overview of differences obtained between WTP values from different survey methodologies). These differences, whichever direction they may go in, are a possible cause for serious concern. Indeed, at stake is the external validity of the data collected via SP experiments and hence confidence in the findings emanating from these data.

Given the divergence of evidence in terms of the existence and direction of any bias, there is particular research interest in first determining 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 approaches may be deployed to bridge this 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. Often, this means presenting respondents with a large amount of information, in line with real world choice settings. 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). On the other hand, 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. 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 either deterministically (Rose et al., 2005; Puckett et al., 2008), or stochastically (Hensher et al., 2007; Hess and Hensher 2010).

What has however received less attention is the actual setting in which the choices are to be made, in other words the presentation, and how this differs from the way in which such decisions are made in a

real world setting. In particular, the way in which alternatives in a SP context are presented in a table listing a number of different attribute values bears little or no resemblance to real world decisions. Clearly, some transport decisions are difficult to mimic in a hypothetical context, e.g. choosing a route, or buying a car. However, an increasing number of 'real world' decisions are made using internet sales channels, which by their nature also make the decision process more 'artificial'. In other words, 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. This presents an opportunity to at least for those kinds of decisions improve the presentation of SP scenarios to align them more with their 'real world' counterparts. 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.

A prominent example in this context is that of air travel behaviour, where online booking engines now account for a majority of all decisions. Users of these websites are routinely faced with a very large number of alternatives, described by numerous attributes. 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. On the other hand, the majority of SP studies in this context still rely on “tried and tested” approaches to presentation, with a very limited number of alternatives and attributes, and no control over the level of detail of the presentation.

The present paper aims to change this by developing a survey for air travel behaviour that mimics the environment of an online booking engine. Alongside the improvement in presentational realism, our survey, in line with real world counterparts, also allows the respondent to control the amount of information in the choice task in a way that is meaningful to them. 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. To allow us to test our assumptions, 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. In this way, we are able to examine whether different results arise from the use of a more realistic choice

environment when compared to a traditional choice experiment. Our results not only showed improved parameter estimates, but also indicate the presence of a lower level of error in our models.

It is worth briefly mentioning the motivation for using a SP experiment modelled on a real world setting, rather than using actual real world data. Indeed, data collected from a real world booking engine may have a high level of response quality, but will be 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), as reflected for example in the RP air travel behaviour studies in Pels et al. (2001) and Hess & Polak (2006a,2006b).

We accept that the ultimate test of realism would be a comparison with real world data, or a validation of our results on such data. Whilst of immense interest, in common with the majority of other SP studies, comparing results with real market data as a test of external validity is however beyond the scope of the current paper (due to availability of suitable data), but remains an important area for future work. We simply strive to show that better performance is obtained in comparison with a traditional SP experiment, based on the ability to achieve meaningful and well estimated parameter values as well as in any reduction in the variance of the random part of utility.

The remainder of this paper is organised as follows. 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

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. Within the marketing literature, Johnson et al. (2004) examined the extent to which search was performed across different websites for several product categories, where one category, travel, included OTA and airline websites. While search levels were low and decreasing over time, search within a website, which in the case of an OTA can span many airlines, was not examined. Outside of the domain of air travel behaviour modelling, Moe (2006)

demonstrates the potential of clickstream data collected from the internet, estimating a two-stage choice model that utilises product view and purchase data to model both of these choices.

Despite the lack of OTA style applications, 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. Theis et al. (2006) examined the impact of connection time at hubs. 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 (2008) 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). Ortuzar and Simonetti (2008) combined SP and RP data in a multimodal choice context. 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.

3. Survey description

The SP scenarios in the current study ask respondents to choose a ticket for return 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 (the distance between seats) 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 choice scenario was also framed in terms of a return ticket, as this is consistent with leisure travel and avoids the complication of multi-stop travel. In the market, purchase of a return ticket requires a decision about both the departing and return flights to be made. To simplify matters, we only required a choice for the departing flight and asked the respondent to assume that the return flight would have similar service levels. For similar reasons, mixed carrier tickets were not presented.

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. Given the likely change over time in the functionality of OTAs, 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 better use 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 typically not included, and would usually need to be sourced from a different website.

Both survey components (i.e., traditional SP and OTA style SP) were shown to each respondent, with the order of the two components randomised, as well as 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[1].

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 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. 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.

[1] 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**

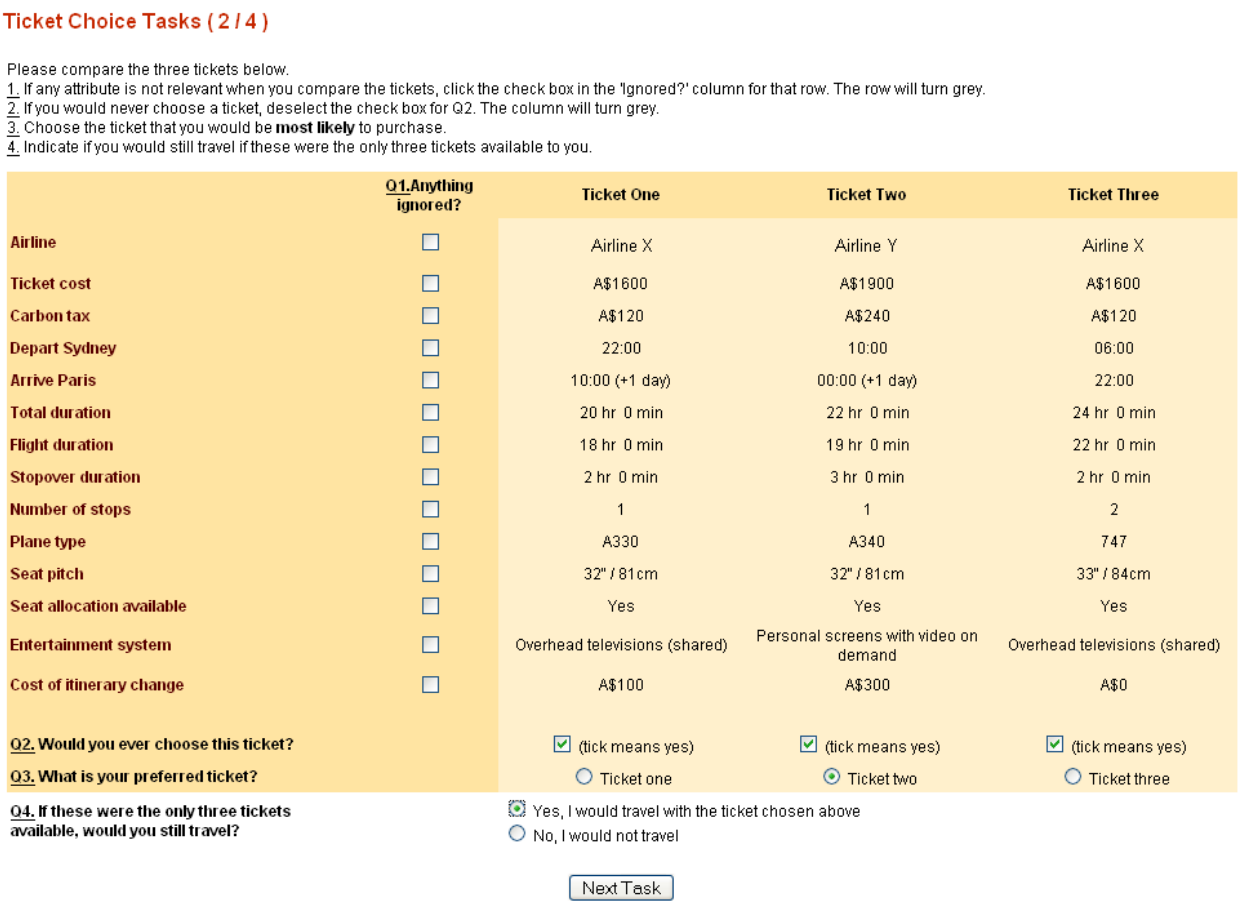
Attribute	**SP levels**	**OTA levels or range**	**OTA: From real flight?**	**Typical online travel agent attribute?**
Price	AUD1600, AUD1900, AUD2200, AUD2500	AUD1809 – AUD6036	Yes	Yes
Carbon tax	AUD0, AUD120, AUD240, AUD360	AUD0 – AUD460.76	No	No
Airline	9 possible	13 possible	Yes	Yes
Departure time	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

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 OTA. To prevent extensive correlations between airlines and service attributes, the plane type, seat pitch, seat allocation (in this study, defined as the ability to select a seat at the time of booking), 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.


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**

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).

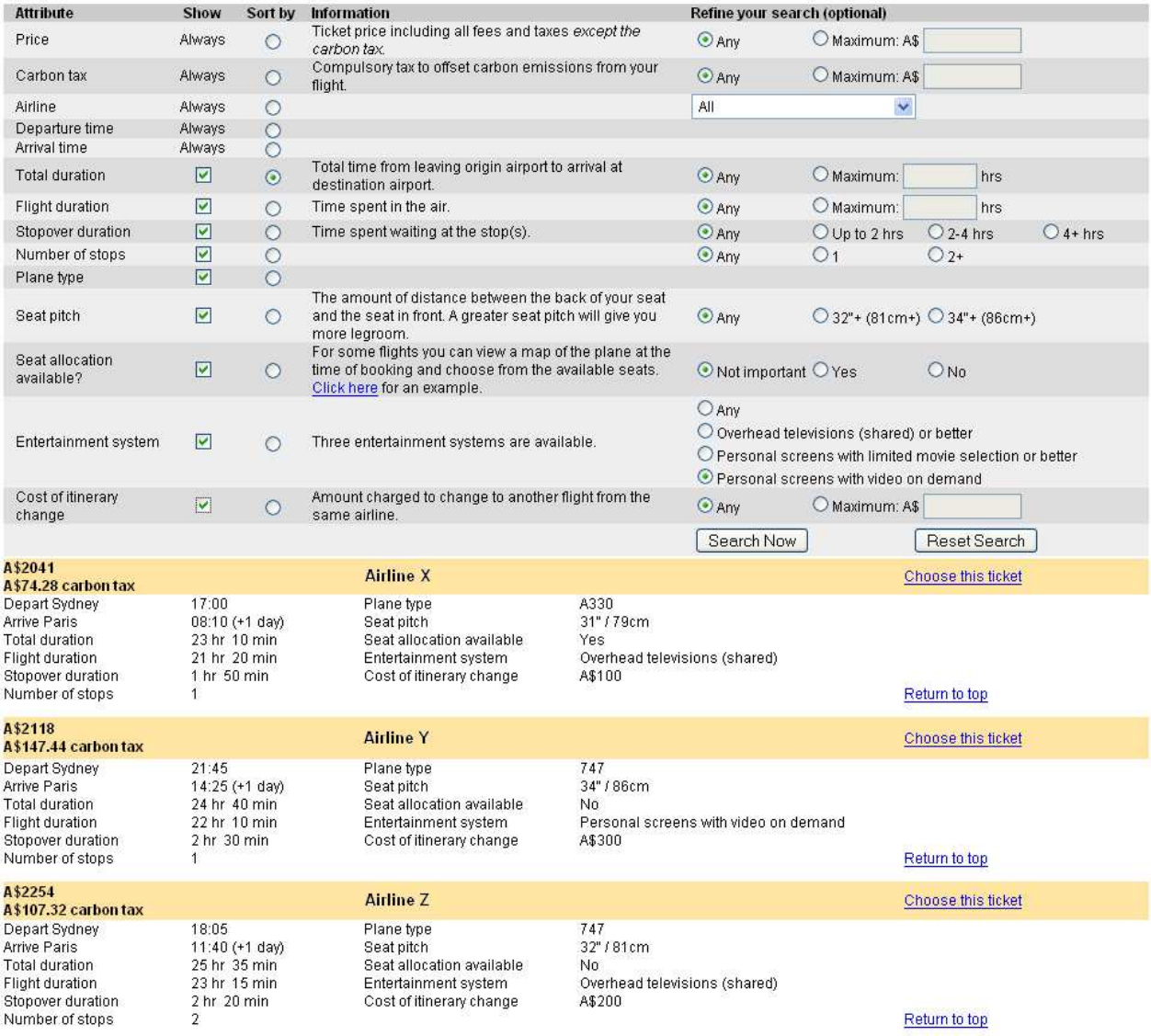


Figure 2: **Search task**

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 alternative with the best value for that attribute being shown first, i.e., 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.

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 and 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 the usual class of ticket purchased for international flights. Data was also collected on a range of socio-demographic indicators, including age, type of employment, pre-tax income, and gender. Finally, it is widely recognised that membership in frequent flyer programmes can have a significant impact on airline choices, especially in

the case of higher tier status (see e.g., Chin, 2002; Proussaloglou and Koppelman, 1999; Hensher et al., 2001; Proussaloglou and Koppelman, 1995; Hess et al., 2007). 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). While our survey did not look at the differentiation in benefits as a result of fares, we collected information on membership in frequent flier programmes with a view to later use in the modelling analysis.

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. Mean survey completion time was 16 minutes, with a median value of 14 minutes.

Table 2: **Socio-demographics of respondents**

Age		**Work type**		**Personal pre-tax income**		**Gender**	
18 to 24	16.7%	Full time	67.7%	Under AUD10,000	3.5%	Male	45.2%
25 to 34	34.6%	Part time (< 30 hours/week)	16.9%	AUD10,000 - AUD19,999	3.7%	Female	54.8%
35 to 44	22.7%	Causal	6.3%	AUD20,000 - AUD29,999	5.6%		
45 to 54	13.4%	Does not work	7.4%	AUD30,000 - AUD39,999	8.4%		
55 and over	12.6%	Undisclosed	1.7%	AUD40,000 - AUD49,999	9.7%		
				AUD50,000 - AUD59,999	12.8%		
				AUD60,000 - AUD79,999	14.9%		
				AUD80,000 - AUD99,999	12.6%		
				AUD100,000 - AUD119,999	8.0%		
				AUD120,000 - AUD149,999	3.0%		
				Over AUD150,000	4.5%		
				Undisclosed	13.3%		

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.

The SP and OTA 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 they are obtained from two different types of data. 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). The equality of relative sensitivities is a testable assumption.

Scale differences between segments of a dataset (in this case the two types of survey task contained in a single pooled database) can be accommodated straightforwardly by estimating the scale of one of the segments relative to the other segment, an approach often based on a Nested Logit specification with dataset specific nests, as discussed by Bradley and Daly (1991) and later also by Hensher and Bradley (1993). A shortcoming of this approach is the inability to directly handle the repeated choice nature of the data at hand, thus almost surely leading to underestimation of the standard errors. The challenge is thus to specify a model that can account both for scale differences and the panel nature of the data.

An alternative approach is to use the scaled MNL (SMNL) model, which is a specific case of the generalized multinomial logit (GMNL) model. The GMNL model was first operationalised by Fiebig et al. (2010) and subsequently by Greene and Hensher (2010). Let $\mathrm{U}_{\mathrm{nsj}}$ denote the utility of alternative *j* perceived by respondent *n* in choice situation *s*. $\mathrm{U}_{\mathrm{nsj}}$ may be partitioned into two separate components, an observed component of utility, $\mathrm{V}_{\mathrm{nsj}}$ and a residual unobserved component, $\varepsilon_{\mathrm{nsj}}$, such that

$$\mathrm{U}_{\mathrm{nsj}} = \mathrm{V}_{\mathrm{nsj}} + \varepsilon_{\mathrm{nsj}}. \qquad (1)$$

The observed component of utility is typically assumed to be a linear relationship of observed attribute levels, *x*, of each alternative *j* and their corresponding weights (parameters), β, such that

$$\mathrm{U}_{\mathrm{nsj}} = \mu_{\mathrm{n}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \beta_{\mathrm{k}} \mathrm{x}_{\mathrm{nsjk}} + \varepsilon_{\mathrm{nsj}}, \qquad (2)$$

where β_{k} represents the marginal utility or parameter weight associated with attribute *k* for respondent *n* and the unobserved component, $\varepsilon_{\mathrm{nsj}}$, is assumed to be independently and identically (IID) extreme value type 1 (EV1) distributed.

As well as containing information on the levels of the attributes, *x* in Equation (2) may also contain up to *J*-1 alternative specific constants (ASCs) capturing the residual mean influences of the unobserved effects on choice associated with their respective alternatives; where *x* takes the value 1 for the alternative

under consideration or zero otherwise. It is clear from Equations (2) that both μ_n and β_k, cannot separately be estimated. As such, most discrete choice models make assumptions about μ_n such as $\mu_n = 1.0$, thus allowing β_k to be estimated.

The utility specification in Equation (2) is flexible in that it allows for a number of different functional forms. The SMNL model assumes that different respondents have the same marginal utilities for each attribute being modelled but different error variances and hence scales, as shown in Equation (3).

$$\beta_{nk} = \mu_n \beta_k, \tag{3}$$

where

$$\mu_n = e^{\left(\bar{\mu} + \sum_{q=1}^{Q} \delta_q w_q + \tau v_n\right)}. \tag{4}$$

$\bar{\mu}$ in Equation (4) denotes a mean parameter of scale, τ a parameter of unobserved scale heterogeneity and v_n a draw from a standard Normal distribution representing the unobserved scale heterogeneity. δ_q in Equation (4) represents parameters associated with covariates w_q which may be used to decompose the scale parameter.

In order for the model to be identified, it is necessary for some form of normalization to take place. This is done by normalizing $E\left[\mu_n^2\right] = 1$, which is accomplished by setting $\bar{\mu} = \frac{-\tau^2}{2}$ such that, ignoring $\sum_{q=1}^{Q} \delta_q w_q$, Equation (4) becomes

$$\mu_n = e^{\left(\frac{-\tau^2}{2} + \tau v_n\right)}. \tag{5}$$

For purposes of estimation necessity, both Fiebig et al. (2010) and Greene and Hensher (2010) truncate the draws for v_n to values between ±1.96, thus limiting potential software overflows which may occur in the exponantiation process.

Estimation of the model requires simulation of the log-likelihood over draws taken from z_n. The log-likelihood function of the model is

$$\log E(L) = \sum_{n=1}^{N} \log E\left(P_n^*\right). \tag{6a}$$

where

$$P_n^* = \prod_{s \in 1}^{S} \prod_{j \in 1}^{J} \left(P_{nsj}\right)^{y_{nsj}}. \tag{6b}$$

and where P_{nsj} are the choice probabilities calculated for the model and y_{nsj} equals one if alternative *j* is the chosen alternative in choice situation *s* shown to respondent *n*, and zero otherwise. Treatment of the log-likelihood function in this manner directly accounts for the panel nature of the SP data (see Revelt and Train 1998).

The empirical focus of this paper is on a possible difference in scale (and hence error variance). A deliberate decision was therefore made not to examine random preference heterogeneity in the current paper. The SMNL model was estimated for the combined dataset and both individual datasets, with 500 Halton draws (Halton 1960). To determine whether the datasets have different scales, a dummy variable w_{OTA} is used to decompose the scale parameter (see Equation 4). In the current context, w_{OTA} is set to one for observations from the OTA dataset and zero for the SP dataset, with an associated parameter δ_{OTA} estimated. This decomposition of the scale allows any differences in scale between the two datasets to be recovered.

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 δ_{OTA} is not statistically significant, then the unobserved effects do not systematically differ between the datasets. If δ_{OTA} is negative and statistically significant, then the OTA tasks have a lower scale and higher error variance than the SP tasks. Such a finding might suggest that people cannot handle the extra complexity of the OTA tasks. If δ_{OTA} is positive and statistically significant, then the OTA tasks have a higher scale and lower error variance than the SP 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.

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 provides information about the extent to which the default presentation was customised by the respondents.

5.1. S*ort 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 how many times each attribute was sorted on at the time of choice (column one), how many individuals sorted on each attribute for all of the tasks they completed (column two), and how many times an actual sort action was explicitly performed (column three). 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**

	1. Tasks with this sort at time of choice		2. Individuals with this sort at choice for all tasks		3. Sort actions performed	
Price	1,019	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	2,310	100%	462		1,380	100%

Examining column one of Table 3, price is clearly 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. Column three shows that there are more sorts on airline than any other non-price attribute, which suggests that some respondents may have strong airline specific preferences. At the individual level, column two 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.

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 AUD1,800 to AUD3,000 with an average of AUD2,482, 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	4.16%
Carbon tax	36	1.56%
Airline	76	3.29%
Total duration	49	2.12%
Flying duration	27	1.17%
Stopover duration	167	7.23%
Number of stops	187	8.10%
Seat pitch	198	8.57%
Seat reservation	258	11.17%
Entertainment system	476	20.61%
Ticket change charge	40	1.73%

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, 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. Additionally, at the individual level, 37 percent of respondents did not have any of the optional attributes shown for any of the five choice tasks. 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	1,034	44.76%
Flying duration	914	39.57%
Stopover duration	945	40.91%
Number of stops	1,023	44.29%
Seat pitch	744	32.21%
Seat reservation	824	35.67%
Entertainment system	951	41.17%
Ticket change charge	698	30.22%

5.4. *Model results for the individual datasets*

Separate SMNL 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 as only part of the information was stored in the database, and six OTA observations were removed as the chosen fares were several times higher than the average, where such outliers would have unduly affected model estimation. 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 hidden by the respondent 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. All of the parameter estimates in the OTA model are of higher statistical significance than their SP equivalent, including carbon tax, seat pitch, and seat allocation, which are 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. Unobserved scale heterogeneity is present in both datasets, as indicated by a very high significance for both τ parameters, with greater heterogeneity in the SP dataset. These two SMNL models represent a significant improvement on the corresponding MNL models, where the log likelihoods of the latter are reported.

A 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 were included in the model, as the remaining 13 dummies were found not to be 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 (e.g., respondents always choosing the cheapest option), and the absence of a treatment of this would have had an undue influence on model estimates (cf. Hess et al., 2010).

***Table 6*: Model Results**

	SP data		OTA data		Combined data		
	Parameter	**(*t*-ratio)**	**Parameter**	**(*t*-ratio)**	**Parameter**	**(*t*-ratio)**	**Dataset**
Price	-0.0034	(-7.79)	-0.0052	(-8.96)	-0.0027	(-14.66)	Both
Carbon tax	-0.0009	(-1.93)	-	-	-0.0006	(-2.19)	SP
Carbon tax	-	-	-0.0041	(-8.39)	-0.0031	(-9.64)	OTA
Charge for flight change	-	-	-0.0026	(-3.89)	-0.0020	(-4.54)	OTA
Travel time	-0.0011	(-3.33)	-0.0020	(-4.52)	-0.0011	(-6.75)	Both
Number of stops	-0.421	(-3.60)	-0.780	(-3.78)	-0.304	(-5.37)	Both
Seat pitch	0.069	(1.92)	-	-	0.050	(2.13)	SP
Seat pitch	-	-	0.598	(7.87)	0.328	(8.16)	OTA
Seat allocation	0.082[1]	(1.81)			0.046[10]	(1.38)	SP
Seat allocation			0.380[6]	(4.47)	0.197[11]	(4.52)	OTA
Entertainment (shared)	-0.308[2]	(-4.18)	-0.419[7]	(-4.80)	-0.215[12]	(-5.94)	Both
Airline constant 1	-0.245[3]	(-3.62)	-0.282[8]	(-5.78)	-0.137[13]	(-5.65)	Both
Airline constant 2	-0.385[3]	(-3.78)	-0.569[8]	(-6.56)	-0.285[13]	(-7.05)	Both
Airline constant 3	-	-	-1.017[8]	(-6.50)	-0.519[13]	(-6.24)	OTA
FF constant 1	0.934[4]	(4.13)	1.221[9]	(6.69)	0.646[14]	(7.16)	Both
FF constant 2	0.444[4]	(4.26)	0.399[9]	(5.59)	0.231[14]	(6.61)	Both
Arrive (9pm – midnight)	-0.203[5]	(-1.78)	-	-	-0.097[15]	(-2.2)	SP
Arrive (1am)	-1.205[5]	(-4.01)	-	-	-0.390[15]	(-3.87)	SP
1st alt. Shown	-	-	3.203	(11.17)	1.680	(7.94)	OTA
2nd alt. Shown	-	-	2.467	(9.52)	1.276	(7.24)	OTA
3rd alt. Shown	-	-	2.021	(7.98)	1.032	(6.55)	OTA
4th alt. Shown	-	-	1.804	(7.15)	0.917	(6.12)	OTA
5th alt. Shown	-	-	1.034	(3.89)	0.536	(3.74)	OTA
6th alt. Shown	-	-	1.083	(4.04)	0.567	(3.94)	OTA
7th alt. Shown	-	-	0.691	(2.29)	0.391	(2.51)	OTA
8th alt. Shown	-	-	0.886	(3.24)	0.488	(3.42)	OTA
SP alternative 1	0.151	(1.50)	-	-	0.114	(1.51)	SP
SP alternative 2	0.231	(2.27)	-	-	0.123	(1.62)	SP
τ	1.009	(7.75)	0.830	(12.74)	0.773	(14.84)	Both
Scale covariate – OTA dataset (δ_{OTA})	-	-	-	-	0.598	(6.62)	Both
Model fits							
LL(0)	-2,022.55		-5,693.70		-7,716.25		
LL(MNL)	-1,706.69		-3,278.66		-5,002.17		
LL(β)	-1,683.80		-3,181.83		-4,860.52		
Number of parameters	16		22		30		
ρ^2	0.167		0.441		0.590		
Adjusted ρ^2	0.160		0.434		0.585		
Observations	1841		1842		3683		
Respondents	462		462		462		

Base levels of effects codes: [1]No seat allocation (-0.082), [2]VOD/personal screen (0.308), [3]Airline constant 4(0.630), [4]No FF membership(-1.378), [5]Other times (1.407), [6]No seat allocation (-0.380) , [7]VOD/personal screen (0.419), [8]Airline constant 4(1.867), [9]No FF membership(-1.620), [10]No seat allocation (-0.046),[11]No seat allocation (-0.197), [12]VOD/personal screen (0.201), [13]Airline constant 4 (0.941), [14]No FF membership(-0.877), [15]Other times (0.487)

A number of qualitative attributes were effects coded in estimation, including seat allocation, entertainment, arrival time, and airline. Effects coding is similar to dummy coding however rather than setting the base attribute level to zero for each dummy variable, the base attribute level is given the value -1. Effects coding was chosen over dummy coding to prevent the base level from being confounded with the alternative specific constants[2]. Care must be taken when interpreting the effects coded parameters, as the base level[3] of utility of an attribute with *L* levels, β_L, is not zero, but rather $-\sum_{l=1}^{L-1}\beta_l$, i.e. the negative of the sum of estimates across all other levels, where β_l is the parameter estimate associated with the l^{th} effects coded variable. 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[4]. 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.

As previously mentioned, the survey contained questions to capture which, if any, relevant airlines the respondent had frequent flyer (FF) membership for, either through airline or alliance programmes. Effects coded interaction dummies were created for each FF programme and airline of interest. For every flight alternative that a respondent viewed, the corresponding FF programme and airline interaction dummy was set to one if that respondent was a member of the airline's FF programme, or a FF programme of an associated alliance. Thus, each FF programme interaction dummy parameter

[2] The estimated marginal utility for the base level of a dummy coded attribute will be confounded with the model ASCs. For example, assume a three level attribute is dummy coded such that the utility specification for that attribute is $\beta_0 + \beta_1 \times dummy_1 + \beta_2 \times dummy_2$. The marginal utility for the base level is therefore equal to β_0. Given the presence of multiple dummy coded variables, the base level of each variable will be similarly confounded with the model ASCs and hence no independent comparison of the marginal utilities of each of the base attribute levels is possible. When effects coding is applied, the marginal utility of the base level is given as $\beta_0 - \beta_1 - \beta_2$ which provides a unique estimate of the marginal utility of the base level of each attribute unconfounded with the model constants.

[3] We use the highest L as the base.

[4] 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 world data in the survey.

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 programme of the airline in question. That airlines represented in FF constants 1 and 2 differ in the effect of their FF programme is curious. Several interpretations are possible, if not identifiable. Airlines under FF constant 1 may have more rewarding FF programmes, 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 programme 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 SMNL model with the scale decomposed by a dataset dummy, as detailed in Section 4. Before settling on a final combined data model to analyse, it was necessary to determine whether the parameters could be treated homogeneously across the two data sets. The first step employed was to plot the SP SMNL parameter vector against the OTA SMNL parameter vector, as detailed in Louviere et al. (2000). The slope of the graph represents the ratio of error variances of the two datasets, with outlying points potentially representing parameters that cannot be considered homogeneous across datasets. The results are shown in Figure 3, where some parameter pairs are scaled by a constant to compress the plot. Carbon tax and seat pitch are possible outliers, although this might be accounted for by the lack of statistical significance at the 95 percent confidence level for these two parameters in the SP dataset. A close examination of seat allocation and charge for flight change is also warranted, as they also are not statistically significant at the 95 percent confidence level in the SP dataset. Various specifications of the combined model were tested, treating the above four attributes as either generic across the datasets of dataset specific.

This combined model is documented in Table 6. It may seem surprising that the fit of the joint model is superior to the combined fit of the two individual models. However, it is important to recognise that in the present context, the combination of the two separate models does not in fact nest within it the simple joint model. Indeed, even though the two separate models allow for dataset specific estimates for all parameters, while the joint model only does so for a subset of parameters, a further difference arises. All three models allow for scale heterogeneity across respondents. However, combining the two separate (OTA and SP) models means that we are working with an overall model that in effect treats the eight choices coming from a single respondent as two sets of four choices coming from two separate respondents. On the other hand, in the joint model, the integration is carried out at the level of the entire sequence of choices, meaning that this model recognises the fact that for each respondent, eight choices are observed, and that the random scale term is invariant across those eight tasks (notwithstanding additionally accommodated scale differences between the two groups of four choices). This treatment of the data as blocks of eight choices facilitates the recovery of scale heterogeneity across respondents, which in this case seems to lead to further gains in model performance.

T statistical significance of the scale covariate δ_{OTA} indicates that in addition to unobserved scale heterogeneity (which, like the individual models, is greater for the SP dataset), the dataset has an impact on scale, with the OTA choice tasks exhibiting *greater* scale (i.e., *lower* error variance) than the SP choice tasks. Carbon tax, seat pitch, seat allocation, and charge for flight change were all identified as being dataset specific. Section 6 contains a full discussion of the differences in error variance, as well as the extent of homogeneity of the parameters. The significance of the parameters in the combined model in most cases represents an improvement on the significance of the respective parameters in the dataset specific SMNL models. This is not surprising, as these parameters are estimated with more observations.

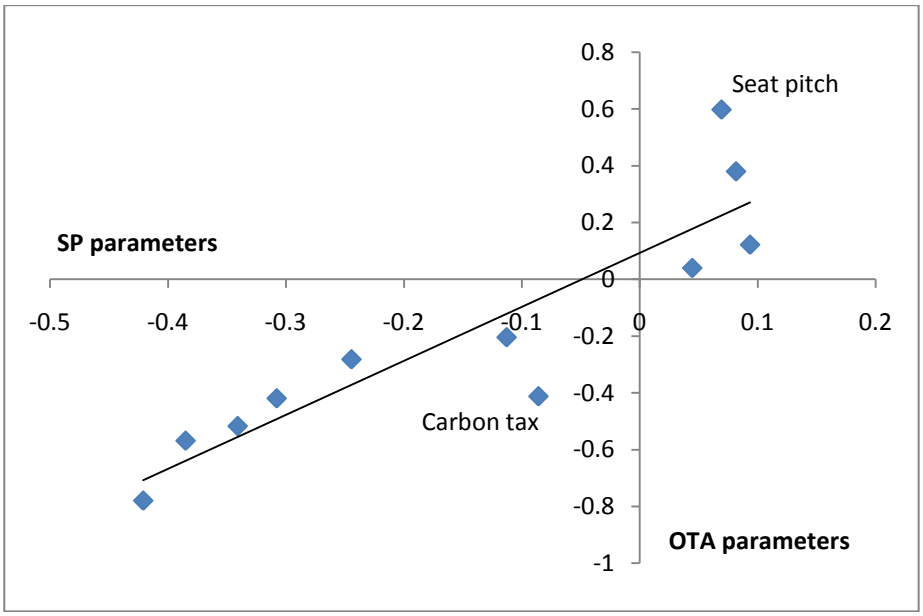


Figure 3: **Parameter plot of SP and OTA parameters**

5.6 *Monetary valuations*

The monetary valuation values derived from the combined model, together with the associated *t*-ratios and 95 percent confidence intervals (CIs) are shown in Table 7. The latter two measures were calculated using the Delta method, which, as discussed by Daly et al. (2010), is an exact method rather than an approximation. 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 from one level of an attribute to a better level 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. Given that the choice task was clearly framed as a return trip from Australia to London or Paris, with the return leg having the same service levels, some monetary valuations need to be halved. For example, respondents were prepared to pay $129.79 to avoid one stop for *each* of the two journeys, or $64.90 per stop *overall*, and $10.90 to avoid an hour of travel time. Other monetary valuations listed in Table 7 are more intuitive when considered as a valuation per hour. In the example of a 48 hour return flight, one inch of seat pitch is valued at $2.24 per hour (within a tested range of 31 to 34 inches), the ability to select a seat is valued at $3.08 per hour, and the presence of video on demand or personal screens instead of shared screens is valued at $3.35 per hour.

While these values may seem high, the flights presented are long distance economy class flights, not short haul hops where people are prepared to sit in uncomfortable seats, and where fare and schedule have an overriding influence on choice. Martin et al. (2008) present findings for legroom that are consistent with the monetary valuation for seat pitch obtained in the present study. They presented an SP task for a short haul flight from Madrid to the Canary Islands. The WTP for an increase in legroom, defined as enough space for the passengers' knees not to touch the seat in front, ranges from €14.23 (those under 45 years of age) to €20.21 (those over 45), for those who are paying their own ticket (as with the present study). This is in the context of a one way flight with a mean fare of €93.25, and flight time of about 3 hours, resulting in WTPs of approximately €4.74 and €6.74 per hour, for young and old respectively, which aligns well with our findings.

While the parameters underlying the effects coded valuations are statistically different, a lack of overlap in the CIs provides us with further confidence that the values themselves are statistically different. Due to the nature of the models applied, these monetary valuations are only point estimates. It is reasonable to assume that some preference heterogeneity exists within the sample. Recovering this heterogeneity remains an area for future work; the ability to recover both preference and scale heterogeneity is a strength of the GMNL model (of which the SMNL model used here is a special case). Further, systematic preference heterogeneity could be recovered through the use of interactions both with socio-demographic information (e.g., Hess et al., 2007), and search behaviour.

Table 7: Monetary valuation results from the combined model for a return flight from Sydney to London or Paris

Attribute	Willingness to pay (improvement)	Willingness to pay (avoiding change)	(*t*-ratio)	C.I. lower	C.I. upper
Travel time (hour)	-	$21.80	(4.91)	$13.10	$30.49
Number of stops	-	$129.79	(4.56)	$74.05	$185.53
Seat pitch OTA (inch)	$107.43	-	(5.43)	$68.68	$146.18
Seat allocation OTA (yes)	$73.94	-	(4.77)	$43.53	$104.35
Seat allocation OTA (no)	-	$73.94	(4.77)	$43.53	$104.35
Entertainment (shared)	-	$80.41	(6.56)	$56.38	$104.43
Entertainment (personal & VOD)	$80.41	-	(6.56)	$56.38	$104.43
Airline constant 1	-	$51.19	(6.15)	$34.89	$67.50
Airline constant 2	-	$106.86	(7.79)	$79.99	$133.73
Airline constant 3	-	$194.28	(6.80)	$138.30	$250.25
Airline constant 4	$352.33	-	(8.86)	$274.40	$430.25

FF constant 1	$241.79	-	(7.97)	$182.35	$301.24
FF constant 2	$86.69	-	(7.24)	$63.22	$110.15
FF base	-	$328.48	(9.16)	$258.22	$398.73

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 from this work may have implications not just for the study of air travel behaviour with SP surveys, but also other areas where extensive use is made of SP surveys to study travel choices that exhibit a high degree of complexity. The usefulness of the results for policy making would obviously need to be judged on a case by case basis.

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 with the 'tools' was clearly more reduced, 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. However, we decided against this, given that the majority of OTAs do sort by price by default (notwithstanding some exceptions).

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) for only 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. Additionally, the models showed significant variations in scale across individual respondents (on top of the between game variation), where this was lower for the OTA game.

We believe that our findings show that the survey methodology presented here 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. Additionally, the extent of 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 socio-demographic characteristics) 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.

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.

Daly, S., Hess, S., & de Jong, G. (2010), Calculating errors for measures derived from choice modelling estimates , ITS working paper, Institute for Transport Studies, University of Leeds.

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.

Fiebig, D.G., Keane, M., Louviere, J.J., and Wasi, N. (2010) The generalized multinomial logit: accounting for scale and coefficient heterogeneity, *Marketing Science*, 29(3), 393-421.

Greene, W.H. and Hensher, D.A. (2010) Does scale heterogeneity across individuals matter? an empirical assessment of alternative logit models. *Transportation*, 37(3), 413-428.

Halton, J. (1960) On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, *Numerische Mathematik*, 2, 84-90.

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. (2010). Hypothetical bias, choice experiments and willingness to pay, in press, *Transportation Research Part B*, 44(6), 435-752.

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. & Polak, J.W. (2006a), Airport, airline and access mode choice in the San Francisco Bay area, Papers in Regional Science, 85(4), pp. 543-567.

Hess, S. & Polak, J.W. (2006b), Exploring the potential for cross-nesting structures in airport choice analysis: a case-study of the Greater London area, Transportation Research Part E, 42, pp. 63-81.

Hess, S. (2008), Treatment of reference alternatives in SC surveys for air travel choice behaviour, Journal of Air Transport Management, 14(5), pp. 275-279.

Hess, S. and Hensher, D.A. (2010). Using conditioning on observed choices to retrieve individual-specific attribute processing strategies, *Transportation Research Part B*, 44(6), 781-790.

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. (2010). Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transportation Research Part D*, 15(7), pp. 405-417.

Johnson, E. J., Moe, W. W., Fader, P. S., Bellman, S. and Lohse, G. L. (2004). On the Depth and Dynamics of Online Search Behavior. *Management Science*, 50(3), 299-308.

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.

Martin, J., Roman, C. and Espino, R. (2008). Willingness to pay for airline service quality. *Transport Reviews*, 28(2), 199-217.

Moe, W. M. (2006). An Empirical Two-Stage Choice Model with Varying Decision Rules Applied to Internet Clickstream Data. *Journal of Marketing Research*, 43(4), 680-692.

Ortuzar, J. de D. and Simonetti, C. (2008). Modelling the demand for medium distance air travel with the mixed data estimation method. *Journal of Air Transport Management*, 14(6), 297-303.

Pels, E., Nijkamp, P. & Rietveld, P. (2001), 'Airport and airline choice in a multi airport region: an empirical analysis for the San Francisco bay area', Regional Studies 35(1), 1–9.

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 Bliemer, 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.

Theis, G., Adler, T., Clarke, J., Ben-Akiva, M. (2006). Risk Aversion to Short Connections in Airline Itinerary Choice. *Transportation Research Record*, 1951, 28-36.

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.