

# Finding Effective Ways to Improve Subjective Probability Predictions through Model Learning

---

by

Xin Wei (Edward Wei)

Submitted to the Marketing Discipline Group, UTS Business School

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy, Marketing

at the

UNIVERSITY OF TECHNOLOGY SYDNEY

March 2014

Permission is herewith granted to UTS to circulate and to have copied for  
non-commercial purposes, at its discretion, the above title upon the request of  
individuals and institutions

## **Certificate of Authorship / Originality**

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text. I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Author.....

Xin Wei (Edward Wei)

# Finding Effective Ways to Improve Subjective Probability Predictions through Model Learning

By

Xin Wei (Edward Wei)

Submitted to the Marketing Discipline Group, UTS Business School, on 12 March 2014, in  
partial fulfilment of the requirements for the degree of Doctor of Philosophy, Marketing

## **Abstract**

Predicting probabilities of marketing events and choices is a primary activity in marketing. Prediction can be from the outcome of a formal model or one's knowledge but the two sources often conflict. Although overwhelming evidence has demonstrated that models often outperform people in terms of accuracy, there is little doubt that decisions are made mostly by people based on their own knowledge. Past research suggests that models and intuition should work together for better outcomes but it is unclear how this may be accomplished other than by using "plain vanilla" style Decision Support Systems (DSSs).

This researcher adopted an alternative approach and believes that the key to solving this problem is to improve people's own knowledge. To do so, people need to gain a substantive understanding of a reliable model to improve predictions. Four generalisable model-learning approaches based on concepts from learning theories in psychology and cognitive science were developed and tested in an experiment to ascertain which approach was more effective in helping learners develop an understanding of the model's parameters and to improve their consequent predictions. The experiment was supported by an online Intelligent Support System (ITS) with both learning approaches and a target model built in. This target learning model is a consumer choice model of airline flights. The system evaluates predictions, estimates learner models, and classifies answers. Moreover, it provides real-time feedback matching the design of each learning approach.

According to the results, the most effective approach for both model learning and prediction improvement is a learning approach generating outcome feedback with correct answers after each experimental design controlled training task. This finding disagrees with a common view of multiple cue probability learning (MCPL). Having regard for effectiveness, the above learning approach is followed by an approach showing feedback with a comparison of estimated learner

model and target model outcomes on all parameters. Both approaches outperformed the approach where learners performed self-regulated learning in a DSS which is actually the *status quo* of decision support nowadays. Another approach tested was to learn a model for a consumer class from the similarities of classes. This approach achieved slow improvement but can be further refined.

In conclusion, this research opens a new path for prediction improvement by combining a learning approach, and methods and technology for experimental design, ITS and DSS.

Thesis Supervisor: Professor Jordan Louviere

Thesis Supervisor: Professor Mary-Anne Williams

Thesis Supervisor: Dr Tiago Ribeiro

## Acknowledgments

During this long journey, I received encouragement, friendship, support and help from so many people around me and this is an opportunity for me to thank them formally.

First and foremost I would like to thank my principal supervisor Professor Jordan Louviere. Jordan started this journey with me, guided me along the whole way, discussed with me a range of big ideas to minor details, and always encouraged me at the right time when he saw I was frustrated. Jordan, it is such an honour to be your PhD student to learn not only knowledge from you but also the ways in which you do things. I am sure the experience with you in the past seven years will greatly benefit me in my work and life.

I would like express my sincere gratitude to my two other supervisors, Professor Mary-Anne Williams and Dr Tiago Ribeiro. Mary-Anne, you have opened the door for me to fields such as cognitive science, machine learning and intelligent systems, and suggested new directions for this research. I will keep improving my understanding of these fields because I truly believe that understanding the process of how people learn is the key to future research in decision support. Tiago, even though you came onto my committee at late stage, your contribution to this research is so significant. You helped me understand the nature of the data I collected and showed me new and different analysis methods with which I was unfamiliar. Without your help, the analysis would not have been conducted so smoothly.

I thank Professor Joffre Swait and Dr Bart Frischknecht for offering several excellent suggestions regarding classification, design and analysis of the experiment. During the entire period of this research development, I also received advice and feedback from CenSoC colleagues and visitors.

In particular, I thank Dr Christian Schlereth, Dr Jorge Arana, Dr Terry Flynn, Dr Elisabeth Huynh and Dr Simon Fifer.

I thank my teachers and colleagues at the Marketing DG for their help during my research and their feedback to my semester presentations. In particular, thanks are due to Associate Professor Sandra Burke, Dr Paul Burke, Dr Christine Eckert, Dr Paul Wang, Professor Grahame Dawling, Dr Chelsea Wise, and Dr Ingo Bentrött. I also thank my former colleagues and PhD colleagues who also studied under Professor Jordan Louviere, Dr David Philens, Dr Con Menictas and Dr Luke Greenacre for their advices and help in the early days of this research.

When I first started this research, I benefited greatly from a conversation with Professor Ray Cooksey from the University of New England who was introduced to me by Jordan.

I wish in particular to thank my CenSoC colleague Gail Bradford who shares the office with me. Gail has been my first reader when I was working on the draft of my thesis. Gail, I truly appreciate all the time you spent in helping me to make my writing clearer for readers. I want to thank my colleague for the longest time at CenSoC, Maria Lambides, for your warm encouragement over all these years. I thank my colleague and former office mate Jane Pong for her advice and great editing work on my early drafts and proposals. I also thank my colleague Karen Cong for helping me with some programming and data recoding.

I thank my thesis editor, Dr Guenter Plum for his excellent editing work to my final thesis. I also thank Marketing DG and the Faculty of Business for providing funding for this research, and Pure Profile for providing the sample for my experiment.

Last, I dedicate this work to my family who supported me greatly during the whole time. To my wife Donna, without your help, tolerance and understanding, this work would not have been possible.

# Table of Contents

---

<b>Chapter 1 Introduction .....</b>	<b>1</b>
1.1 Background and Motivation .....	1
1.1.1 Overview.....	1
1.1.2 Bridging the Gap between Predictions by People and Models.....	2
1.2 Research Problem and Research Hypotheses .....	6
1.3 Overview of Experiment Plan.....	12
1.4 Justifications for Research.....	16
1.4.1 Learning Choice Models to Make Predictions on Choice Probabilities .....	16
1.4.2 Two measurements, Two Questions and One System Used in Prediction.....	20
1.4.3 Implications in Marketing.....	21
1.5 Linking Theories and Methodology.....	22
1.6 Plan of Thesis.....	23
<b>Chapter 2 Developing Learning Approaches .....</b>	<b>25</b>
2.1 Overview.....	25
2.2 A Review of Learning Theories.....	27
2.2.1 Definitions of Learning and Taxonomy of Learning Theories.....	27
2.2.2 The “Behavioural” School of Learning Theories.....	28
2.2.3 The “Mind” and “Machine” Schools of Learning Theories .....	31
2.3 Key Attributes to Characterise Learning.....	36
2.3.1 The Selection of Key Attributes .....	36
2.3.2 Attribute One – Self-Regulated and Design Controlled Learning.....	40
2.3.3 Attribute Two – Feedback .....	45



2.3.4 Attribute Three – Knowledge Representation .....	49
2.3.5 Attribute Four – Categorisation.....	52
2.3.6 Summary.....	56
2.4 Building Learning Approaches under Extended Framework .....	57
2.4.1 Extended Framework for Learning.....	57
2.4.2 Designing Learning Approaches for Testing.....	59
<b>Chapter 3 Evaluation, Estimation and Classification .....</b>	<b>63</b>
3.1 Overview.....	63
3.2 Evaluating Probability Predictions.....	65
3.2.1 What is Good Probability Prediction? .....	65
3.2.2 Using Scoring Rules to Evaluate Accuracy of Probability Predictions .....	69
3.3 Preparing Target Learning Models and Estimating Learner Model .....	74
3.3.1 People’s Knowledge in Probability Predictions and Discrete Choice Models.....	74
3.3.2 General Framework to Model Choices.....	75
3.3.3 Model Type One - Aggregated Choice Model.....	78
3.3.4 Model Type Two – Choice Models with Latent Classes .....	80
3.3.5 Estimating Learner Model for Real-Time Feedback.....	87
3.4 Classification .....	90
3.4.1 Selecting the Appropriate Classification Method.....	91
3.4.2 Bayesian Classifier and Choosing Appropriate Likelihood Function .....	94
3.5 Summary .....	96
<b>Chapter 4 Design an Empirical Study .....</b>	<b>97</b>
4.1 Overview.....	97
4.2 Overview of Stated Preference and Experimental Design Methods .....	98
4.3 Designing Stage 1 Consumer Survey .....	104

4.3.1	<i>Choosing the Product Category .....</i>	104
4.3.2	<i>Selecting Attributes.....</i>	105
4.3.3	<i>Attributes, Experimental Design and Choice Tasks for Stage 1 Survey.....</i>	107
4.4	Summary of Stage 2 Training and Learning Experiment .....	110
4.4.1	<i>Overview.....</i>	110
4.4.2	<i>Experimental Design for Learner Experiment .....</i>	112
4.4.3	<i>Four Experimental Conditions Matching Four Learning Approaches .....</i>	114
4.4.3.1	Experimental Condition 1 (EC1) .....	115
4.4.3.2	Experimental Condition 2 (EC2) .....	116
4.4.3.3	Experimental Condition 3 (EC3) .....	117
4.4.3.4	Experimental Condition 4 (EC4) .....	119
4.5	Summary .....	122
<b>Chapter 5</b>	<b>Methodologies and Results of Analysis .....</b>	<b>124</b>
5.1	Introduction .....	124
5.2	Analysis Methods and Results for Stage 1 Consumer Survey.....	125
5.2.1	<i>Analysis Methods .....</i>	125
5.2.2	<i>Target Model One – Fixed-Effects MNL Model for All Consumers’ Choices.....</i>	126
5.2.2	<i>Target Model Two – Fixed-Effects MNL Models of Consumer Classes.....</i>	131
5.3	Analysis Methods and Results for Stage 2 Learning Experiment .....	133
5.3.1	<i>Response Rate, Duration and Learners’ Feedback.....</i>	133
5.3.2	<i>Testing Prediction Accuracy.....</i>	136
5.3.3	<i>Results - Prediction Accuracy .....</i>	142
5.3.3.1	Prediction Accuracy for “Preferred” Option .....	142
5.3.3.2	Prediction Accuracy on Target Consumer Group (Learning Approach Four) ...	145
5.3.3.3	Prediction Accuracy on Probabilities of Options .....	147

5.3.4 Summary of Hypotheses H1a, H2a and H3a.....	155
5.3.5 Testing Target Model Parameter Learning.....	157
5.3.5.1 The Nature of Testing Model Parameter Learning.....	157
5.3.5.2 Models and Analysis Methods in Testing Model Parameter Learning.....	161
5.3.6 Results – Target Model Parameter Learning.....	166
5.3.6.1 Initial Analysis - Individual Level Coefficients $\widehat{\Delta\beta^t}$ .....	166
5.3.6.2 Follow-up Analysis - Using $\widehat{\Delta\beta^t}$ to Test Model Parameter Learning.....	170
5.3.7 Summary of Hypotheses H1b, H2b and H3b.....	183
5.4 Summary .....	185
<b>Chapter 6 Conclusions and Implications .....</b>	<b>187</b>
6.1 Introduction .....	187
6.2 Conclusions of Research Problem and the Four Learning Approaches.....	187
6.2.1 Conclusion regarding the Research Problem.....	187
6.2.2 Conclusion on Learning Approach One.....	189
6.2.3 Conclusion on Learning Approach Two .....	191
6.2.4 Conclusions on Learning Approach Three .....	193
6.2.5 Conclusions on Learning Approach Four.....	195
6.3 Contributions and Implications.....	198
6.4 Limitations and Future Research .....	201
<b>Appendix 1 Experimental Design for Stage 1 Survey .....</b>	<b>203</b>
<b>Appendix 2 Experimental Design for Stage 2 Tasks .....</b>	<b>205</b>
<b>Appendix 3 Stage 1 Consumer Survey Screenshots.....</b>	<b>207</b>
<b>Appendix 4 Stage 2 Learning Experiment Screenshots .....</b>	<b>221</b>
<b>Appendix 5 Socio-Demographic Background of Respondents .....</b>	<b>270</b>

<b>Bibliography .....</b>	<b>274</b>
---------------------------	------------

---

# List of Figures

---

Figure 1.1 A screenshot of the “plain vanilla” MDSS used in this research .....	9
Figure 1.2 An example training task in the Stage 2 training program .....	14
Figure 1.3 The position of this research in marketing decision making.....	23
Figure 2.1 Extended S-P-R-O Learning Framework.....	58
Figure 2.2 Four Learning Approaches.....	60
Figure 3.1 All data points fall into a convex hull defined by three archetypes (This example is from Eugster & Leisch, 2009, p. 18) .....	84
Figure 3.2 Three Classifiers (based on Figures 3 to 5 in Lippmann 1994, pp. 87–88) .....	92
Figure 4.1 Example choice and prediction task in Stage 1 online consumer survey .....	109
Figure 4.2 EC2 - an example of outcome feedback .....	117
Figure 4.3 An example feedback comparing “fare” in a learner model and the target model .....	119
Figure 4.4 EC3 - relative importance of attributes .....	119
Figure 4.5 EC4 - an example feedback after each task.....	120
Figure 4.6 EC4 - an example feedback after each Session.....	121
Figure 5.1 Proportions of respondents who predicted correctly in S1 and S2 by set.....	144
Figure 5.2 Transformations of Hellinger Distances (HDs) .....	148
Figure 5.3 Checking square root and original HD against normal distribution .....	148
Figure 5.4 Prediction accuracy improvements from starting session to end session by approach.....	153
Figure 5.5 Model learning process – converging to a fixed parameter.....	159
Figure 5.6 Distributions of individual coefficients of attributes for starting session.....	167
Figure 5.7 Distributions of individual coefficients of attributes for end session .....	168

---

Figure 5.8 Difference between end and starting coefficients, over starting coefficient (Qantas) .....	171
Figure 5.9 Difference between end and starting coefficients, over starting coefficient (Virgin).....	172
Figure 5.10 Difference between end and starting coefficients, over starting coefficient (Jetstar).....	173
Figure 5.11 Difference between end and starting coefficients, over starting coefficient (\$400 fare) .....	174
Figure 5.12 Difference between end and starting coefficients, over starting coefficient (\$460 fare) .....	175
Figure 5.13 Difference between end and starting coefficients, over starting coefficient (\$520 fare) .....	176
Figure 5.14 Difference between end and starting coefficients, over starting coefficient (4 hours) .....	178
Figure 5.15 Difference between end and starting coefficients, over starting coefficient (change allowed).....	179
Figure 5.16 Difference between end and starting coefficients, over starting coefficient (free food) .....	180

---

# List of Tables

---

Table 3.1 Examples of three “scoring rules” when actual probabilities are 1 or 0 .....	73
Table 3.2 Examples of three “scoring rules” when actual probabilities are not 1 or 0 .....	73
Table 4.1 Model results of the study conducted by CenSoC in 2010 .....	106
Table 4.2 Four Experimental Conditions.....	114
Table 5.1 Conditional logit model for cross-country flight offer choices .....	127
Table 5.2 Mother Logit model including all cross-effects (CEs) .....	128
Table 5.3 Model predictions and learner predictions vs. actual choice probabilities.....	130
Table 5.4 Results of archetypal analysis.....	132
Table 5.5 Results of three fixed-effects models for the three classes by archetypal analysis .....	132
Table 5.6 Duration for Stage 2 Learning Experiment.....	134
Table 5.7 Respondents’ feelings about the four learning approaches after Session 2.....	135
Table 5.8 Respondents’ willingness to participate in future training after Session 2.....	135
Table 5.9 Correct predictions of preferred options for total predictions.....	143
Table 5.10 Respondents’ performance in predicting preferred options .....	143
Table 5.11 Learners’ performance in correctly predicting target consumer group .....	145
Table 5.12 Prediction success achieved by learners (% of total responses) .....	145
Table 5.13 Proportion of correct predictions (groups belonged to by groups to predict) .....	146
Table 5.14 Means of transformed HD by learning approaches & sessions .....	149
Table 5.15 Results of regression analysis with transformed HD as dependent variable.....	150
Table 5.16 Comparing prediction accuracy by learning approaches & sessions .....	151
Table 5.17 A summary of model fits for regression analysis using mean HDs.....	152
Table 5.18 Proportions of learners who improved their prediction accuracy .....	154

---

Table 5.19 Average transformed HDs by target segments for Approach Four .....	155
Table 5.20 Means and standard deviations of individual coefficients for starting and end sessions .....	169
Table 5.21 “Qantas” parameter learning by learning approach .....	171
Table 5.22 “Virgin Australia” parameter learning by learning approach .....	173
Table 5.23 “Jetstar” parameter learning by learning approach .....	174
Table 5.24 “\$400 fare” parameter learning by learning approach .....	175
Table 5.25 “\$460 fare” parameter learning by learning approaches .....	176
Table 5.26 “\$520 fare” parameter learning by learning approach .....	177
Table 5.27 “Flying time 4 hours” parameter learning by learning approach .....	178
Table 5.28 “Ticket change allowed” parameter learning by learning approaches .....	179
Table 5.29 “free food & beverages” parameter learning by learning approach .....	181
Table 5.30 Summary (mean) rankings of learning approaches .....	181
Table 5.31 Average learning of model parameters by approach & session .....	182
Table 5.32 Average learning of model parameters by session & target segment (Approach Four) .....	183

---



# Chapter 1 Introduction

## 1.1 Background and Motivation

### 1.1.1 Overview

Predicting probabilities of events occurring is challenging and important. As best put by Winkler (1996, p.1), “*since probability is the mathematical language of uncertainty, it is natural in modelling inferential and decision-making problems to represent our uncertainty in terms of probabilities*”. It is not hard to see that asking people to predict the probability of an event occurring is to assess their degree of belief in, or uncertainty towards, these events in quantitative form. For example, economists forecast likely movements in the economy and make recommendations based on those forecasts to government and organisations to determine policies. Doctors assess the risks and predict likely consequences for patients arising from the application of certain treatments. Examples of where probability predictions are made explicitly or implicitly can be found in all aspects of life. In marketing, practitioners constantly face situations of decision making in uncertain environments. Consumers also make decisions in purchasing all manner of goods and services from small items to large investments such as a house. Learning to effectively and accurately predict probabilities of certain occurrences may reduce the risk of making poor, or even wrong, decisions.

To make predictions, one either has to apply an inductive approach by closely observing the frequency of occurrence of events so as to establish the probabilities of their recurrence, or else has to understand the underlying model determining outcomes to make effective predictions based on expected probabilities deductively. This researcher is particularly interested in approaches that best support the second type of predictions, that is, learning a model. Although the first approach may be an approach of choice in predicting simple events, in predicting events influenced by, or associated with, other more complex phenomena, it is difficult if not impossible to take an

inductive approach because observed frequencies are often not available for the making of inferences. Studying how people predict and learn probabilistically is hardly a new research topic. What is new in this research is the identification of a direction that can most effectively facilitate learning by people from a complex discrete choice model so as to make probability predictions on a discrete number of choice options when these options vary. This research applies a new type of experiment which treats model learning, task training, feedback and other elements all through an intelligent tutoring system free from intervention by human trainers.

### **1.1.2 Bridging the Gap between Predictions by People and Models**

Considering how people learn to predict probabilities and how to evaluate their predictions, abundant research has been conducted on this and related topics in psychology, judgment and decision making (JDM), and social judgment theory (SJT) (e.g. Bower & Hilgard 1981; Camerer & Johnson 1997; Cooksey 1996). For example, Estes and other psychologists developed learning theories articulating associations between recurring situations and subsequent events in probability learning (Estes 1950; Estes & Burke 1953; Estes 1972). Decision theory researchers and statisticians discussed and tested many “scoring rules” on how to measure people’s probability predictions (Friedman 1983; Gneiting & Raftery 2007; Nau 1985; Winkler 1996; Winkler & Murphy 1968). Hammond and others studied probability learning based on single and multiple cues (Cooksey 1996; Hammond & Stewart 1975). In related fields to probability predictions, Meehl and others compared predictions made by experts and models in different problem areas and concluded that models overwhelmingly outperformed people on accuracy (e.g. Dawes 1971; Goldberg 1970; Grove et al. 2000; Meehl 1954). Tversky and Kahneman (1974) initiated the trend to study heuristics and bias that influences people’s judgment process. These theories on probability learning and related fields provide a solid foundation to guide empirical studies in various fields including marketing. For example, Meyer’s study on learning multi-attribute

judgment policies is theoretically based on a multiple-cue probability learning (MCPL) approach under SJT (Meyer 1987).

As distinct to studying people's probability prediction activities, advancements in computers and other technologies bring with it an enormous capacity and capability for research to discern probabilistic data from the resources available. This area of research covers model estimation, predictive learning in computer science and artificial intelligence, and data mining (Cherkassky & Mulier 2007). In marketing, learning from data has become increasingly popular, performing well in areas such as forecasting, product and price optimisation, and consumer studies (e.g. Dzyabura & Hauser 2011; Lilien & Rangaswamy 2002). Availability of large databases such as scanner panel data and survey data also make learning and predicting possible. Apart from these examples, in various operations certain problems may be supported by imperfect information, while certain problems may require a large budget to process, and certain problems can only be solved on a sequential, step by step basis. In solving these classes of problems, learning from data plays an important role (Powell & Ryzhov 2012).

Looking at the foregoing streams of research, it is not difficult to see that there exists a gap between probability predictions made by people using knowledge they possess and those predictions made by models gained from data computation and analysis. For example, the probability of consumers choosing a product at a certain price level can be intuitively predicted by marketers. Such choice can also be predicted from sales and research data. In reality, the two predictions can be quite different. In this case, which prediction should a company rely on? This problem is not simply a matter of trust, but relates to reasoning and sources of disagreement behind the two approaches. For example, people do not use every piece of information as do models and apply shortcuts or "configural rules" in the predictions (Camerer & Johnson 1997). On the other hand, models can

outperform humans in prediction accuracy because models can represent relationships of variables more efficiently, even if relationships are not represented accurately (e.g. Grove et al. 2000; Hastie & Dawes 2001). One of the common thoughts among researchers seems to be: even though people are not as accurate as models for various reasons, they can benefit from models as decision aids to improve their own predictions. This belief that *people can improve* their predictions via learning from models is the fundamental premise of this research.

In marketing, this gap remains problematic in both theory and practice waiting for a better solution (Hoch & Kunreuther 2001; van Bruggen & Wierenga 2010). Some researchers have made their positions on this quite clear. For example, Blattberg and Hoch (1990) suggested that the ideal strategy is to apply both people's intuition and the model in predictions and overall decision making. Dhir (2001) stressed the importance of managers to have a "convergent understanding" of models by focusing on the reasoning behind these models. The reality is, even though models may generate more accurate predictions than people, decisions are still largely made by decision makers themselves based on their own experiences. This is widely acknowledged from studies performed on managers at various levels (e.g. Burke & Miller 1999; Covin, Slevin & Heeley 2001; Vanharanta & Easton 2010; Woiceshyn 2009). Studying how to best improve decision-makers' judgments by learning from models may be as important as studying the models themselves. In principle, since decisions are made by decision makers, the quality of decisions is determined by the decision makers' own mental models. Therefore, the contribution made from an external model to the quality of a decision is equivalent to the degree of influence by the model on the decision maker.

Few researchers in marketing so far have started to think carefully about training people in the use of models. An example is the study by Kayande et al. (2009) where training features were built into

decision support systems to teach users. However, research on decision support systems often limits itself to particular problems or features, hence its findings may not be widely applicable or suited to generalisation. Certainly practical features can improve the effectiveness of a system, although it is more important to identify theoretical directions which may lead to future developments to broaden the role that a decision support system can play. For example, some directions can be deduced from learning theories, progress made in model estimation, statistical learning and other areas. Meyer's study on multi-attribute learning suggests a useful direction in applying learning theories in multi-attribute probability predictions (Meyer 1987). Dzyabura and Hauser's (2011) study on learning consumer decision rules is supported by algorithms developed from machine learning.

In light of the foregoing, this research aims to narrow the gap between probability predictions made by people and by models. This is done by identifying the most effective model learning approach which can maximise learners' understanding of target models and increase the accuracy of their predictions. In building these approaches for testing, learning approaches, prediction evaluation, model estimation and classification are applied through an intelligent tutoring system. This research is not about particular design features such as graphic user interfaces or the visualisation of certain information. The learning approaches tested are driven by underlying theories. The objective is to identify effective ways for people to learn to operate a model and make better predictions. This researcher believes, until a direction is identified to effectively train people on models, a gap between the two sources of predictions based on probabilities will remain, and judgments and decisions by people and models will be inevitably disparate, if not conflict.

## 1.2 Research Problem and Research Hypotheses

Section 1.1 explained the background of the present research and the researcher's motivation in conducting it. This section provides a more detailed discussion of the research problem and hypotheses. At the outset, the research problem is stated as follows:

What is the most effective approach for training people to gain a substantive understanding of a model in making accurate predictions of related probabilities?

The study tests four model-learning approaches. Testing these approaches is effected in an online experiment using an intelligent tutoring system. The system has real-time evaluation, model estimation, classification and feedback generation capacities without the need for human intervention. A learner can work directly with the system to perform model learning and probability prediction in training tasks. The learning environment includes instructions and feedback processes that are designed specifically for each training approach. Learners are asked to study a particular model and learn how to make probability predictions according to the model's rules of operation. The learners interact with the system by receiving information, instructions and trial tasks. They then make a series of probability predictions dependent on the state of their learning of the target model. The system performs evaluations, estimations or classifications depending upon the particular experimental conditions imposed. The object is to identify the approach that causes learners to successively improve both prediction accuracy and model learning through the course of the experiment. By analysing data gained from the experiment, the most effective approach can be identified.

The type of model selected to empirically test this research problem is the discrete choice model (DCM) applied to consumer choices in a selected product category (cross-country airline travel). The reasons for selecting DCM for this study are threefold. First, DCMs are well developed both

theoretically and practically. They are an important tool in marketing science where new research proliferates. Second, the consumer choice problem is a complex phenomenon and there are various possibilities related to how consumers make choices. Nonetheless, consumer choices can also be simplified by DCMs into a probability prediction problem which matches this research topic. Third, although the current model-based decision support systems can make predictions, they are not designed to train people in how models work. Using DCMs for this research is an appropriate choice as these models reflect the complex nature of consumer behaviour and yield results in probability prediction terms. On the other hand, due to complexity, the DCM lacks an effective approach for teaching people how these models make probability predictions.

The four learning approaches tested in this research are as follows:

- Approach One: learners (selected subjects) are asked to learn directly how a DCM operates from an interactive decision support system which provides probabilities in response to any scenarios selected by the learners;
- Approach Two: learners are asked to learn from outcome feedback which gives correct answers to probability predictions made by learners during the training tasks;
- Approach Three: learners are asked to learn from feedback comparing their own models with the target model, attribute by attribute;
- Approach Four: learners are asked to learn from classification feedback indicating whether they have predicted probabilities according to a model of a particular class that they are asked to predict, accompanied by information on the differences between several classes.

Among these four approaches, learners involved with the first three approaches, learn from the same model, and learners in the fourth approach receive information related to several classes, but only one class is the correct one that they should learn.

Learners under Approach One are considered the control group. Under this approach, learners can choose any scenarios relating to the target model with any attribute combinations. A decision support system can immediately give correct probabilities of events or options in the chosen scenarios. Learners can perform self-regulated learning by going through any scenarios in which they are interested.

Feedback in Approach Two is widely applied in traditional probability learning studies (this will be discussed in Chapter 2). This type of feedback is called “outcome feedback”, which refers to the correct answer to a training task being given immediately following a learner’s completion of the prediction task. Learners are expected to correct what they believe may be causing discrepancies to occur iteratively through a series of tasks followed by feedback (Cooksey 1996).

Approaches Three and Four can both be considered as “cognitive feedback”, a concept established in MCPL studies. Cognitive feedback is a type of feedback that may contain, but not exclusively, the following types of information: statistical information about task characteristics, information about learners’ cognitive and judgment characteristics, and information that compares learners’ outcomes and system’s outcomes (Cooksey 1996). In brief, cognitive feedback works at a deeper level in a learners’ cognition. Approach Three supports attribute by attribute learning of the target model. Approach Four focuses on similarities and dissimilarities of several different classes (in the experiment, different consumer groups). The approach intends to establish a new and different type of cognitive feedback by highlighting their similarities/dissimilarities visually and structurally without explaining attributes one by one. This approach can be linked to such theories as Prototype, Imagery and Conceptual Spaces influential in cognitive science (e.g. Gärdenfors 2000; Kosslyn 1981; Rosch 1973).



Approach One is the default approach practised in marketing when a decision support is given to marketers. Marketers receive a marketing decision support system (MDSS) with a black-box style model built in. This system has computation features but no communication features (Kayande et al. 2009). Eisenstein and Lodish (2002) termed it the “plain vanilla” system, describing it as:

(plain) vanilla MDSS play a passive role in the human-machine interaction. They may execute computations, present data, and respond to queries. But they cannot explain their logic, deal with incomplete information, or make logical inferences ... they are incapable of even simple reasoning. Hence they do not serve as intelligent assistants to a decision-maker. (p. 439)

This approach is considered a reference point so the any different learning approaches can be compared against this approach to see if any improvements are achieved. This is equivalent to comparing a new hypothetical decision support approach to current practice. A screenshot of the “plain vanilla” MDSS used in this research is shown in Figure 1.1.

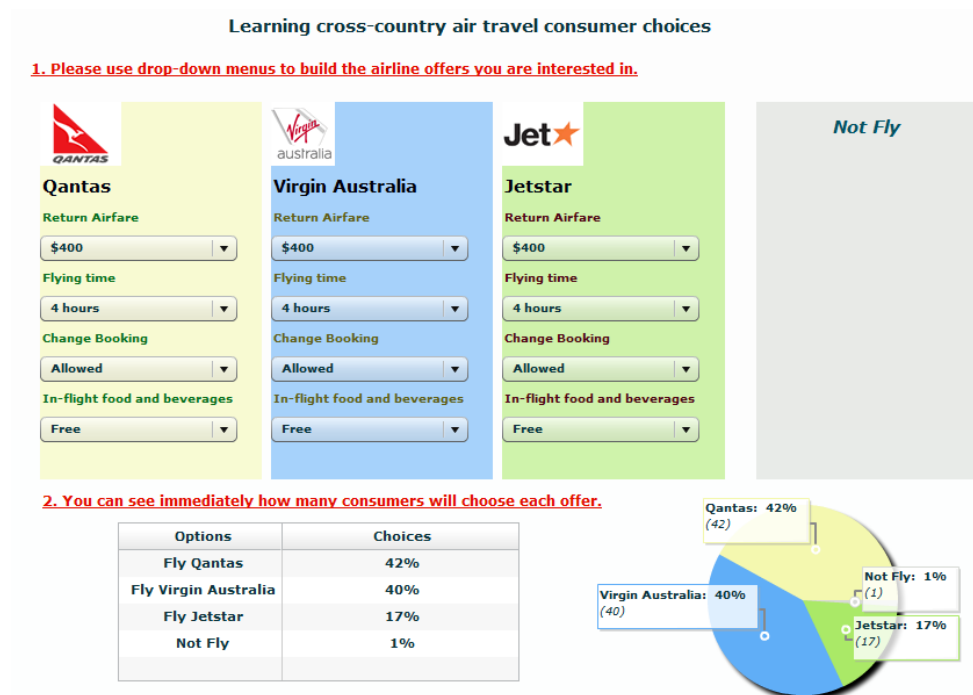


Figure 1.1 A screenshot of the “plain vanilla” MDSS used in this research

In this study, Approaches Two and Three are compared with Approach One. Different from most studies on probability learning in which only prediction accuracy matters, this research also examines how well learners can capture model parameters and apply their understanding in predicting probabilities. Prediction accuracy is regarded as the “normative” standard corresponding to subjective probability prediction. Understanding the model can also be referred to as “substantive learning” which implies probability predictions are made with learners having knowledge of the problem (Hogarth 1975; Winkler 1996; Winkler & Murphy 1968). Therefore, each main hypothesis is split into two separate hypotheses matching these two performance measurements.

The reason that Approach Four is not directly compared with Approach One is that target information and the learning objectives are different. While Approaches One to Three aim to understand a single target model to make probability predictions, Approach Four asks learners to learn from the differences of several different class models in order to converge to one particular class. Because the learning objectives and type of information are different, it is not reasonable to directly compare the performance of learners under Approach Four with Approach One, at least not in the same experiment. It is natural to believe that the task that learners are asked to perform is more difficult than the one required in the other three approaches. Instead of comparing them, a more appropriate hypothesis may be to focus on whether learners have greater success in their prediction and converge to the most appropriate class model with more tasks; for example, comparing a second training session with the first training session.

Hypotheses for this research are summarised below:

**H1a (on prediction accuracy with Approach Two versus Approach One):** Learners who receive outcome feedback after each training task make more accurate probability predictions than those who perform self-regulated learning using a "plain vanilla" MDSS.

**H1b (on model learning with Approach Two versus Approach One):** Learners who receive outcome feedback after each training task have a better understanding of the target model than those who perform self-regulated learning using a "plain vanilla" MDSS.

**H2a (on prediction accuracy with Approach Three versus Approach One):** Learners who receive diagnosis of their own model after training tasks make more accurate predictions than those who perform self-regulated learning using a "plain vanilla" MDSS.

**H2b (on model learning with Approach Three versus Approach One):** Learners who receive diagnosis of their own model after training tasks gain a better understanding of the target model than those who perform self-regulated learning using a "plain vanilla" MDSS.

**H3a (on prediction accuracy with Approach Four):** Learners who receive class information and classification feedback after training tasks improve their predictions of probabilities matching a particular class with more tasks and feedback given in training.

**H3b (on model parameter learning with Approach Four):** Learners who receive class information and classification feedback after training tasks gain a better understanding of a target class' model with more tasks and feedback given in training.

In testing these hypotheses, it is important to notice that different estimation approaches are applied using different measurements for prediction accuracy and model learning. Measurements for the two purposes come from different research areas.

To evaluate probability prediction accuracy, a widely applied and accepted measure in subjective probability evaluation is “scoring rules”. Scoring rules are well suited for evaluating the differences between two probability distributions with categorical choices. Simply speaking, for each prediction, a numerical score can be given to indicate the accuracy of a prediction. If such a scoring rule is “strictly proper” then better predictions will always yield better scores. Among many scoring rules, logarithm rules such as Shannon entropy and cross entropy (Kullback-Leibler divergence) are often applied. Other scoring rules such as the Brier score are also often applied. Gneiting and Raftery (2007) provided a complete review with mathematical proofs covering all commonly used scoring rules in probability prediction evaluation.

To test model learning, some measurements need be established to model the predicted probabilities of discrete choices or events. Methods to test model parameter learning are not as universal as scoring rules. Measurements and estimation approaches can be developed from discrete choice model literature. For example, given a learner’s prediction of choice probabilities for several options, if we can consider one option as the base option, then ratios of probabilities of other options over the probability of this base option yields a set of odds which are comparable across different scenarios. Odds and odds ratios are commonly applied concepts in discrete choice models and categorical data analysis (e.g. Agresti 2002).

### **1.3 Overview of Experiment Plan**

The experiment for this research is conducted in two stages. Stage 1 is a survey to study people’s choices with respect to a product, in this case, cross-country airline travel. Data for this survey is then used to develop real consumer choice models as target models for learning in Stage 2. Stage 2 is an online training experiment. The same respondents who completed the Stage 1 survey are invited to participate in this online training experiment. Learners are randomly assigned to one of the four conditions matching the four learning approaches discussed in Section 1.2. They complete




two sessions of training tasks, each with 16 prediction tasks under different scenarios. Data collected in Stage 2 is used for testing hypotheses H1a to H3b.

The Stage 1 consumer survey adopts a proven experimental design method (orthogonal main effects design or OMEP), and a randomly selected sample of consumers are invited to complete the survey. Two types of models are developed from this survey. The first type is an aggregated multinomial logit model (MNL) summarising the choices of the entire sample in Stage 1; the second type includes several class models in MNL, each representing unique choices of a particular class of sample or consumer group. A consumer group refers to a representative consumer group with more homogenous preferences and choices.

Using these two types of models, an online tutoring system is built to conduct the Stage 2 experiment. Those who completed the Stage 1 study are invited to complete the training program. Each respondent (learner) is randomly assigned to one of the four experimental conditions matching four learning approaches. In the program, there are two sessions of 16 training tasks identical for all approaches and learners. Each training task is similar to the one shown in Figure 1.2. In each task, respondents need to make a categorical choice (predict consumers' most likely choice) and a full probability prediction of consumers choosing each option including the "none" option.

As discussed, there are four experimental conditions (matching four learning approaches) for testing. Learners in Conditions 1 to 3 are asked to make predictions about **all** consumers by learning from the aggregated consumer model. Learners in Condition 4 are asked to make predictions about **one of the three** consumer groups by learning from the similarities and dissimilarities of the three consumer groups.

**Task 1 (Tasks 1 to 16 in Session One):**

	<div>Qantas</div> 	<div>Virgin Australia</div> 	<div>Jetstar</div> 
Return Airfare	\$580	\$580	\$400
Flying Time	4 hours	4 hours	4 hours
Change Booking	Not allowed	Not allowed	Allowed
In-Flight Food and Beverages	Free	Free	Not Free

Q1. If the above offerings are made in the market, which of the following choices do you think that consumers would **MOST LIKELY** make?

- ☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Q2. Out of **100** consumers, how many of them do you think would choose each option?

*Please enter a number in every box and make sure the sum is 100.*

people would fly Qantas  
 people would fly Virgin Australia  
 people would fly Jetstar  
 people would not fly

= 0

Figure 1.2 An example training task in the Stage 2 training program

Learners in Condition 1 are asked to learn from an MDSS directly through self-selected “what-if” analysis. Learners in Condition 2 receive outcome feedback after each task, informing them the correct answer to the task. Learners in Condition 3 receive a full model diagnosis after each session of 16 tasks comparing their own models with the target model. Each learner’s learning model is processed in real-time using prediction data from the previous session. Learners in Condition 4 are first asked to learn from information about the three consumer groups before starting their training tasks. Information they receive contain visual and verbal descriptions of three consumer groups describing their similarities and dissimilarities. They then receive feedback after completing each training task informing them whether they have predicted the correct group. Feedback is generated from a built-in classifier. Results of the Stage 2 test research hypotheses on two performance measurements discussed earlier. All the foregoing discussed procedures will be further described in Chapter 4.

In designing this experiment, two arguments arose from the feedback received on the proposal. The first regards the necessity of the Stage 1 survey and the second regards the type of sample for Stage 2. For Stage 1, some argue whether this stage is even needed and ask why the simulated models cannot be used directly for the target model for learning. The answer to this argument is simple. Since this research is aiming to search for a better learning approach to improve learners' understanding of a model to improve predictions, target models used to correct learners' misconceptions and errors have to be more successful predictors and closer to reality than the learners' own models. Simulation models cannot guarantee this outcome. It can be said that to correct learners' errors in thinking about real choices with unreal and inaccurate simulation models is contrary to the purpose of improving predictions through model learning.

Some may also question why marketers or managers are not used in the sample because it is obvious that learning consumer choice models is of more interest to them. Besides the obvious reason that using a sample of managers for a proof-of-concept study is costly, a more important reason is that it is contrary to the study's objective to reduce rather than increase the impact of the learners' experience and prior knowledge on learning. This researcher aims to search for a learning approach which can be generalised to solve a wider array of problems. Research of this nature should not limit itself to a particular group or context so research findings are not influenced by the unique knowledge that learners possess. In fact, whether the empirical problem is related to airline offers or any other categories, or whether learners have knowledge of the category, should not play a significant role in the research findings. It is the difference in learners' performance before and after the learning experiment that is important. Therefore, it may be of benefit to use general respondents and non-experts who hold no, or limited, perceptions of the product instead of those who are influenced by their own market knowledge. This relates to the external validity of the research findings. It is common in marketing to use a general sample and a novel category

in studies on learning. For example, Meyer used students to learn a product category (copper wires) of which they had no prior knowledge in his study. The results of his study can be said to be due to applied learning approaches (Meyer 1987).

## **1.4 Justifications for Research**

As discussed in Section 1.2, the research problem aims to identify effective ways to improve people's probability predictions by learning models. This research problem is general in nature because it does not specify the type of probability and the type of model involved. The research proceeds by conducting an experiment testing people's performance in predicting consumers' choice probabilities through learning choice models. Why is this experiment an appropriate empirical case for the research problem in marketing? A justification for this appropriateness is the main feature discussed in this section.

Besides the justification, it is also worthwhile to discuss the uniqueness of the methodology applied in this research. There are two noticeable differences from past studies on probability learning in marketing. First, two measurements are used in testing performance instead of one: prediction accuracy and model parameter learning. Past probability learning studies focused only on prediction accuracy. Second, the experiment is conducted with learners and the system interacting directly in an intelligent tutoring system. This method has not been applied in past studies in marketing. The final justification is about the potential areas in marketing concerned with the implications of this research. However, it is more appropriate to discuss this justification in more detail in the final chapter of the thesis following a discussion of the results.

### **1.4.1 Learning Choice Models to Make Predictions on Choice Probabilities**

Two reasons determine why improving predictions of choice probabilities by learning choice models is a good empirical case in marketing. First, making choices is one of the most complex



activities in marketing and it is difficult to predict probabilities of choices. Second, there exist well recognised barriers for people predicting choices accurately based on intuition.

Theories of choice differ greatly as to whether people's choices are economically "rational". If choices can be considered rational, then it is easier to gain more accurate predictions of choice probabilities. Theories of choice can generally be categorised into a normative stream and descriptive stream. The normative stream considers that choices follow a utility maximisation rule therefore choice probabilities can be predicted quantitatively. By way of contrast, the descriptive stream thinks people's choices are driven by a variety of heuristics thus violating the utility maximisation rule (Bell, Raiffa & Tversky 1988; Hastie & Dawes 2001; March 1988).

The central concept of the normative framework is that if a decision maker's choices obey some well-defined axioms such as transitivity and independence, then both utilities and probabilities of choices can be predicted using deterministic functions (Goldstein & Hogarth 1997; Luce 1959; Luce & Raiffa 1957; Savage 1954; Simon 1997; Thaler 1980; von Neumann & Morgenstern 1944). This idea provides viable ways to further develop advanced models. A major break-through came with the development of discrete choice modelling methods consistent with Random Utility Theory (e.g. Ben-Akiva & Lerman 1985; Ben-Akiva et al. 1999; Louviere, Hensher & Swait 2000; Manski 1977; McFadden 1974; Thurstone 1927; Train 2009; Yellot 1977). On the other hand, researchers from the descriptive stream have demonstrated that people can be inconsistent with one or more rational rules in making choices. Researchers found that individuals rely on simplified decision strategies adopted to overcome cognitive and computation limitations (Payne, Bettman & Johnson 1993; Slovic, Fishhoff & Lichtenstein 1977; Tversky & Kahneman 1974). For example, Kahneman and Tversky (1979) proposed that individuals identify a psychological "prospect" that

translates objective values of goods into personal values, thus their decision rules are subjective and may not be economically rational.

As Louviere and Meyer (2008) stated, although behavioural researchers focus on examples to show that people do not make choices consistent with normative theories, they have not offered an alternative paradigm that can be applied in solving real modelling problems. In contrast, random utility models have shown great accuracy in predicting people's choices. They further pointed out that research on choice is advancing towards the integration of normative and behavioural decision theories to enhance both statistical accuracy and the understanding of individual differences. There is no doubt that researchers have developed many useful methods along the way in predicting choices by including factors such as individuals' psychological and social differences. The field is extending and developing towards more robust models in this direction (Ben-Akiva et al. 1999; Ben-Akiva et al. 2002; de Palma et al. 2008).

Based on the foregoing discussion, it is clear that *choices are complex and difficult to predict*. Predicting choices requires an understanding of the models that are developed under a normative framework, as well as understanding the differences among individuals due to social and other factors. It is difficult to imagine that any experts, no matter how much knowledge they may have, could make accurate predictions about choice probabilities without the assistance of a model. This reality justifies the key assumption that people need to learn from a model to make accurate predictions about choice probabilities.

Moreover, people also face fundamental barriers in making accurate probability predictions. These barriers are associated with psychological and cognitive processes that they cannot overcome. Research has demonstrated how people may experience difficulty in dealing with probabilities

related to concepts such as distribution, independence and randomness. These findings are well documented in early subjective probability prediction literature (e.g. Hogarth 1975; Slovic & Lichtenstein 1971). Generally speaking, people draw information from incomplete cues and tend to search for information contingently and non-systematically in making judgements (e.g. Camerer & Johnson 1997; Hammond 1955). Cognitive scientists support these views with evidence showing how human minds may work. First, the human mind does not work as a general problem solver but as a set of adaptations. Second, features of the human brain are designed to achieve adaptation conveniently rather than optimally. Limitations of the human brain are common, as Luger et al. described (1994, p. 143): “although extraordinarily adept at some tasks, such as pattern recognition, it (the human brain) is slow and cumbersome at other tasks, such as numerical computation”. Unfortunately, numerical computation is also the key capacity needed in predicting probabilities. The issue is, can people be trained to improve such capacities through model learning, or alternatively, substitute the need for numerical computation with other enhanced capacities, such as using mental images, or make judgements from similarities. Although this is hard to answer, especially it lacks experimental means to directly observe what happens in people’s minds and make such connections, it is possible to test whether prediction accuracy improves, or not, given certain conditions facilitated by these potential approaches.

Due to barriers in making predictions, it is agreed among researchers that people require decision aids such as decision support systems. For example, Hoch, Kunreuther and Gunther (2001) dedicated a whole volume to this particular topic. One interesting view is that people’s intuition and the analytical model can enter into the decision process at different stages. In the initial problem framing stage, intuition plays a more important role. At a later prediction stage, formal models are more useful and accurate. This view is also supported in Blattberg and Hoch (1990).

To summarise, choices are complex and people are not inherently equipped to make accurate predictions in context. Learning from more accurate models will help them to make more accurate predictions about choice probabilities. This essentially matches the main research problem regarding improved predictions through learning the relationship between model outcomes and model parameters. Therefore, testing how to best learn from choice models to improve predictions of choice probabilities is an ideal empirical study to answer the main research problem.

#### **1.4.2 Two measurements, Two Questions and One System Used in Prediction**

As mentioned in Section 1.2, past studies on probability learning only consider prediction accuracy as the indicator in judging performance, most notably in MCPL literature (e.g. Cooksey 1996). In this research, both prediction accuracy and model learning are used to measure learners' performance.

In order to measure both performances, in the Stage 2 learning experiment, two types of questions are asked in the training tasks. The first question is a “categorical” prediction question. When learners face a scenario consisting of several options, a combination of various attributes and levels, they are asked to predict a single best option. A question like this is commonly used in prediction tasks and can be used as an indicator of prediction accuracy. The second type of question is a “numerical” prediction task which is unique to this research. According to this researcher's knowledge, such a task has not been asked in past probability learning research. This question asks people to predict choice probabilities of all product options. This question is more difficult to answer than the first question and provides much more information. It is easy to see that besides providing good information on prediction accuracy, full probability prediction provides a strong indicator to test model parameter learning. Arguably, other than learners having good knowledge of the target model, it is difficult to produce accurate full probability predictions covering all choice options. It is possible to randomly guess a single best category and still make it correct, but it is

not possible to have a close to true probability distribution prediction without a good understanding of the target model.

In this research, a dynamic and automatic evaluation, estimation and classification mechanism is also introduced into the experiment. This mechanism is built into a tutoring and feedback system which can arguably be considered an Intelligent Tutoring System (Woolf 2009). Computation using task results to generate feedback is effected without intervention from human tutors or operators. The objective of this mechanism is that it can provide target and individualised feedback to each learner without delaying continuous learning. Dynamic feedback without delay can be considered comparable to a human tutor providing learners with relevant feedback.

In summary, this research is unique with regards to methodology in two ways: first, two measurements of prediction accuracy and model learning, and questions are designed to ensure the two measurements can be tested; second, an intelligent tutoring system is facilitated by dynamic computation and feedback generation. If both approaches are working effectively in this research, this researcher expects further use of this method in future research to study probability prediction and model learning.

#### **1.4.3 Implications in Marketing**

As mentioned earlier, “plain vanilla” decision support systems only perform model computations for prediction but lack any communication or training capacities. This type of system is unlikely to be defined as a complete decision support because other than making its own predictions, it does not support a person’s decision process with comprehensive guidance and insight. This is the precise area in which this research can help to make a difference. Research findings may have direct implications for decision support and staff training. Knowing effective ways to learn models to improve probability predictions is useful because it can reduce uncertainty and risk in decision

making thus improving decision quality overall. By applying effective learning approaches, marketers at various levels can benefit from better decision support.

Moreover, the system applied in this research can be extended to an alternative decision aid tool. It can either be applied alone for training, or combined with existing decision support systems as a new type of integrated decision support system. If combined with an existing “plain vanilla” system, the combined system can play several roles: to make model-based predictions, to communicate with users about model parameters and insights, and can provide ways to train users. The guiding principle is that decision support does not simply mean the provision of an accurate model. Having an accurate predictive model is only the starting point. What is more crucial to marketing decision-making is to effectively transfer this knowledge from a model to the marketers’ minds. This helps marketers to become better decision makers overall even in situations where a physical decision support system is not available.

As mentioned earlier, more discussion of the implications of this research will be provided in the last chapter following the provision of the research results.

## **1.5 Linking Theories and Methodology**

Figure 1.3 illustrates connections between the background, research problem, main justifications and more detailed theoretical points crucial to this research. As shown in Figure 1.3, the parental discipline targets the conflict of model and intuitive predictions of probabilities. As discussed in Section 1.1, the main background is the existing gap between the subjective and model predicted probabilities. This causes conflicts in decision making which indicate the path(s) a decision maker should follow. The belief shared by the present researcher and other researchers on this issue is that the gap can be narrowed by integrating the model into people’s intuitive outlook to improve the quality of predictions made by people. This is predicated on the evidence that people are the

main decision makers not models, hence knowledge integration should be performed by people learning about the relevant models. This is the main research problem. It aims to identify the most effective way that a model can be learned. The research targets this problem from several aspects. First, it examines existing learning theories. The resultant review in Chapter 2 serves as the foundation in establishing directions for learning approaches which will be tested in this research. Second, Chapter 3 will discuss theories, methods and practices in building evaluations, model estimations and classifications into an intelligent tutor. The ensuing discussion will cover issues such as the ways to perform evaluations in a tutoring system during a learning experiment to motivate and inform learners how to make better predictions, ways to better design a training experiment, ways to construct comparable measures for testing prediction accuracy and model parameter learning, and ways to perform analysis on learners' performances.

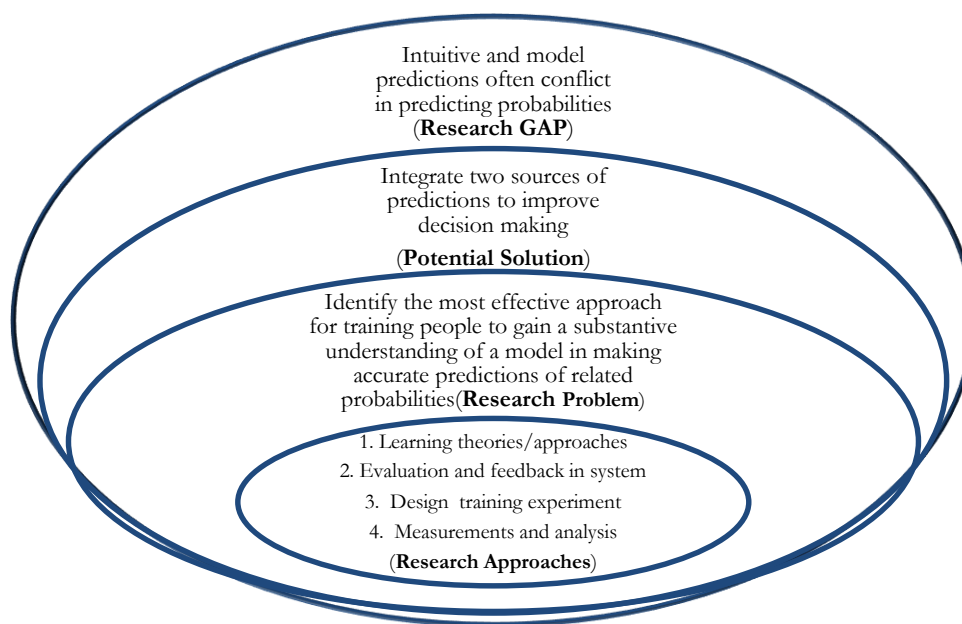


Figure 1.3 The position of this research in marketing decision making

## 1.6 Plan of Thesis

This thesis contains six chapters. Chapter 1 introduced the background of the research and provides an overview of the research problem, research hypothesis, measurements, justifications

and methodology. Chapter 2 concentrates on the literature of learning theories and the key characteristics of learning to establish the main learning approaches for testing. Chapter 3 covers theories and methods on prediction evaluation, model estimation and classification. These methods are built into the proposed intelligent tutoring system to generate real-time feedback. Chapter 4 gives details of the empirical study covering experimental design, the survey instrument, and the fieldwork plan. Chapter 5 will present and discuss the analysis results. Chapter 6 will conclude with theoretical and practical contributions, suggesting implications and limitations of the research findings, and propose future research directions.



## Chapter 2 Developing Learning Approaches

### 2.1 Overview

Chapter 1 addressed a fundamental point of this research: people can learn from models to improve their predictions of probabilities. Regardless of how a model may be taught to learners, the exact process of learning in people's minds remains largely unknown and not directly observable. What researchers can collect for analysis are people's responses to training tasks before and after learning. They then can identify the relationship between what has been taught and the learning results to try to understand the mystery of what has been learned. The mainstream of learning theories focuses on studying such relationships between stimuli and responses since both stimuli and responses can be clearly observed and measured. Other theories make conjectures about what may have happened implicitly in the learning process. However, theories targeting the overall learning process lack observation and measurement means hence they remain exploratory. It is beyond the scope of this research to study the learning process, though it is crucial to identify key characteristics common to many learning problems from existing theories that have an impact on learning performance. By identifying such characteristics, different learning approaches can be designed and tested in an experimental setting to ascertain which approach works most effectively in terms of the performance measures. This is the objective of this chapter.

There is little doubt that information in an analytical model such as its coefficients is just a raw form of knowledge because there exist many different ways of representing and teaching this knowledge. These different ways can be easily misunderstood as simply visualisation techniques or instructional design problems, the fields that have been studied by information visualisation and instructional design researchers (e.g. Merrill 1994; Tufte 1983; Tufte 1997). It is certainly true that knowing how to design information visualisation and how to arrange instructions is important. However, there are more fundamental issues related to how knowledge should be taught and

learned by people. For example, should we expect learning to occur with many small steps of improvements in trial-and-error fashion, or should we expect learning to result from several major steps of improvements with plateaus between steps? Answers to such questions are closer to the core of learning. As Norman (1985) indicated, the main reason that many studies of learning have come to “nought” is because learning involves many aspects, and failure to identify them can cause researchers to come to wrong conclusions. According to Norman, learning is about knowledge representation, input and output, thought, inference and many other things. Many implicit aspects affect the success of learning. To name a few, in learning a model, how should the model’s knowledge be best represented and taught to learners? What kind of input should learners receive before, during and after making predictions for trial tasks? How should they receive them? These questions are above the level of design features and instructions but relate to methodologies for teaching and learning.

Before discussing some important aspects of learning, the chapter will first provide a brief review of learning theories. This review introduces some concepts and background for establishing the theoretical foundation of learning approaches. As mentioned in Chapter 1, it is important to develop learning approaches that are theory-driven, because practical approaches that are problem specific cannot be generalised for a wider array of problems. This will be followed by a discussion on four key aspects of thinking about learning approaches, namely: 1) self-regulated or self-controlled learning versus guided experimentally controlled learning; 2) feedback and their effects; 3) knowledge representation, and 4) categorisation. These aspects shape the learning mechanism designed in this research and support the development of research hypotheses and empirical tests.

## **2.2 A Review of Learning Theories**

### **2.2.1 Definitions of Learning and Taxonomy of Learning Theories**

Learning covers many phenomena and is usually accompanied by a form of training in practice. For example, technicians follow training courses to learn new technical skills, doctors learn how to perform surgery by observing and assisting experienced doctors, and business professionals learn how to improve their understanding of the market by continuously following market information. Generally speaking, people observe external events and receive inputs from the environment, process such information and take actions. They then assess outcomes returned from the environment, either explicitly or implicitly, and repeat the above described process. Understanding this learning process is not universal hence research about learning is not restricted to a single theory, even a single discipline.

Indeed, many phenomena are studied under the banner of “learning”, making it an ambiguous concept to define. Mowrer and Klein (2001) defined learning as “a relatively permanent change in the probability of exhibiting a certain behaviour resulting from some prior experience (successful or unsuccessful)” (p. 2). Some definitions of “learning” are more specific and stress the positive outcome only. According to these definitions, not only should some domain knowledge be gained from experience, learners should also show improved performance when the same situations occur again (Simon 1983). Whether learning suggests improved performance is an open question for many. It is probably convenient from an experimental perspective to think that the main objective of learning experiments is to gain improvement, as this researcher aims to do. However, it is also important to acknowledge that improvement is only one of the possible outcomes from learning. As pointed out by Luger (2009, p. 388), “learning research must address the possibility that changes may actually degrade performance”. In reality, people can learn from degraded performance too. Phenomena observed during learning trials such as a plateau of no performance improvement

followed by a sudden performance improvement have been often observed in experiments (Estes 1972).

Considering different definitions of learning, it is not a surprise to see that theories of learning also vary greatly. There are in general three schools of learning theories: learning theories investigating the “behaviour” of learners (prominent in psychology), learning theories investigating the process of the “mind” (prominent in cognitive science), and learning theories investigating how a “machine” can learn as humans (prominent in artificial intelligence). These three schools of theories together define the major landscape of learning theories. They differ with regard to theoretical foundation and empirical practice, but are related on many fronts. Learning theories initially were studied in psychology. Further development came with advancements in cognitive science and AI, accompanied by increasingly advanced technology and improved research methodology (Bower & Hilgard 1981; Gärdenfors 2005; Luger et al. 1994; Luger 2009; Mowrer & Klein 2001). These theories are reviewed in the following sections.

### **2.2.2 The “Behavioural” School of Learning Theories**

Learning as a research problem has long been studied in psychology. Both theories and experimental methods were first developed in psychology (Bower & Hilgard 1981; Kimble 1985). Psychological learning theories involve the measurement of learners’ performance and responses to learning tasks. Two kinds of information are often considered crucial. The first kind of information is materials that are designed to trigger learners’ reactions and actions, the “stimulus”. The second kind of information is the learners’ responses to stimuli. Learning can be viewed as a repeated loop consisting of a sequential process from stimuli, to responses and follow-up evaluation of outcomes, or simply called “Stimulus-Response-Outcome” (S-R-O) mechanism. This mechanism is often applied as a standard framework in studying learning (e.g. Bower & Hilgard 1981).

Psychologists working in the field of learning are mostly empiricists and behaviourists who share a strong and common view that learning occurs as a result of experience and the only acceptable way to measure learning is through observed behaviour. Therefore, the purpose of learning theories under this framework is to identify key factors such as the stimuli, conditioning and feedback influencing behaviour of learners. These theories, regardless of their differences, have been loosely termed “Behavioural-Associationist” theories (Bower & Hilgard 1981; Hulse, Egeth & Deese 1980; Mowrer & Klein 2001). The term explains two distinct characteristics of these theories in general. First, ignoring specific mental processes, these theories emphasise people’s behaviour demonstrated and observed in outcomes. Whether people are really gaining experience explicitly or implicitly and whether people’s minds work in a certain way is not of interest. Second, these theories are interested in the association of stimulus and response. Rather than considering responses as being driven by clear purposes, these theories more or less suggest that responses are habits or natural reactions to stimuli which are distant from some deeper level mental activities (Bower & Hilgard 1981; Mowrer & Klein 2001).

It is important to note that early work in the field of probability learning by Estes and others was developed with this background (Estes 1972, 1994). According to Estes’ Stimulus Sampling Theory (SST), what learners learn and how they respond are under the influence of the sample consisting of different combinations of stimulus elements. By varying these combinations of stimulus elements, learners’ responses may change accordingly. Stimulus elements can change over time, across trials, or under the influence of external environment. Further sampling certain elements will result in further learning of these elements. Thinking about this theory in the context of the present research, it can well support the need to use an experimental design method in learning. Only by applying a carefully selected experimental design for stimulus elements (in this

case attributes and levels for making choices), attributes and their relationships to choice probabilities can be learned in a more explicit way.

Since the 1970s, many studies of probability learning have been conducted under the Multiple Cue Probability Learning (MCPL) umbrella. In these studies, psychologists conducted experiments to test how learners performed in learning probabilities of events that consisted of multiple cues. Frameworks of these studies are consistent with the traditional S-R-O framework, and little attention is paid to the mental processes of the learners. In terms of stimuli, not only are multiple cues sampled and presented, the conditioning of different feedback forms is also added. In terms of responses, researchers are interested in observing the incremental learning of probabilities over trials. By analysing the relationships between observed responses on a performance level and experimental conditions relating to presented cues and feedback, researchers are able to detect the impact of different conditions on performance (e.g. Brehmer 1987; Castellan 1974; Cooksey 1996; Edgell 1978, 1983; Klayman 1984; Steinmann 1976). These methods have been applied in studying learning performance in many different areas, including marketing and management (Eisenstein & Hutchinson 2006; Meyer 1987).

In summary, the traditional learning theories in psychology have provided a useful and important framework that incorporates input (stimuli), output (response), evaluation and conditioning for learning. It is also reasonable to say that designing learning studies within this framework is both clear and practical, which also explains why most empirical work related to learning has been conducted in this tradition. The limitations of these theories are also clear. There are some fundamental aspects neglected in this framework. For example, these theories do not discuss the nature of knowledge and its representation in learning. Although less obvious and difficult to observe, these aspects are nonetheless critical to the success of learning processes. Should any

knowledge be considered as just stimulus elements and should people's understanding of different knowledge be considered natural responses to these elements? Are there different ways that learning can occur, whereby different thought and inference processes are triggered, resulting in different outcomes? These missing parts work in the minds of learners and are not obvious or easily separable from the responses. These neglected aspects of traditional learning theories have sparked the development of learning theories in cognitive science (Gärdenfors 2008a; Mowrer & Klein 2001).

### **2.2.3 The “Mind” and “Machine” Schools of Learning Theories**

Research about learning in cognitive science and artificial intelligence (AI) are discussed in this section together. This is because they are well connected in many ways from their origin to current development. They are new scientific disciplines and both are yet to fully identify their theoretical boundaries and directions (Luger 2009; Luger et al. 1994). There are two overarching goals in cognitive science according to Gärdenfors (2000). One is explanatory, the other is constructive. The explanatory goal aims to explain and theorise aspects of human cognition and what occupies human minds. The constructive goal is to develop systems that can accomplish cognitive tasks by simulating or emulating human thought process (Gärdenfors 2000). Whilst cognitive science may focus more on the explanatory goal, AI focuses on the constructive goal to study what can be achieved by computer science and other technology in simulating human thought process.

According to Luger et al. (1994), cognitive science is a new science developed with efforts by researchers from different disciplines including psychology, linguistics, philosophy, computer science, neuroscience and many others. The common belief driving this development is that the science of intelligence can be regarded as common rules and principles existing in human intelligence that can be identified and utilised for constructive purposes. Different from the psychological learning theories which attempt to encompass the entire domain with a global

framework characterised by S-R-O association, cognitive science specifically targets the process of the human mind. The motivation behind this move is that existing psychological learning theories cannot give explanations for many phenomena observed in learning (Bower & Hilgard 1981; Cohen & Lefebvre 2005; Gärdenfors 2000, 2005, 2008a; Harnad 2005; Luger 2009).

An early movement towards the study of the human mind was initially made by psychologists themselves. For example, Tolman's (1948) study on animals and humans shifted away from S-R-O association and considers learning as a purposeful and active search process. The process of learning constructs a so-called "cognitive map", which serves to reach different rewards instead of passive and reactive responses to stimuli (Tolman 1948). Another example is the rise of Piaget's theory on Constructivism, which has since made a great impact on education. Piaget's theory reveals how differences in the way children and adults learn are due to structural differences in their mental processes, and how knowledge is constructed instead of being passively taught (Bransford et al. 2006). A third example of the break away from the traditional S-R-O paradigm is the development in studying human information processing to construct computer systems that can solve problems in a similar fashion. This is led by Simon and his associates, whose work revolves around the development of early AI applications for problem solving. Simon (1979) discussed constructs in information processing such as knowledge domain, semantic memory, adaptive production, pattern induction, motivation and emotion, which are not a part of concepts in traditional learning theories in psychology.

It can be said that developments of cognitive science and AI constitute a paradigm shift in studying human learning. In addition, researchers are also interested in building practical systems with capabilities to learn, which requires "mind" and "machine" schools of learning theories to work coherently. The key belief is that, if a computer system could perform in the way a person would



have responded in facing the same situations, then an abstract form of the learning and decision process can be identified and replicated external to the person. As a result, the underlying process in such a system can represent key characteristics of the mental model of the person (Luger et al. 1994). This is also the original idea of the “Turing Test” proposed by Alan Turing (Gärdenfors 2005). In reviewing both human cognitive learning and machine learning theories, it is noticeable that the taxonomies of both theories are almost identical.

Common ideas of cognitive science and AI focus on symbolic and connectionist learning. Luger (2009) and Luger et al. (1994) reviewed both classes of learning in the contexts of cognitive science and AI. Recently, a third class of learning has emerged: concept learning. The argument is that concept learning cannot be treated as symbolic or connectionist learning as it has more to do with similarities of concept properties and dimensions of these concepts (Gärdenfors 2000, 2005, 2008b; Gärdenfors & Williams 2001). Fundamentally, these three classes of learning established learning theories in cognitive science and AI and they differ in many aspects such as knowledge representation and information processing. Below is a brief introduction to the principles of each class of learning.

The central tenet of *symbolic learning* considers knowledge as symbols, and learning occurs via processing equipped with manipulation methods such as logical “if-then” rules. However, the symbol itself is not considered meaningful. What makes it meaningful is the operation of symbol manipulation. One of the key ideas is that natural languages can be represented and programmed recursively with a central processor. The central processor then determines rules which can be applied to manipulate symbols to yield outcomes (Luger et al. 1994; Gärdenfors 2000, 2005, 2008a). Symbolic learning has been the central focus especially in designing AI systems but the idea has also been criticised for some fundamental problems in the context of human learning.

Existing scientific findings relating to the human brain suggest that the notion that the human brain works as a central processor to manipulate symbols, simply does not hold. There is no such thing as a central processor in the human brain; quite the opposite in fact. Neuroscience has found that the human brain works in similar fashion to a well distributed parallel network consisting of many modules (Gärdenfors 2000, 2005, 2008a). Another fundamental problem of symbolic learning can probably be best explained by Searle's famous thought experiment called the "Chinese room" scenario (Searle 1980). In this scenario, a mechanism (person or computer program) with no knowledge of Chinese is able to demonstrate the ability to construct responses in Chinese by following strictly syntactic rules by treating Chinese characters as symbols. However, no knowledge in Chinese is gained whatsoever by the mechanism. The mechanism is simply simulating the construction process. It is not strange that computer systems favour a symbolic learning approach because of well-defined syntactic rules. However, it is a different case for human learning when gaining knowledge, rather than simulating certain capabilities, is the purpose of learning.

The motivation for *connectionist learning* theories is closely associated with the biological construction of the human brain, as mentioned previously. The human brain works as a parallel system with simultaneous activities of numerous neurons. The connectionist learning approach is mostly applied in AI and the main system developed under this notion is Artificial Neural Networks or ANN (Gärdenfors 2000, 2005, 2008a; Luger et al. 1994, 2009; Pandya & Macy 1995;). Although considered a highly useful approach in many areas such as data mining, ANN also comes with many limitations. As with symbolic learning, connectionist learning is not an approach targeted for human learning. Systems designed under this principle can perform tasks such as automatically identifying patterns when large amounts of training data are available. Even with a sophisticated computer system, such a pattern identification task is complex and slow. More crucially, it is

suitable only if a large amount of data is available for processing, and its central aim is to extract combinations of data to derive features for modelling or classification purposes (Hastie, Tibshirani & Friedman 2009; Pandya & Macy 1995; Woolf 2009). Therefore, it is difficult to apply this approach to human learners.

Concept learning proposes a process that involves learning complex concepts from their similarities and dissimilarities. A concept is defined here as a broad term meaning the basic building blocks of knowledge, and may cover many subjects such as a complex task, a skill, a knowledge domain and, in the context of this research, a model predicting probabilities. The Conceptual Spaces (CS) theory is a framework designed by Gärdenfors (2000) specifically for this type of learning. As Gärdenfors (2000) explains, the motivation for the development of this theory is the lack of understanding and methods in both symbolic learning and connectionist learning of learning concepts. In symbolic learning, the basic building block is a symbol, whereby semantic knowledge cannot be represented until operational rules for manipulation are in place; in ANN, the neuron is the basic building block but it is also not meaningful until a discernible pattern is recognised. This theory targets the level of concept in its representation and categorisation. It proposes that concepts can be learned by defining “quality dimensions”, which are semantic scales for measuring features or attributes of concepts. As suggested by Gärdenfors (2000, p. 1), *“Conceptual spaces are geometrical structures based on quality dimensions”*. In simple terms, concepts can be visualised as objects in a geometrical structure so that similarities of concepts can be represented by spatial distances in conceptual spaces. The idea is that by treating concepts in this way, a learning task can be largely simplified. This theoretical idea is different from traditional learning theory but it is a simple and clear idea if one can agree with the key assumption that a knowledge domain and its basic building blocks can be represented in this way.

The Knowledge Representation (KR) literature seems to support this belief of concept learning (Markman 1999; Rumelhart & Norman 1985). In KR, knowledge and concepts are considered as the “represented world” and this framework constructs an approachable “representing world” in geometrical space with concepts showing as objects. There are no restrictions to request that the represented and representing world be in the same form, as long as psychological interpretations and semantic connections can be established. Moreover, support may also be drawn from theories in using images for feature representation and learning concept connections directly from connected quantitative graphics (Kosslyn 1981; Tufte 1997; Wilhelm 2005).

A key problem faced by this researcher in applying these theories in cognitive science and AI is simply, that these theories lack experimental and empirical evidence to show they actually work in human learning problems. This applies even more to the field of probability learning which is occupied by methodologies developed in the context of traditional psychological learning theories. Nonetheless, knowing these theories of learning process can help this researcher to pursue directions to develop learning approaches somehow matching the underlying principles of these learning theories.

## **2.3 Key Attributes to Characterise Learning**

### **2.3.1 The Selection of Key Attributes**

In Section 2.2, three schools of learning theories were reviewed. The review made clear that psychological learning theories focus on the behaviour of learners while other theories focus on the mental process. To better connect these theories to the reality of how learners engage in learning in real-life, it is practical to treat some key features or concepts studied in these theories as attributes in the design of learning approaches.

Since there are many characteristics in the theories for selection, it is important to consider the objectives of this study and the environment in which it will be conducted. From discussions in Chapter 1, it is clear that learners will be trained to learn how to operate a model represented as a DSS and to perform both categorical and full probability prediction tasks in a computer tutoring environment. The selection of learning characteristics should relate to how to improve the learners' mental models by either self-learning or computer-assisted learning. The tutoring system needs to provide everything that is required for learning without intervention from a human trainer. Learning outcomes can only be determined by what learners are able to do and what the system structure and parameters allow to be done.

Improvements in the mental models require learners to continuously practise, either guided by their own strategies, or directed by the system. As Johnson-Laird (1988, p. 130) pointed out: "once you have some internal model of what ought to happen, you can learn by practising the skill until your performance converges on the desired model". It is self-evident that a single exposure to a probability problem cannot converge to a desired model because probability itself is established on multiple incidences. Making generalisations from multiple incidences of trials is fundamentally an inductive problem. In thinking about tasks for practice, the first thing is to refine tasks given to learners. Instead of giving learners hundreds or thousands of tasks, tasks can be designed more efficiently to reduce the effort required in converging to an ideal mental model. To achieve such efficiency, two directions may be considered. First, as suggested in some education literature, learners can be given the freedom to determine their own learning strategies to overcome inefficiency caused by using the same training program for students who have different knowledge and styles. This is the idea of self-regulated or self-directed learning (e.g. Garrison 1997; Zimmerman & Martinex-Pons 1990). Second, using an experimental design method, training tasks can be refined to represent the whole problem space with statistical efficiency such that the number

of training tasks is reduced to a smaller set of tasks but each task is more efficient. This is similar to the situation applied in discrete choice experiments where efficient experimental designs require much fewer tasks to elicit people's preferences. Similarly, the same method can be considered to apply in training. This generates the first attribute to test in terms of the learning approach: self-regulated learning versus experimentally-designed learning.

In thinking about the attributes for learning, it is difficult to miss “feedback”. Feedback is one of the most commonly discussed and well-studied subjects in learning studies. An entire literature in MCPL focuses on comparing different types of feedback and their effectiveness in learning (e.g. Cooksey 1996). Indeed, in traditional learning theories in psychology, feedback is arguably the most important factor in the process of constructing associations between stimulus and response. Learning converges on existing training tasks, so learners need to have access to evaluation for better convergence. Without feedback on performance, learners are left in the dark without further direction (Johnson-Laird 1988). In cognitive science, feedback is not discussed as explicitly as in psychology and it is often embedded in discussions of broader topics. Feedback is regarded by cognitive scientists as the foremost criterion to separate supervised learning from unsupervised learning (Harnad 2005). This view is supported by statistical learning theories. In statistical learning, supervised learning generates inferences and functions as feedback from existing or test cases and applies them to new cases. Two common approaches are regression and classification models. On the other hand, the focus of unsupervised learning is to identify hidden structures in data. It includes methods such as cluster analysis (Hastie, Tibshirani & Friedman 2009). In supervised learning, the discovery of functions and inferences is achieved by analysing existing cases by a mechanism, whether it is a human brain or a computer system. The analysis results are then evaluated by certain objective standards such as being able to further reduce errors in a

function. Before new trial cases enter the learning process, evaluation results are returned to the mechanism as feedback.

Next, in mediating between a stimulus and the response to create an association between the two, there needs to be a representation of both external knowledge and internal states of mind. Rumelhart and Norman (1985) believe that the primary task of knowledge representation (KR) is to elicit internal mental models, whereas representing external objects is a secondary task. The literature in cognitive science provides detailed knowledge classifications to represent different types of knowledge such as features, propositions, imagery, structure, concept and network (Cohen 1983; Markman 1999). KR is regarded crucial to understanding a particular knowledge domain and most relevant to establish a foundation for developing methods for computation and problem solving in that domain. KR serves two purposes in practice: first, to ensure that what learners are assessing are indeed the right underlying domains that they are supposed to learn; and second, to know whether learners are learning more effectively from a particular form of representation. As a common and important branch in cognitive science, KR has been studied widely and serves an important role in improving the mental models of learners. There are different views about how to classify representation forms. However, one view is agreed by all that the choice of KR in any particular domain makes a great difference in learning (Markman 1999; Rumelhart & Norman 1985).

One objective of learning is to generalise knowledge from many trials and group these trials into instances of same categories. In this spirit, learning is to construct a new program consisting of fewer numbers of generalised categories. In the view of cognitive scientists, categorisation is a broad concept grounded in interactions between agents and environment, and reflects different classes of environmental situations (Cohen & Lefebvre 2005). It can be linked to practical methods

such as classification and clustering as mentioned in machine learning and statistical learning literature (e.g. Hastie, Tibshirani & Friedman 2009; Mitchell 1997; Pandya & Macy 1995; Pekalska & Duin 2005). It can also be linked to theories in cognitive science and psychology and discussed in those contexts. For example, in Rosch's Prototype Theory, categorisation is considered the process of identifying *natural categories* that learners automatically try to match in learning new concepts (Rosch 1973, 1975). A large volume edited by Cohen and Lefebvre (2005) gives a good review of many works written by cognitive scientists on categorisation. Categorisation overall is regarded highly as the central theme of cognition. As put by Harnad (2005, p. 20), "to cognize is to categorize". The notion of categorisation is to find identical categories based on similarities or dissimilarities of instances.

In summary, four attributes are considered basic building blocks to construct learning approaches for testing. They cover different aspects of learning: 1) should learners conduct self-regulated learning or should a tutoring system control the way learning tasks are selected via experimental designs; 2) which type of feedback is more effective in learning; 3) how to represent target knowledge; and 4) will categorisation help learners to generalise findings?

### **2.3.2 Attribute One – Self-Regulated and Design Controlled Learning**

The fundamental problem in comparing self-regulated learning and experimentally designed controlled learning is whether learners can converge to an ideal mental model effectively with limited exposures to training stimuli and trial tasks. An analogy for this problem is finding a path to a destination while driving on the road. Self-regulated learning is like finding a path purely based on intuition and self-developed strategies, without external guidance. Design controlled learning is similar to using a GPS; it tells you the shortest way to reach the destination while ignoring other possible paths that may also lead to the destination. In the context of this research, if learners use a "plain vanilla" style decision support system without an effective communicating target model,



they may need to conduct many rounds of practice trying to find the patterns of the model. The advantage of this learning is that it is freely controlled by learners to construct any scenarios that they are interested in learning, but the disadvantage is that they are not able to find those most representative scenarios to maximise the effectiveness of the learning. On the other hand, experimentally designed controlled learning will provide a small but efficient set of scenarios so learners can maximise what they can learn from each scenario.

There are three aspects to the comparison of self-regulated versus experimentally designed controlled learning. The first aspect relates to the process of generalising from examples to form accurate assessments of probabilities. The second is about what learners do when they are conducting self-regulated learning. The third has to do with selection bias when learners self-select tasks as sample stimuli.

The generalisation process of learning from multiple incidences is necessary. Without this process, examples remain as unique as they are. Rules, principles and other connections cannot be derived from examples (Johnson-Laird 1988). In psychological learning theories, probability learning relies on inductive inference. The basic assumption is that people are able to learn probabilities from many examples. The research problem of interest in probability learning is more about whether and how people *can* actually improve through practice, not whether people can learn in the first instance. Hogarth (1975) gave an extensive review of many cases to show that due to limited mental capacity and the influence of a variety of biases, people are unable to assess probabilities accurately regardless of the amount and quality of information received. The only remedy that can improve this situation is to provide decision aids during probability learning. In contrast, Edwards argued that people are not so limited in their mental capacity that they cannot perform complex mental arithmetic as well as mental note-taking. According to Edwards, if people have ever shown poor

performance or lack of learning in probability assessment experiments, it is not because people have any inherent disadvantages. Rather, it can only be caused by poorly designed experiments (Edwards 1975).

In empirical studies, psychologists have been applying trial-and-error methods by giving learners repeated tasks, sometimes in hundreds, or even thousands. This is most notable in probability learning literature (e.g. Camerer & Johnson 1997; Cooksey 1996; Estes 1972; Goldstein & Hogarth 1997; Hogarth 1987). In the marketing literature such as in managerial and consumer learning, it is also obvious that training methods involve some form of inductive learning through experience accumulation (e.g. Eisenstein & Hutchinson 2006; Hoch & Schkade 1996). It can be said that the underlying expectation of this method is for learners to combine induction from examples and falsify temporary hypotheses they have formed during learning until correct hypotheses are found. In designing a pedagogical tutoring system to support this process, systems are required to offer multiple tasks and interactive feedback (e.g. Gulz 2008; Woolf 2009). The question is: should information and training tasks be determined by learners or designed by trainers?

In the field of education, the idea of self-regulated learning is popular. It has been applied in both classroom environments and computer-assisted environments (Barber et al. 2011; Butler & Winne 1995; Garrison 1997; Zimmerman 1990, 2008; Zimmerman & Martinez-Pons 1990). The idea is simple: learners can determine their own goals that they wish to achieve, develop their own learning strategies by self-selecting tasks, and monitor learning outcomes. In this process, they are actively directing their learning so they are not passive recipients of knowledge. The role played by trainers becomes that of the facilitator. Researchers have been focusing on issues such as motivation, self-monitoring, self-efficacy and learners' internal feedback in an attempt to form a comprehensive model of such self-directed learning (e.g. Butler & Winne 1995; Garrison 1997). As pointed out

by Zimmerman and Martinez-Pons (1990), in self-directed learning, performance and self-efficacy are determined by strategies that learners choose by themselves and these strategies can vary greatly due to individual differences. Performance and self-efficacy in turn also influence the further selection of goals and strategies of learning. In addition, external task characteristics also influence learners' choices of goals and strategies (Butler & Winne 1995; Barber et al. 2011). According to these researchers, the main problems of self-regulated learning are clear and include misspecification of goals and tasks, and the misperception of cues and performance. Although self-regulated learning is attractive because it allows learners to have freedom in determining their goals and strategies, the outcomes of this method may not be ideal. Even more, due to the possible misspecification of goals and tasks, relying on learners themselves to design their own learning program can be quite ineffectual.

An important point that has not been covered by any researchers in the field of self-regulated learning is the problem of selection bias in self-selected tasks. If a target learning problem involves a large number of scenarios with enormous numbers of combinations, it is impossible to train with all possible scenarios and only a subsection can be selected for training. This is a common phenomenon not limited to the topic in this study. For example, different combinations of training can take place in teaching people how to operate advanced equipment or a program in difficult software languages. Training materials commonly only cover a well-selected small set of cases to maximise what learners can learn from each case. To find a good set of training tasks is not an easy task itself though it can determine the outcomes of learning. But what are the criteria for selecting a small but representative set of cases to allow efficient learning?

In original work by Heckman and others on selection bias, there was not much fundamental interest in what caused the sample to be selected with bias. The interest was to fix the biased

sample problem with a statistical solution. Selection bias has been discussed in the context of biased samples in social and economic research. For example, the work on sample selection bias by Heckman was discussed in the context of an economic study of wages and labour supply amongst females (Heckman 1979; Vella 1998). It has since been applied in assessing social programs such as training programs for unemployed people (e.g. Heckman & Hotz 1989). Regardless how sample selection bias may be caused, when it does occur, there are unobserved characteristics involved. Failing to identify or estimate the effects of these missing characteristics can cause huge errors in estimation.

In the context of this research, learners need to sample from a large combination of available training tasks; this is equivalent to a market researcher selecting samples from a population. For example, assume that a consumer choice model that learners want to learn is in the simple form denoted in Equation 2.1, where the whole population of tasks is denoted as  $I$ . If all tasks are observed, in the learner's mental model, there should be a record or observation  $X_i$  for each task.

$$E(Y_i|X_i) = X_i\beta \quad (i = 1, \dots, I) \quad (2.1)$$

For the sake of argument, we assume there is a “perfect” learner who can learn and record every task from 1 to  $I$  to draw a correct conclusion, so the consumer model function should be recovered without any extra errors besides initial in-built errors in the model. Because it is impossible to learn all tasks from 1 to  $I$  and only a subset of tasks can be selected and used in learning, the actual function that this learner can realistically learn is not the function in Equation 2.1, but the function in Equation 2.2.

$$E(Y_i|X_i, \text{sample selection rule}) = X_i\beta + E(U_i|\text{sample selected}) \quad (2.2)$$

In this function, the expected unobserved error component  $E(U)$  is conditioned on the sample (task) selected. As will be further discussed in Chapter 3, the purpose of experimental design is to minimise the unobserved error component, so it is reasonable to think that if the tasks are selected

following this principle,  $E(U)$  can be reduced to a minimum. In this case, the estimation of  $\beta$  from the subset of tasks from the experimental design is only a “loss of efficiency” (Heckman 1979, p. 155). In contrast, there is no guarantee that self-selected tasks can follow this principle of experimental design; therefore, the expected error component becomes much larger for self-selected tasks than with experimentally-designed tasks.

From a learning perspective, what we learn and how we respond are subject to samples of stimulus elements shown to us. If there are missing elements or combinations of elements in samples, learners are less likely to learn effectively (Estes 1950).

In summary, this section discussed two opposing views in selecting learning information and training tasks. This served to introduce the first attribute in designing learning approaches.

### **2.3.3 Attribute Two – Feedback**

As discussed earlier, feedback plays a crucial role in learning. Apart from other attributes such as knowledge representation and categorisation, which are rarely mentioned in learning literature, feedback has always been the focal point for studies in learning. There are many types of feedback; for example, feedback on outcomes, feedback on task characteristics, feedback as procedural instructions, rewards and punishments on performance and many more. It is fair to say that the types of feedback and the number of studies about feedback are enormous. Several thorough reviews and meta-analyses of feedback in learning are given by Kluger and DeNisi (1996), Hattie (1999) and Hattie and Timperley (2007). To have an idea of the volume of studies in this area, in Kluger and DeNisi’s meta-analysis, 607 effect sizes and 23,663 observations were covered. Even more, Hattie’s 1999 review covers 500 meta-analyses representing a variety of types of feedback in learning.

Feedback is considered as “information provided by an agent (e.g. teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding” (Hattie & Timperley 2007, p. 81). Kluger and DeNisi (1996) referred to it as *feedback intervention* (FI), but offered a definition very similar to Hattie and Timperley. In general, the feedback discussed here refers to external feedback on either the learning performance or process. It does not include internal feedback that learners generate by themselves during learning. It is agreed by the above-mentioned authors that not all feedback is good for learning and performance improvement. Indeed, quite the opposite; certain feedback can have negative effects on performance. For example, extrinsic rewards and tangible rewards can significantly undermine the intrinsic motivation of learning interesting tasks. Feedback such as “you should have performed” in a particular way or at certain level is considered as threatening to learner’s self-esteem and is shown to cause even worse performance (Hattie & Timperley 2007). As Kluger and DeNisi (1996) concluded, feedback intervention in general improves performance by about 40% but over one-third of the feedback types can decrease performance. The question is: *which types of feedback should learning approaches adopt in this study?*

Views held by Kluger and DeNisi (1996) and Hattie and Timperley (2007) on the role of feedback in learning are similar. According to them, there exists a hierarchy of feedback. The hierarchy starts from feedback on task learning, and moves up to feedback on task procedure, task motivation and finally on self-related characteristics. Moving up this hierarchy, the positive effects of feedback decrease. This means, the most effective type of feedback for performance is closely related to the actual task at the bottom level in the hierarchy. It may sound strange because one would think that feedback on individual characteristics should be quite effective. As explained by the foregoing authors, feedback about self-related characteristics actually does not offer useful task information for learners to correct judgements.

So what kind of feedback is task related? Feedback on tasks mainly includes: 1) corrective feedback or “outcome feedback”, which is a correct answer given to a learner after a predictive answer to a task question; and 2) cognitive feedback, which includes some or all statistical information about formal task characteristics such as models and decision insights behind predictions, and information about learners’ cognitive characteristics and other relationships (Cooksey 1996; Kluger & Denisi 1996). These two types of feedback have been extensively studied and compared in MCPL literature (Cooksey 1996). MCPL, or multiple cue probability learning, as summarised by Slovic and Lichtenstein (1971), is part of probability learning research to see how people can learn the relationships between a set of cues and environmental situations (criteria). Functional relationships between cues and criteria are typically statistical, as in a regression model. In the process of learning these functional relationships, either outcome feedback or a variety of cognitive feedback can be applied to improve performance. While outcome feedback simply reports the correctness of answers, leaving cognitive activities up to learners, cognitive feedback can provide suggestions on how and why decisions should be made. A large number of empirical studies have compared the two types of feedback and most researchers agree that cognitive feedback may be more effective than outcome feedback.

For example, Hammond, Summers and Deane (1973) compared outcome feedback with cognitive feedback. They found that not only was outcome feedback unnecessary for performance improvement, it actually impeded performance improvement. According to Schmitt, Coyle and King (1976), improvement by outcome feedback is significantly less consistent than improvement by cognitive feedback with task information. They found that cognitive feedback with both cue utilisation and task information produces the best matching between answers and predictions. One advantage of cognitive feedback over outcome feedback is that it can better identify irrelevant information which may hinder performance (Sengupta 1995). Researchers then went further to

test and compare different types of cognitive feedback. Steinmann (1976) tested feed-forward information in combination with cognitive feedback. In his study, statistical information about the task is given as feed-forward information before tasks are provided. Statistical information about subjects' own performances in addition to information about the tasks are included as part of the cognitive feedback subjects received. Task complexity was then varied on cues and their relationships to the criteria. Steinmann found that, in both simple and complex tasks, subjects perform well with this combined feedback. Klayman (1984) argues that cognitive feedback focuses only on relationships between cues and criteria and cannot capture an important aspect of probability learning, namely, the generation of new predictive cues. Hence, such feedback encourages an inappropriate deterministic mental set. Balzer, Doherty, and O'Connor (1989) decomposed cognitive feedback into three components: information about the task system, information about the subject's cognitive system, and information about the relationship of the task system to the cognitive system. Their review suggests that information about task is mainly responsible for performance improvement while the other two components have a less significant effect.

A more useful comparison for this study was given by Arunachalam and Daly (1996). They compared two types of cognitive feedback, judgement policy feedback and model prediction feedback. Judgement policy feedback is feedback that directly answers the accuracy of stated judgement policies such as cue weights and functional forms applied by learners. Model prediction feedback is based on predictions made according to learners' judgement policies. This is equivalent to a learner model, the results of which are given back to learners. Arunachalam and Daly (1996) found that model prediction feedback is more effective than judgement policy feedback, because it exactly shows how learners' policies differ from correct policies in terms of prediction results.



In terms of which feedback may perform better in testing, this researcher holds the view that it may be conditional on the type and complexity of the learning problem. It may also depend on which learning style a learner adopts. For example, in really simple tasks, such as in predicting an on and off signal in multiple trials, learners may simply take a frequency count of each event in their minds and follow frequencies mentally. This was observed in probability learning studies and indicated from early learning theory literature (e.g. Estes 1976). If this is the case, outcome feedback can actually be quite effective and anything more complex than a simple frequency count in cognitive feedback may not work better. Another possible condition under which outcome feedback may be more effective is when learning problems are too complex and cognitive feedback too difficult to comprehend; for example, when a learning model has many parameters. In cases when there are a limited number of relationships that can be clearly explained, cognitive feedback should be more effective because it provides more task related information. In this research, both types of feedback and the different types of cognitive feedback are tested.

#### **2.3.4 Attribute Three – Knowledge Representation**

Attributes One and Two largely associate well with the traditional S-R-O framework in psychological learning theories. The underlying question relating to Attribute One is how to cause learners to learn effectively from stimuli organised by learners or by controlled experimental design. Attribute Two uses learners' responses for evaluation and provides evaluation feedback to learners as new stimuli. The ideas of Attributes Three and Four discussed in the following two sections are knowledge representation and categorisation. They do not match directly with the S-R-O framework based on observed behaviour. In a way, these two attributes make assumptions about potential mental processes that learners may employ in learning if a particular learning approach is applied.

Attribute Three is knowledge representation or KR. It is a central issue of cognitive science in studying cognition. This attribute is not clear because it has not been tested in the context of learning real-life problems by researchers in psychology, education, business or other disciplines. So far, it has largely remained as a theoretical concept rather than an empirical research subject. In the context of learning, it is crucial to think about how knowledge is represented, taught and stored. Knowledge representation, as Rumelhart and Norman (1985) put it, is a notation of mental models and mental activities. Perhaps a better way to describe knowledge representation is the relationship between the “represented world” and the “representing world”. Knowledge to learners can be considered as the “represented world”, or the actual domain knowledge regardless of any particular ways in which it is manifested. Information that learners actually receive can be considered as the “representing world”, or knowledge in a particular form that works on learners’ cognition (Markman 1999). What learners see is only a chosen form of the underlying knowledge but not the knowledge itself. The two worlds are only probabilistically related. Therefore, the objective of designing an effective learning approach should aim to most effectively reflect the underlying knowledge that learners are supposed to learn.

There are many categories of KR and researchers have similar and different views on how knowledge should be represented. Since this is a vast and relatively inconsistent field, this research will focus only on one particular argument of KR, namely, the argument between advocates for propositional and imagery KR (Anderson 1978; Cohen 1983; Kosslyn 1985; Markman 1999; Pylyshyn 1973; Pylyshyn 1981; Rumelhart & Norman 1985). Without spending substantial effort trying to define what is propositional and imagery KR in abstract terms, it is more meaningful to compare their differences concerning relevant ways to represent an analytical model to make its knowledge accessible for learning. One approach is to list a combination of propositions regarding features, rules and relationships identified in the model. These propositions are addressed formally

and explicitly, in either verbal or graphic formats. A different way is to express model relations in spatial forms such as in two-dimensional graphics with each dimension representing a model property. In the former case, graphics are used simply as a way of presentation, parallel to verbal expressions. In the latter case, graphics are used for demonstrating structural relationships in the model without formally announcing them. The desired cognitive activity is for people's mental images to be created seamlessly. In practice, both representations can be used jointly, with more or less emphasis on one of the two representations.

Researchers supporting the propositional representation insist that the represented world should be reflected by propositions. These propositions can be any formal statements on logical relationships, structures, connections, concepts or any other forms of propositions. Cognitive operations are regarded as *activation* and *manipulations* of these propositions (Cohen 1983; Pylyshyn 1981). In research supporting propositional representation, visual images can be used but they are not independent constructs, so have no major roles to play in cognition. This view is consistent with symbolic learning and connectionist learning discussed earlier in cognitive learning theories. Propositions can be naturally considered as major building blocks in learning. Activation and manipulation of these blocks define cognitive activities. In contrast, researchers who support imagery representation (or "imagists" in some literature) believe that there exists a functional role for mental imagery. Not only can actual images in the represented world be represented and directly mapped to the representing world as mental images, spatial relationships such as those between objects and geometrical structures, concepts and conceptual relationships, can all be represented as images and map to mental images. The leading researcher in developing this theory is Kosslyn (1981, 1994), but spatial representation is just one way of thinking about imagery KR (Markman 1999).

The real challenge in the study of KR is how to conduct suitable empirical tests. As pointed out by Cohen (1983), both KR approaches seem resistant to experiments because both views can find explanations in opposite results. For example, in a study done by Kosslyn, Ball and Reiser (1978), people were asked to first learn the locations of objects on a map, then asked to search their mental images to recall distances between two locations. Results show that subjects seemed to use mental maps effectively to retrieve distances between objects without actually looking at physical maps. Such results clearly favour the imagery approach. However, researchers who support propositions argue that subjects were simply responding to demand characteristics and they would first need to interpret the map and decode distances between the subjects into formal propositions (Cohen 1983; Markman 1999). Some believe this argument regarding two different KR approaches cannot be solved experimentally, because neither can find distinct properties which can be isolated by behaviour. No psychological data can inform us whether people's internal representation is in this and not the other form (Anderson 1978; Cohen 1983).

Although it may be difficult to identify which KR approach is actually implemented in people's mind through experiments, it is however possible to test the effectiveness of a learning approach facilitated to support a particular KR approach rather than supporting other approaches.

### **2.3.5 Attribute Four – Categorisation**

Classifying learners based on their predictions using rules such as Bayesian classifiers is common. Thinking widely about real-life applications such as in e-commerce, it is fair to say that classifying people based on their patterns of behaviour is a common application. There is a key difference between categorisation and classification although the two terms are often used interchangeably in practice. According to Markman and Ross (2003), categorisation is a broader concept referring to a mental process of acquiring and using categories. Categories are defined as “groups of distinct abstract or concrete items that the cognitive system treats as equivalent for some purpose” (pp.

592–593). On the other hand, classification refers to methods used to reach classification models in order to identify categories. Three common classes of classification models are widely recognised. They are *exemplar models*, *prototype models* and *rule-based models*. The exemplar and prototype models are similarity-based models whilst exemplar models require information integration from examples. Rule-based models focus on identifying rules and classifying instances into classes according to identified rules (Markman & Ross 2003). One aspect of classification is to identify classification models for the purpose of category learning.

Categorisation plays a profound role in cognitive activities and has many cognitive functions. For example, using the above mentioned three classes of classification models, cognitive neuroscientists have found, from functional magnetic resonance imaging (fMRI) studies, that when people are conducting categorisation learning, different brain regions are activated. Results even show that for exemplar-based categorisation learning, human brains can have different patterns depending on the number of examples for information integration (Ashby & Ell 2001). This suggests that not only are humans capable of categorising stimuli and learning categories such as discussed by Medin and Shaffer (1978) and Nosofsky (1986), there is indeed solid neuroscience evidence demonstrating that category learning is a *natural* cognitive activity.

Concepts are often considered as mental representations of categories. Therefore, category learning can be addressed as concept learning. Komatsu (1992) provides a thorough literature review of the three major schools of concept learning theories covering *feature abstraction*, *prototype* and *instance comparison*. Linking these theories to classification models, it is obvious that these theories match well with the three types of models: rule-based models (feature abstraction), prototype models (prototype) and exemplar models (instance comparison) respectively. These three theories shape the picture of category or concept learning.

To explain very briefly following reviews by Komatsu (1992), feature abstraction theories are developed under the influence of the traditional S-R-O framework. Theories formed under this assumption focus on extracting common features from incidences in order to form concepts. Mental representation in this concept acquisition form is the process through which a list of common features (attributes) is developed. New incidences are then compared according to this feature list. The main criticism of feature abstraction theories is that categories often have fuzzy boundaries with regard to features, thus it is questionable whether using features to represent categories is an effective method.

As distinct to feature abstraction which is an approach working from the bottom level, the prototype approach works from the top level. Prototype theory developed by Rosch and her colleagues is a classical theory of categorisation (Rosch 1973, 1975; Rosch et al. 1976). The focus of this theory is around the concept of “prototype”. A prototype is not a list of features or any particular example, but the “*central tendencies of categories*” (Rosch 1973, p. 328). It is non-arbitrary and has most of the attributes common to all members in a category (thus “highest cue validity”). This theory holds that a category reflecting a prototype is considerably easier to learn than a category violating it. This theory articulates a clear structure of categorisation levels with prototypes. The defining hierarchy of categorisation levels consists of a basic level, a subordinate level and a superordinate level. The basic level is the class that people naturally assign an object to, for example, any tables that people see belong to the class “table”. The subordinate level can be considered as the property dimensions of classes, for example, for “table”, property dimensions may include height, weight, and style. The superordinate level is a higher class consisting of many classes with common attributes. For example, “table” can be grouped into furniture. Though intuitively easy to understand, the main problem of prototype theory relates to empirical tests. Hampton (1995) gave several cases showing difficulties in testing prototypes. For example,

features defining prototypes may be fuzzy or not independent. Therefore, there are no clear boundaries between prototypes and for similarity-based categorisation results can be different due to non-independent features defining prototypes. It is important to note that experiments used to establish the theory, such as those conducted by Rosch and her colleagues, did not have complex concepts. Whether or not the theory will work equally well in learning complex concepts requires further testing.

Putting into the perspective of empirical testing in this research (a learning consumer choice model), using feature abstraction and prototypes for category learning is equivalent to learning consumer choices through product attributes and consumer segments. Product attributes can be considered as a list of features in feature abstraction. Consumer segments can be considered as prototypes. The difference in concept learning literature is that features and prototypes are less complex. Therefore, this study provides a good test bed by applying elements of these theories in learning. Moreover, these two underlying learning approaches also relate to using proposition or imagery representations. The idea of feature abstraction strongly implies that the format of knowledge representation is a list of formal propositions, and learning occurs when this list is looked up and refined. On the other hand, the idea of categorisation levels in prototype theory suggests that some key dimensions are used to define prototypes, and a way to represent these dimensions is through the use of multidimensional techniques to reduce problem dimensions and visualise key properties of prototypes. In doing this, similarities of categories (prototypes) can be converted to distances in spatial representation. This is also the key idea of Conceptual Spaces Theory (Gärdenfors 2000). The theory holds that concepts can be visualised as objects in a geometrical structure so that similarities of concepts can be represented in a “conceptual space” defined by “quality” dimensions. This further proves that a learning approach based on similarity-

based visual representation is a different learning approach overall to the traditional feature abstraction approach originating from the S-R-O association idea in psychology.

Instance-comparison categorisation learning mainly works by categorising new training cases based on existing identified categories. Therefore it can be regarded as an extra support to similarity-based category learning, and negates the need for a feature list. The theory holds that for many complex concepts, the only way to acquire them is through constantly comparing similarities of new instances to existing categories. It argues that learning analytically, such as through a feature look up table, makes it difficult to define new training cases' category memberships (Komatsu 1992). Since this is regarded as an ancillary idea to the prototype approach, it will not be further discussed here.

To summarise, through the foregoing discussion, two distinct category learning approaches have been clarified, namely feature abstraction learning and prototype learning.

### **2.3.6 Summary**

Regardless of how many different theoretical views are covered in this section, the purpose of the discussion remains clear: to design and test theory-based learning approaches to improve probability predictions through acquiring knowledge from a model. As pointed out earlier, if the idea is purely to develop a practical tool without a theoretical foundation, at best such a tool can work well to solve a particular problem. By reviewing rich theoretical views and arguments relating to four key attributes of learning, it is possible to develop strong theory-based learning approaches with generalisation value and avoid the problem of relying on vague and often incorrect assumptions about how learning may occur.



## 2.4 Building Learning Approaches under Extended Framework

### 2.4.1 Extended Framework for Learning

This chapter reviewed different learning theories and frameworks from different disciplines under which learning is studied. One important framework is the traditional S-R-O association framework treating learning as a repeated loop from stimulus to response then to outcome. Another example framework is the concept learning framework developed in cognitive science using similarity-based concept learning with objects in quality dimensions to show differences of concepts. Among these frameworks, the key difference is whether a framework addresses the learning process. In behavioural learning theories under the S-R-O framework, process is not covered because under this framework only observed behaviour can be considered evidence for learning. The learning process is not directly observable therefore theories regarding process are conjectures. Thinking empirically, although we cannot measure the learning process directly, we can however design learning approaches which are best suited for a particular learning process to be adopted by learners. For example, we can provide a formal list of features and rules in stimuli to support feature abstraction learning, or provide descriptions of categories to support category learning, or provide outcome feedback to support trial and error learning.

The discussions in this chapter provide adequate evidence to extend the traditional S-R-O framework. This framework is labelled the “stimulus-process-response-outcome” framework which can be abbreviated as the “S-P-R-O” framework. There are two important associations here. This first association is the one between stimulus and process. This researcher believes, the design of a stimulus should not be considered as purely task information or instructional information, but rather the stimulus should best facilitate a particular target learning process considered as most effective for the task. The learning process relies on a stimulus and eventually determines the response and outcome. For example, giving propositions in verbal forms suggests the learning

process will involve understanding, memorising and implementing. Providing graphics with relationships of classes or features illustrated triggers a learning process of building mental images. This part of stimulus design can be considered as “process design” which is to determine the closest aligned possible learning process.

Another association is the one between process and response. If a different learning process is applied, different responses may be triggered. A learning process leaning towards applying rules may provide better responses if answers require computation and logical operation, but it may be less effective in giving answers in other forms such as visualised forms. In contrast, a learning process based on mental images may give clear directions or make good predictions about how distant are two objects, but can be inadequate in giving answers to other forms of questions.

Besides the above two associations, the third association between response and outcome are well covered in traditional learning theories, especially in the literature on feedback. Together the three associations define this extended learning framework and form a loop structure. Once responses are evaluated as outcomes, information is returned as feedback as a part of new stimuli for the next round of learning. This association between outcome and stimulus is considered to be the fourth association. The extended framework is depicted in Figure 2.1.

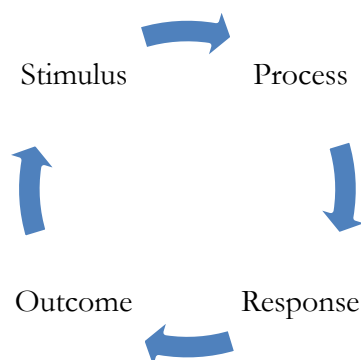


Figure 2.1 Extended S-P-R-O Learning Framework

We have discussed four attributes derived from learning theories. Attribute One (self-regulated versus experimentally design learning) clearly targets the association between stimulus and process. The assumption is that if we design stimuli efficiently using experimental design, we can expect an effective learning process which will lead to better responses. Attribute Two (feedback) targets the associations between response and outcome and between outcome and stimulus. Once responses are provided, either simple or complex evaluation is conducted to determine outcomes. These outcomes are then used to enhance stimuli for the next round of learning. Attribute Three (knowledge representation) targets associations between stimulus and process. By treating stimulus as the “represented world” with some forms of representation as the “representing world”, different learning process may be applied. Attribute Four (categorisation and classification) again targets the association between stimulus and process, by designing a stimulus best suited for categorisation, the assumption being that learners may adopt categorisation in their learning process.

#### **2.4.2 Designing Learning Approaches for Testing**

Under the extended framework and using the four discussed attributes, it is possible to develop many different learning approaches. However, the purpose of this research is not to exhaust all attributes relating to learning and all combinations of attributes to identify the most effective combination for learning. This would be an impossible task because the number of attributes and the number of combinations are both practically *infinite*. However, it is possible to choose a small set of representative learning approaches using identified attributes and test them in an experiment. The results of such experiment will provide insights as to how each key attribute may work in supporting learning. The effects of some of the attributes may be confounded in this context, however, once a proof of concept study has been conducted to show that some learning approaches work more effectively than others and provide better learning and more accurate predictions of probabilities, then it is possible to generalise the method and learning approaches

for further research. Four such representative learning approaches have been developed for testing in this study and they are shown in Figure 2.2.

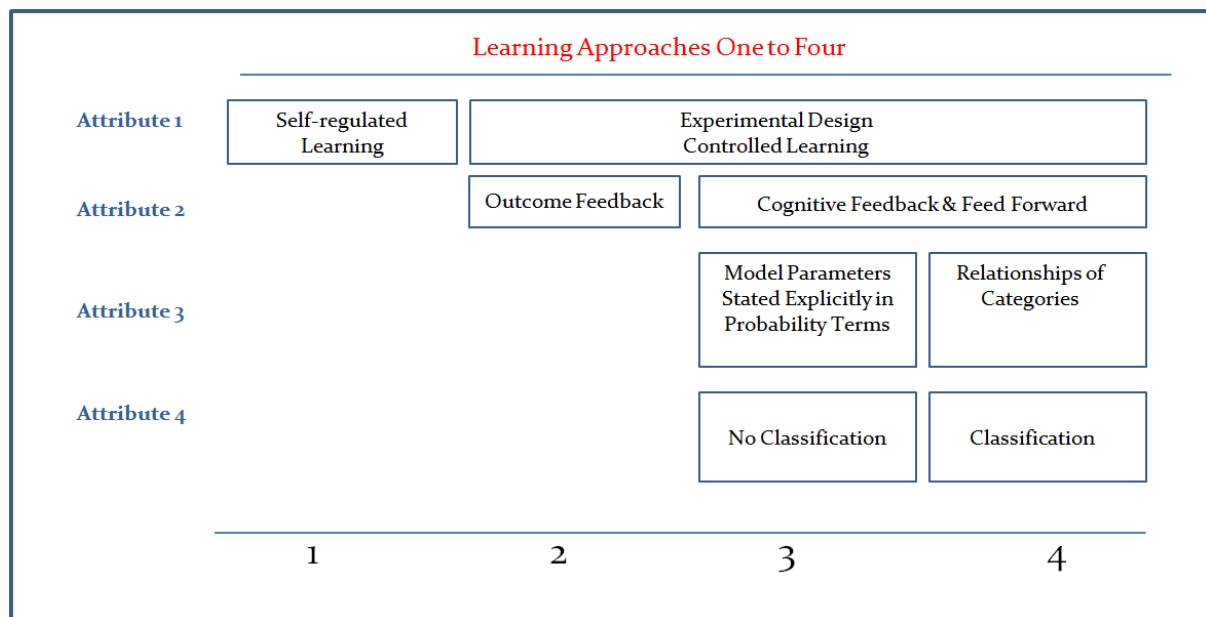


Figure 2.2 Four Learning Approaches

As shown in Figure 2.2, each learning approach matches with one experimental condition for empirical study. The following paragraphs will summarise these four learning approaches separately.

Learning Approach One (Experimental Condition 1): In this approach, learners design their own stimuli. The combinations and the time of learning these stimuli are controlled by learners. They do not receive feedback after training tasks to inform them about their performance and the stimuli are not designed to support any particular learning process. Put simply, learners are self-supported. In the context of the empirical study in this research, learners have to learn a consumer choice model by constructing examples of choice scenarios by themselves. They learn from the outcomes of these scenarios without explanation, and no extra feedback is provided after completing the training tasks.

Learning Approach Two (Experimental Condition 2): Learners are given a limited set of experimental design controlled training tasks to make predictions. These tasks cover the problem space efficiently to allow learners to learn from a limited number of combinations. After making predictions in each training task, learners receive feedback informing them about the correct answers. This learning approach matches with what is traditionally used in probability learning studies in psychology.

Learning Approach Three (Experimental Condition 3):

Learners are given the same experimental designed training tasks as in Learning Approach Two. Learners are given feed forward information before starting training tasks and in-depth cognitive feedback after each session of training tasks. The characteristics of information are demonstrated explicitly with model parameters shown in probability terms in both verbal and graphic formats. Classification methods are not applied, and learners are most likely performing the process to develop and refine a list of key features important for outcomes. The expectation is that learners are able to integrate feature-level information into the overall concept to improve their prediction performance.

Learning Approach Four (Experimental Condition 4): Learners are given the same experimental designed training tasks as in Approaches Two and Three. The focus is to learn from feed forward information before tasks and cognitive feedback after tasks. Categories and their similarities are illustrated and explained in chart and words showing relationships of categories based on class model parameters. Cognitive feedback informs learners whether they understood the target category and its model. The main expectation is that learners are able to differentiate categories through learning the relationship chart and classification outcomes in feedback.

Of all the discussions in this chapter, one area has not been addressed but plays a key role in learning. This is the discussion regarding methodologies used in evaluation, estimation and classification. Learning approaches can be directly applied and felt by learners as conditions for learning, but evaluation, estimation and classification occur in the background of a tutoring system. This mechanism works through the tutoring system and plays a key role to support and facilitate proposed learning approaches. For example, estimation results of learning models provide learners with individual level feedback, and classification results inform them whether they are learning the correct category. These features are key components of Learning Approach Three and Learning Approach Four. Even for a simpler learning approach such as Learning Approach Two, the main feature (outcome feedback) needs to demonstrate correct answers to training tasks based on an estimated model. Chapter 3 discusses these methodologies in the context of this thesis.

## Chapter 3 Evaluation, Estimation and Classification

### 3.1 Overview

Chapter 2 discussed learning theories and proposed four learning approaches under an extended learning framework. This chapter discusses methods to evaluate probability predictions, estimate learner models and classify responses to known classes. These methods cover both the experiment stage by using learners' responses to training tasks to provide immediate feedback for further learning, and post-experiment analysis using collected data to test hypotheses of prediction accuracy and model learning. To be more specific, in the experiment stage, for Learning Approaches Three and Four, a combination of probability evaluation, model estimation and classification are required to generate more complex and dynamic feedback specific to each learner. In the subjective probability prediction literature, researchers often use the terms *ex ante* and *ex post* to describe the stages pre- and post-prediction (e.g. Friedman 1983; Winkler 1996; Winkler & Murphy 1968). Methods applied in the *ex ante* stage should encourage reliable and coherent predictions, whereas methods applied in the *ex post* stage should examine whether experiments have achieved expected objectives such as improving accuracy and model learning outcomes.

As mentioned in Chapter 1, in the learning experiment, the mechanism performing evaluation, model estimation and classification is built into an intelligent tutoring system. According to Woolf (2009), several components are key characteristics of intelligent tutoring systems: *knowledge base* or what is being taught to learners; *learner model* (also called “student model”) representing the learners' status of knowledge to support their responses for training tasks; *evaluation agent* performing evaluation, estimation, classification and other functionalities; and *communication approach* providing information and instructions to help learners. Among the four components, learning approaches discussed in Chapter 2 can be considered as the *communication approach*. In this research, the knowledge base is established from the results of the consumer choice models as discussed in

Section 1.3. The remaining two components, *learner model* and *evaluation agent*, are the focus of this chapter.

The *learner model* makes continuous learning possible without delay and satisfying the objective of advanced feedback can communicate with learners about their own models. The idea is that by comparing differences between their own models and a target model, learners can correct errors. The *evaluation agent* is the mechanism performing various functions to capture learner models. A similar concept is a “learning machine”, a common term used in the statistical learning theory literature (e.g. Cherkassky & Mulier 2007; Vapnik 1999). Exactly as the term suggests, its task is to learn from training responses that learners give in training tasks.

The learner model can provide results that are individual and task specific. It can be communicated back to learners via feedback designed as learning approaches. Recall the proposed S-P-R-O association framework illustrated in Figure 2.1, these methods performed by the evaluation agent (learning machine) are the key functions linking response and outcome (R-O) and provide new sources of information linking outcome and stimulus (O-S) in a continuous learning loop.

In the post-experiment analysis, identical or more robust analysis approaches to evaluation and estimation can be applied. Results can provide evidence to test hypotheses and support research findings. Without the restriction of fast computation typically required for dynamic feedback, better statistical fits can be expected by using more robust methods. For example, the maximum likelihood estimation (MLE) method is more suitable for post-experiment analysis whilst the ordinary least squares (OLS) method is better for fast computation. Among the methods of evaluation, estimation and classification, classification is used only during experiments for Learning



Approach Four designed to use the categorisation learning approach (learning from classes). Evaluation and model estimation are important for both stages.

This chapter is organised into three sections. Section 3.2 discusses subjective probability prediction theory and evaluation methods using “scoring rules”. Section 3.3 discusses estimation approaches developed on the basis of discrete choice model methods. The objective of this discussion is twofold. First, it provides theories and methods for developing two types of models for learning experiment. One type is an aggregated model and the second type is a model of latent classes. Second, the discussion also covers how to estimate each individual learning model. This is applied in both pre-experiment and post-experiment analysis. Section 3.4 discusses the classification approach relating to categorisation learning in Learning Approach Four. Together these three types of analysis define the role performed by the evaluation agent in the tutoring system applied during learning and the analysis of collected data in post-experiment hypotheses testing. To summarise, the chapter will provide theoretical and methodological discussions of methods and models for both the pre-experimental and post-experiment analysis stages. It is important to note some exact analysis approaches used in the empirical study will be better articulated in Chapters 4 and 5 when the experiments are discussed. This chapter focuses on theories and methods supporting exact approaches.

## **3.2 Evaluating Probability Predictions**

### **3.2.1 What is Good Probability Prediction?**

Subjective probability is a field with a long history of research. Research in this field generally has two objectives: first, it aims to elicit people’s subjective probabilities consistent with common probability principles and encourage people to provide honest and accurate predictions; second, it studies how to best evaluate people’s subjective probability predictions. It is commonly

acknowledged that the foundation for this field was established by de Finetti and Savage during the 1930s to 1960s. They jointly developed a theoretical framework which was adopted by researchers who studied subjective probabilities. This framework consists of axioms to serve as principles for subjective probabilities such as comparability and transitivity, and their applications with regard to a behavioural perspective (e.g. Fishburn 1986; Savage 1971, 1972; Suppes 1994). The initial focus of the theory is on how best to elicit subjective predictions *ex ante*. For many situations where probability predictions were discussed in the early literature, true probabilities are not known *a priori* at the prediction stage, as for example in forecasting weather events in meteorology (e.g. Winkler & Murphy 1968). What are available are two probabilities: the predicted probabilities, and the best expected probabilities from assessors. In evaluating predictions, it is often the case that these two probabilities or probability distributions are compared. Later research focused on how best to evaluate subjective probability predictions by comparing predictions with actual probabilities by using a variety of “scoring rules”. The properties and effectiveness of various scoring rules became the key subject in this field. Scoring rules can be applied in both *ex ante* assessment and *ex post* evaluation (Friedman 1983; Gneiting & Raftery 2007; Savage 1971; Winkler 1996). Applying scoring rules to evaluate the accuracy of predictions is the main interest of this research in relation to this theory.

Consider a general standard to apply regarding whether a subjective probability prediction is a good prediction; Winkler and Murphy (1968) proposed two concepts in measuring the “goodness” or the quality of subjective probability predictions. The first standard is a *normative* standard examining whether predictions are coherent with people’s true expected probabilities (if actual probabilities are not available) or actual probabilities (if they are available). The second standard is *substantive learning*, measuring the knowledge of a person applied in making predictions. Just as statistical models can be judged on both prediction accuracy and statistical goodness-of-fit,

subjective probability predictions can be judged on their prediction accuracy and knowledge. Prediction accuracy can be directly observed by comparing predicted probabilities with actual probabilities and applying scoring rules. However, quantifying the knowledge used by a person in making predictions is not as simple. This is because knowledge in one's mind neither can be directly observed nor easily elicited because even learners themselves may find it difficult to describe explicitly what knowledge is being used in making the predictions. People cannot describe tacit knowledge or simply are not aware of the particular knowledge being used. This requires a way to develop a model that can approximate their knowledge. This chapter leaves the discussion about understanding substantive learning for Section 3.3. Section 3.2 only focuses on evaluating the accuracy of using scoring rules.

Knowing the two standards, before thinking about evaluation and estimation methods, the first problem is how to best design an elicitation task to capture more information from people. In other words, the type of task people are asked to perform to make predictions should be adequate to allow the measurement of accuracy and knowledge. Tasks asking people to provide full probability predictions on discrete and mutually exclusive events or options can give more information than tasks simply asking people to predict one choice or one probability for a single event. This is because “probability forecasts can explicitly recognize and quantify uncertainty ... probability forecasts are more valuable to decision makers than categorical forecasts” (Winkler 1996, p. 2). Full probability prediction tasks for discrete and exclusive events also match with the fundamental representation given by Savage to evaluate category-based probability predictions (Friedman 1983; Gneiting & Raftery 2007; Savage 1971; Winkler 1996). This is discussed in Section 3.2.2 regarding scoring rules.

The idea of asking people to provide full probability predictions for tasks is also supported by researchers working in the areas of expert judgement and discrete choice models. For example, O'Hagan et al. (2006) gave detailed descriptions of a variety of methods and processes for eliciting expert subjective probabilities. One of the methods they proposed is to use a “constant-sum scale” that asks subjects to allocate percentages to different choices in given choice scenarios. This method was also proposed by Louviere and Woodworth (1983). This method is equivalent to asking people to predict probabilities for categories, the main difference being the use of percentages instead of probabilities as the former is an easier concept for people to grasp. This method is also one of the methods discussed by Louviere and Islam (2008) in comparing importance weights and measures. From both mathematical and decision-making perspectives, it is not hard to see why full probability prediction is more informative in providing evidence of substantive learning because it is much harder for people to answer without the support of knowledge. Assume there are four mutually exclusive events for people to predict the probabilities that each will occur. The simple task is to predict which event will occur. Whether predicted correctly or incorrectly, information gained by this method is limited. To evaluate how a person is performing overall requires researchers to accumulate predictions for many tasks to establish the probabilities of correct predictions. On the other hand, if the person is asked to predict probabilities of all four events such that they sum to 1, it is possible to draw an inference about this person's performance based on the closeness of the predictions to actual probabilities of occurrence related to even one task. In the simple category prediction task, a correct prediction can be the result of a random pick or the application of other strategies. On the other hand, making correct or even close predictions for the four probabilities is impossible without good knowledge of the four events being predicted. According to O'Hagan and others, in expert elicitation tasks, eliciting multinomial probability distribution is much harder than eliciting single binary probability. In return, information gained from the former elicitation is also much greater (O'Hagan et al.

2006). In this research, the prediction task of full probability distribution of all options can provide data to test both prediction accuracy and substantive learning.

### **3.2.2 Using Scoring Rules to Evaluate Accuracy of Probability Predictions**

This section addresses the key issue of how to evaluate prediction accuracy using scoring rules. Using scoring rules to evaluate prediction accuracy is a well-accepted and highly developed method. Research on scoring rules started early in the 1950s under the framework of subjective probability theory. For example, Brier (1950) proposed the Brier score used in weather forecasts to test the accuracy of predictions of mutually exclusive categorical events. Research on scoring rules started to evolve from the 1960s. Many forms of scoring rules, properties of scoring rules such as strict properness and effectiveness and applications in various fields have been studied (e.g. Friedman 1983; Hogarth 1975; Winkler 1996; Winkler & Murphy 1968;). Particular topics such as which scoring rule is the most appropriate one for a certain situation are still evident in recent research (e.g. Bickel 2007). Researchers from other disciplines such as computer science also are interested in scoring rules. For example, researchers studied the use of scoring rules in fields such as machine learning, forecasting systems and online algorithm learning (e.g. Dawid, Lauritzen & Parry 2012; Skouras & Dawid 1999; Vovk 2001). Two thorough and influential reviews were provided by Winkler (1996), and more recently by Gneiting and Raftery (2007) of the types of scoring rules and their properties and usage.

To define a scoring rule, Gneiting and Raftery (2007, p. 359) state that, “scoring rules assess the quality of probabilistic forecasts (predictions), by assigning a numerical score based on the predictive distribution and on the event or value that materializes”. While this definition may sound somewhat technical, a simpler expression is that scoring rules are scores generated as a result of comparing predicted and actual probabilities. Earlier literature, such as Winkler (1996), discusses scoring rules that focus on probabilities of discrete events with a finite number of probabilities.

Later literature, such as Gneiting and Raftery (2007), discusses probability distributions and densities and the concept of scoring rules used to predict continuous variables. The present research is only interested in using scoring rules to evaluate probability predictions of categorical variables, in particular a discrete number of decisions and choices in decision making, the context of this thesis.

Starting with Savage's popular representation characterising scoring rules for probability predictions for categorical variables, three popular scoring rules of quadratic, spherical and logarithmic form are discussed (Gneiting & Raftery 2007; Savage 1971; Winkler 1996). For a categorical variable containing  $M$  finite number of mutually exclusive options  $\{o_1, o_2 \dots o_m\}$ , the probability predictions for  $M$  are a probability vector  $P$  consisting of  $\{p_1, p_2 \dots p_m\}$ , and the actual probabilities for  $M$  are another probability vector  $R$  consisting of  $\{r_1, r_2 \dots r_m\}$ . To simplify the discussion, assuming there are only two options in  $M$ , the general form of the expected score for predicting the two options is (Winkler 1996):

$$E_p[S(r)] = p_1 S r_1 + (1 - p_1) S r_2 \quad (3.1)$$

with  $S r_1$  as the score for  $r_1$  if  $o_1$  occurs, and  $S r_2$  as the score for  $r_2$  if  $o_2$  occurs. If there are  $i$  number of multiple options, the general form then becomes:

$$E_p[S(r)] = \sum_i p_i * S(r_i) \quad (3.2)$$

Applying the three common scoring rules (quadratic, spherical and logarithmic) to Equation 3.2, the general form of three scoring rules with  $i$  options are (Winkler 1996):

$$\text{Quadratic: } 1 - \sum_i (r_i - p_i)^2 \quad (3.3)$$

$$\text{Spherical: } \sum_i r_i p_i / (\sum_i p_i^2)^{1/2} \quad (3.4)$$

$$\text{Logarithmic: } \sum_i r_i * \log(p_i) \quad (3.5)$$

All three scoring rules have been mathematically proven to be “strictly proper”. A strictly proper scoring rule means a person who makes predictions can maximise this score relative to actual

probabilities, and the maximum is unique (Gneiting & Raftery 2007; Winkler 1996). There is consensus by all researchers in this field that this is the required property for considering a scoring rule. In other words, if a scoring rule cannot be maximised or the maximum is not unique and the true value, it cannot be considered as an appropriate scoring rule. The reason why is because it can neither be regarded appropriate for eliciting honest predictions nor be regarded appropriate for rewarding people who make better predictions.

A key question is which scoring rule is the most appropriate to use, but there is not a simple answer. In some cases various forms of quadratic rules such as the Brier score and various forms of spherical rules such as Hellinger distance (a pseudo-spherical scoring rule) are appropriate. By nature, quadratic rules are Euclidean distance measures, which are suitable when probability predictions and actual probabilities are continuous, not dichotomous numbers. Friedman (1983) and Nau (1985) discussed “effectiveness” of scoring rules, which is another property besides strict properness. Friedman (1983) defines this as the appropriate scoring rule function being monotonic and Nau (1985) defines it as functions that meet the transitivity principle. Put simply, measures such as the Brier score using squares or square roots of probabilities are appropriate if distance measures between probability vectors can be applied. In Chapter 5, a type of distance measure (Hellinger distance) will be further discussed and used in analysis.

Logarithmic scoring rules are considered appropriate when there are two or more options involved and when models with likelihood ratios associated with maximum likelihood estimation are involved (Gneiting & Raftery 2007; Winkler 1996; Winkler & Murphy 1968). Winkler and others suggest additional reasons as to why the logarithmic form may be more appropriate in evaluation (Winkler 1996). For example, in situations when actual probabilities are only ones and zeros, the way to calculate logarithmic based evaluation scores is completely *local*; that is, the evaluation of

the subjective probability for any particular choice does not rely on probabilities of other choices and so it does not rely on assumed probability distributions. Another advantage of logarithmic scoring rules is that they are *sensitive to distance*. For example, in situations where there are two predictions, the prediction closer to the actual probability will always have a higher logarithmic score even if the difference of the two predictions is small. More recently, Bickel (2007) suggested that the logarithm rule is the least affected in the *ex ante* stage by tasks using nonlinear functions or when the rank order of predicted options is involved. Bickel also suggested that the logarithm rule performs better than quadratic and spherical rules in post-prediction analysis when nonlinear utility functions and rank orders are involved.

In Tables 3.1 and 3.2, two examples are given to show how the three scoring rules work for Persons A and B, who provided predictions for three mutually exclusive options of  $\{o_1, o_2, o_3\}$ . In Table 3.1, it can be seen that the logarithm scoring rule concludes that a better prediction is made by Person A when  $o_1$  actually occurs (probability is 1) and the scoring rule becomes *local*, which means the final score is only determined by the actual and predicted probabilities of  $o_1$ . Quadratic and spherical scoring rules use all probabilities for evaluation therefore concluding that Person B performed better. If we only consider the actual event  $o_1$ , then Person A performs better because of the higher probability predicted. In Table 3.2, when actual probabilities are not dichotomous, the outcomes of all three scoring rules conclude that Person B is a better performer, which is evident by direct inspection of the closeness of the predictions. In the case of Table 3.1, it can be said that the logarithmic (L) score is appropriate because predicted probabilities for the options did not occur (i.e. probability = 0), which has little meaning and can be ignored in evaluation. In the case of Table 3.2, when full probabilities are used, all three scoring rules give rise to clearer distinctions. Both quadratic (Q) and spherical (S) rules also show clear distinctions as to better



performance. In this case, it is acceptable to select any of the three scoring rules. These two examples are modified from an example given in Winkler and Murphy (1968).

Table 3.1 Examples of three “scoring rules” when actual probabilities are 1 or 0

	$o_1$	$o_2$	$o_3$	$Q(r,p)$	$S(r,p)$	$L(r,p)$
Person A	0.35	0.6	0.05	0.215	0.503	-1.050
Person B	0.3	0.35	0.35	0.265	0.518	-1.204
$r_1, r_2, r_3$ as correct predictions	1	0	0			
Better performance				Person B	Person B	Person A
Forms of Scoring Rules				$1 - \sum_i (r_i - p_i)^2$	$\sum_i r_i p_i / (\sum_i p_i^2)^{1/2}$	$\sum_i r_i \log(p_i)$

Table 3.2 Examples of three “scoring rules” when actual probabilities are not 1 or 0

	$o_1$	$o_2$	$o_3$	$Q(r,p)$	$S(r,p)$	$L(r,p)$
Person A	0.35	0.6	0.05	0.860	0.503	-1.375
Person B	0.3	0.35	0.35	0.965	0.566	-1.119
$r_1, r_2, r_3$ as correct predictions	0.45	0.3	0.25			
Better performance				Person B	Person B	Person B
Forms of Scoring Rules				$1 - \sum_i (r_i - p_i)^2$	$\sum_i r_i p_i / (\sum_i p_i^2)^{1/2}$	$\sum_i r_i \log(p_i)$

There are many forms under each of the scoring rules. For the common forms discussed above, the range for quadratic rule is between -1 and 1, the range for spherical scoring rule is between 0 and 1 and logarithmic scores are unbounded on the lower end with highest end as 0, just as the log-likelihood produced by maximum likelihood estimators. Besides the clear connection with log-likelihood in models using maximum likelihood methods, logarithmic scoring rules also relate to entropy minimisation. As pointed out by Gneiting and Raftery (2007), logarithmic scoring rules have a direct corresponding connection with Negative Shannon Entropy and are associated with Relative Entropy (also called Kullback-Leibler Divergence or KL Divergence). It is a preferred scoring rule in several disciplines, such as information technology, because both Negative Entropy and Relative Entropy are common measures of information loss.

For the purpose of this research, there is no need to further explain how scoring rules are derived mathematically. More advanced mathematics on this topic largely relate to asymptotic probability

distributions and do not relate to a finite number of probability predictions on a limited number of categorical options for single prediction incidence. In the four learning approaches proposed in Chapter 2, scoring rules are only related to Learning Approach Four in classification at the experimental stage (a logarithmic rule is used for classifying responses to classes). However, at the post-experiment analysis stage, scoring rules are used as the most important measure of the accuracy of probability predictions without involving attributes of options.

When attributes of choice options are used to see how much people understand them in making predictions, a learner model needs to be identified using these attributes as predictors and predicted probabilities as the dependent variable. This is the topic of Section 3.3.

### **3.3 Preparing Target Learning Models and Estimating Learner Model**

#### **3.3.1 People's Knowledge in Probability Predictions and Discrete Choice Models**

Section 3.2 discussed using scoring rules to evaluate probability predictions within the framework of subjective probability theory. This section discusses methods for estimating people's knowledge in making probability predictions about discrete options consisting of attributes. The objective of estimation is to understand the knowledge used in predicting probabilities, but the fundamental method and theory come from estimating people's preference for and choice of discrete options (e.g. Train 2009). Perhaps the only difference between the two is that in the former case, people are making inferences based on their knowledge of attributes; in the latter case, people are making choices based on their preferred attributes. In both cases, people need to consider the attributes of the options.

As mentioned earlier in discussing scoring rules, this research is only interested in discrete and mutually exclusive categorical options that have wide implications in judgement and decision

making in marketing and other fields. In thinking of the type of models applied in modelling discrete categories with probability predictions as outcomes, the commonly acknowledged approach is to use discrete choice modelling (DCM) methods such as Multinomial Logit (MNL) or other types of DCM models (Louviere, Hensher & Swait 2000; Train 2009). Although DCM is not often applied in modelling people's probability predictions of choice options but their own preferential choices, these models can provide a clearer understanding of people's preferences for the attributes. In most DCM studies, the response data is stated preference (SP) data. In this thesis, the response data is people's predictions. There are no real theoretical differences from a choice modelling standpoint. A model that represents people's knowledge in making predictions requires estimation of attribute parameters from prediction responses. Similarly, a model that represents people's preferences and choices would be reflected in the same parameter estimates derived from choice responses.

### **3.3.2 General Framework to Model Choices**

The section starts with a short introduction to Random Utility Theory, the underlying theory for choice modelling. Instead of discussing many types of models, this thesis focuses on two classes of discrete choice models relating to the learning approaches discussed in Chapter 2. The first class is an aggregated model that can be used as the single model for training (Learning Approaches One, Two and Three), such as commonly applied MNL model. The second class takes into account differences between people by developing separate parameter estimates for classes or segments of people. This type of model can be used for categorisation learning (Learning Approach Four). Besides being used as target models in preparing for the learning experiment, DCM are also used in the learning experiment (Stage 2 of experiment). For example, in Learning Approach Three, a real-time MNL model using a weighted least squares method is applied to represent the learner model in feedback in the experiment. In the analysis stage of hypothesis

testing, individual MNL models are used to generate an individual attribute parameter matrix for further testing people's model learning experience under each experimental condition.

DCM is a common approach to model choices. It targets situations where people choose from a discrete number of choice options (Train 2009). In marketing, DCM is used to study consumer preferences and choices of products and services, which are common types of decisions that consumers make. Random Utility Theory (RUT) is a widely accepted theoretical framework to study such problems (e.g. Ben-Akiva & Lerman 1985; Louviere, Hensher & Swait 2000; Manski 1977; McFadden 1974; Train 2009; Yellot 1977). Under this common framework, choice behaviour, the estimated choice model, and stated preference (SP) or reveal preference (RP) data all can be studied and compared (Louviere, Hensher & Swait 2000). The concept of RUT can be explained by a simple idea. Assume an individual  $n$  in population  $N$  and a choice option  $i$  from a finite set of choice options  $C$ . The utility of choice  $i$  is associated with what the individual is expecting to gain by choosing  $i$ , represented by a random variable or a latent index  $U_{in}$ . This term consists of two components. The first component is called the “representative utility”, which is a systematic and observable component of the utility  $V_{in}$ . The second component is a random and unobservable error component  $\epsilon_{in}$ . The RUT utility function is therefore:

$$U_{in} = V_{in} + \epsilon_{in} \quad (3.6)$$

According to Ben-Akiva and Lerman (1985) and Train (2009), if option  $i$  is chosen over option  $j$  it means:

$$P_n(i|C) = Pr(V_{in} + \epsilon_{in} \geq V_{jn} + \epsilon_{jn}) \quad (3.7)$$

Stated formally, among a set of choice options  $C$ , the probability of an alternative  $i$  being chosen by individual  $n$  over any other choice alternative  $j$ , equals the probability of the utility of  $i$  being greater than the utility of  $j$ . This also implies that the sum of the two components of utility for  $i$  is greater than the same utility for alternative  $j$ .

It is important to note that in the literature, no restrictions are imposed on which particular statistical model should be applied to model  $V_{in}$ . Different models also have different assumptions regarding  $\epsilon_{in}$ , which lead to different model forms. RUT remains the conceptual framework in thinking about and studying choices, and it provides a general framework for deriving and estimating different models.

To further specify the  $V_{in}$  one can use generalised regression forms to model choices under RUT. For example, Ben-Akiva and Lerman (1985) expressed this general form as:

$$V_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \beta_3 x_{in3} + \dots + \beta_k x_{ink} \quad (3.8)$$

In this form,  $k$  is the number of product attributes. Similarly, Louviere and Woodworth (1983) and Train (2009) also used similar forms:

$$V_i = \beta_{0i} + \sum_k \beta_{ki} * x_k \quad (3.9)$$

$$V_{in} = x_{in} * \beta + k_i \quad (3.10)$$

Ben-Akiva and Lerman (1985) pointed out that although it may be intuitively convenient to think about  $V_{in}$  as the mean for  $U_{in}$ , it is theoretically misleading. In Equations (3.9) and (3.10), the matrix  $x$  refers to explanatory variables or predictors. These variables can appear in many different forms, such as in linear or exponential forms, covering both quantitative and qualitative information (Louviere, Hensher & Swait 2000). These general forms to model  $V_{in}$  can be further specified into different models. Although the general forms often look as if they are linear models, functions for model  $V_{in}$  are not restricted to linear terms. Nonlinear and non-additive models can be accommodated (Ben-Akiva & Lerman 1985; Louviere & Woodworth 1983; Train 2009). As Ben-Akiva and Lerman (1985, p. 63) stated, people often do not realise that “linearity in parameters is not equivalent to linearity in the attributes”. Any transformation of choice related attributes can be valid to estimate parameters and various estimation methods also apply (e.g. maximum likelihood estimation, weighted least squares estimation and simulation).

With regard to the error component, many assumptions can be made about error distributions as well as ways to specify their relationships. Early econometric and psychometric literature discussed the case when an error is an independent and identically distributed (IID) random variable. The Multinomial Logit Model (MNL) follows this assumption strictly (Manski 1977; McFadden 1974; Train 2009). Other models have been developed since the 1970s. Train (2009) provides a detailed discussion of different models including Probit, Nested Logit, and Mixed Logit. These models vary by the assumptions and treatments of the error term. For example, MNL assumes the error distribution is an IID extreme value Type I or Gumbel distribution. Different assumptions about statistical properties of the error distribution lead to different models and estimation approaches.

To summarise, RUT is a conceptual framework used to study choice. Important theoretical concepts of RUT include: 1) the utility of choice is random and consists of two components: the representative component and error component; 2) different explanatory variables can be studied and analysed using different models; and 3) the random error component follows a particular distribution. Models differ largely due to different assumptions about the error terms and the distribution (or lack thereof) of the parameters.

### **3.3.3 Model Type One - Aggregated Choice Model**

MNL is the most widely applied model for analysing choices. The theoretical foundation of this model was established by McFadden (1974). Ben-Akiva and Lerman (1985) and Train (2009) provided detailed discussions of the specifications and properties of this model. MNL can be denoted as follows:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \quad (3.11)$$

In this form,  $C_n$  represents a set of discrete choice alternatives. If the number of choice alternatives is two, MNL is reduced to a binary logit model in which there are only two choice alternatives,  $i$  and  $j$ . In most situations, there are more than two choice alternatives.

There are a few key properties and assumptions that distinguish MNL from other RUT models. If these conditions are violated, then other models should be considered in estimating choice probabilities. First, the error term  $\epsilon_{in}$  as shown in the random utility model (form 3.6) is Extreme Value Type I or Gumbel distributed. For each choice alternative, the error term is IID. This distribution can be characterised by key parameters  $\mu$  and  $\eta$  with the mode equal to  $\eta$ , and mean equalling  $\eta + \lambda/\mu$  ( $\lambda \sim 0.577$ ). The variance is  $\pi^2/6\mu^2$  (Ben-Akiva & Lerman 1985; Evans, Hastings & Peacock 2000; Train 2009). It is important to note that the above properties of the error distribution for MNL are not directly measurable in empirical studies. The important value of these properties is that they are the starting point from which to derive the main expression of MNL (equation 3.11). Train (2009) and Ben-Akiva and Lerman (1985) provide detailed mathematical proofs for Equation 3.11.

The second property of MNL, namely, the Independence from Irrelevant Alternative (IIA) property, is important empirically (Hausman & McFadden 1984). The IIA property means behaviourally that the probability ratio or relative odds of any two alternatives  $i$  and  $k$  in any choice set are not influenced by adding or removing any other alternatives from that choice set. If a new alternative is added to the set it will not influence the ratio of choice probabilities of the two existing alternatives. According to Train (2009, p. 49), the IIA expression can be directly derived from the probability function of MNL:

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}} = \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni} - V_{nk}} \quad (3.12)$$

If IIA is violated behaviourally (identified by available test methods) MNL does not hold and alternative models need to be considered.

As mentioned earlier, MNL is the most commonly applied model in studying consumer choices. This is largely due to its ease of estimation and widely available software. However, it does have major limitations. One such limitation is that it is estimated as an aggregated model and does not take into account consumers' preference heterogeneity (Train 2009). There are more robust but also more complex models that take into account preference heterogeneity, nested structure, latent variables, latent classes and unequal error variances among sampled consumers (e.g. Ben-Akiva et al. 2002; Fiebig et al. 2010; Revelt & Train 1998; Walker 2001).

In this thesis, the main objective is to test whether people can learn from a model to help their subjective probability predictions, which suggests starting with a simpler model that is more intuitive to learn. In this case, MNL has an advantage over more complex models because it can be easily explained and interpreted. It is highly likely that a simpler model like MNL will increase the likelihood of learners gaining knowledge quickly to improve predictions. Therefore, a simple MNL model is considered as the first candidate to serve as a target model for learning in the learning experiment (Learning Approaches One, Two and Three), even though this implies preference heterogeneity is neglected for learning.

### **3.3.4 Model Type Two – Choice Models with Latent Classes**

The last section discussed the basic choice model. This section discusses choice models with several latent classes, each showing its unique choice behaviour. Each class represents a group of people with more homogeneous choice behaviour that differs from the other classes. From a model learning perspective, learning a model with several latent classes should increase the level of difficulty for learners. Learning choices for one group of consumers may only require focusing



on explaining the effects of attributes in making choices. Learning the choice behaviour of several latent classes may require integrating and differentiating knowledge about different classes and may also involve different learning techniques such as categorisation learning. In this section, the objective is to examine how to estimate models with latent classes.

As discussed in the last section, aggregated MNL does not model preference heterogeneity. Researchers have developed different models to account for heterogeneity. One method is to extend fixed parameters of attributes to parameter vectors following some continuous distributions. These models compute continuously distributed parameter vectors and provide means and standard deviations of the parameters. The most widely known model of this kind is the Mixed Logit model (Revelt & Train 1998; Train 2009). Unlike aggregated MNL, mixed models assuming parameter distributions can capture individual differences. However, from a model learning perspective, continuous distributions on parameters should be difficult for humans to interpret and predict. Imagine the burden a learner may encounter if the learning task is to learn parameters to make choice predictions, but each parameter varies over a distribution. No previous research has demonstrated that people can learn the fixed parameters of an aggregated choice model to make probability predictions. Thus it should be even more difficult and also unclear as to whether people could understand more complex parameter distributions. Thus, it is reasonable to believe that such tasks are much more difficult than learning fixed parameters in aggregated models.

On the other hand, learners should be able to better understand intuitively if a complex model can be separated into several simple models to represent the choice behaviour of several different classes. Within each class people are considered homogeneous in their choice behaviour and their attribute parameters are fixed. For example, consumers buying electronic goods can prefer a

cheaper price or a famous brand. Considering these two attributes as the main ways that two classes of consumers differ, if learners can be told about them, it should be easier for learners to learn and compare the two classes and make predictions accordingly. For example, people can predict the former class to choose products with cheaper prices whilst predicting the latter class to choose famous brands.

Many methods in statistics can be used to generate classes of subjects based on certain characteristics, such as cluster analysis (e.g. Field 2005). In the DCM literature, a widely cited model is the Latent Class Model (LCM) (e.g. Kamakura & Russell 1989; Louviere, Hensher & Swait 2000; Train 2009; Wedel & Kamakura 2000). In the econometric and statistics literature, the LCM is also studied, though with greater focus on estimation and less on behavioural interpretation as in the DCM literature (e.g. Agresti 2002; Greene 2003). Apart from LCM researchers in choice models, economics and statistical learning also are developing new ways to classify and segment people based on preference heterogeneity. One of these methods is archetypal analysis initially developed by Cutler and Breiman (1994), which has been further investigated in choice modelling and statistical learning contexts (e.g. Carson, Bordes & Pailthorpe 1997; Eugster & Leisch 2009; Hastie, Tibshirani & Friedman 2009; Li et al. 2003). This section briefly reviews these two methods as main candidates to gain a model with latent classes in this research.

Please note, as will be discussed in Section 5.2.2 of Chapter 5, archetypal analysis was found to generate consumer classes with better statistical fits and clearer explanations than LCM. Therefore, only results from archetypal analysis will be presented in Chapter 5. However, for discussion purposes, LCM is briefly covered here.

### Archetypal Analysis

The original article on archetypal analysis by Cutler and Breiman (1994) proposed an approach to model data points as convex combinations of a few extreme points lying on the boundary of a convex hull. These points are named “archetypes” or “pure” types. The central point of this idea is that these extreme value points have more influence on other data points; instead of using common factor or cluster methods, data points can be classified according to probabilities that fall into these “pure” types (Eugster & Leisch 2009). In statistical learning literature, archetypal analysis is used as a factorisation or classification method in unsupervised learning (Hastie, Tibshirani & Friedman 2009). The main difference in using this method over normal statistical clustering analysis is that each data point has different probabilities of falling into each archetype instead of falling into a cluster completely.

Cutler and Breiman (1994) did not suggest how this approach can be applied in marketing or consumer choice studies. However, using the idea of archetypes in consumer choice problems is not difficult to understand. Li et al. (2003) shows an application of this approach in modelling consumer choices. In thinking about archetypes, it is natural to assume these pure types are those extreme individuals who have the most unambiguous preference rules as to how they make choices. For example, extreme types may be those who are clearly cost driven or brand driven. These extreme consumers are the archetypes that fall on the boundary of the convex hull to define the whole space of consumers. All other consumers fall within the convex hull defined by these extreme individuals and can be identified by their combinations of probabilities relating to these archetypes. In a geometrical form, Figure 3.1 demonstrates an example of a convex hull defined by three archetypes on the boundary with all data points falling within the convex hull.

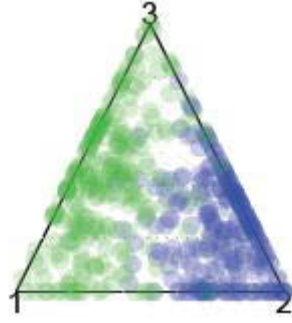


Figure 3.1 All data points fall into a convex hull defined by three archetypes (This example is from Eugster & Leisch, 2009, p. 18)

In the context of consumer choice, a consumer choice data set  $X$  (e.g. a “most preferred” choice data set) contains  $n$  consumers and  $m$  attributes. Data  $X$  is an  $n \times m$  matrix. Archetypal analysis assumes that given  $k$  number of archetypes, data can be modelled with two matrices  $\alpha$  ( $n \times k$  matrix representing coefficients of archetypes) and  $\beta$  ( $m \times k$  matrix representing coefficients of the data set). The aim of the analysis is to identify convex combinations of the data set  $Z$  ( $m \times k$ ) which is equal to  $X^T \beta$  such that the residual sum of squares (RSS) is minimised. This is shown in the following form of a matrix norm  $\| \cdot \|_2$ :

$$RSS = \|X - \alpha Z^T\|_2 \text{ with } Z = X^T \beta \quad (3.13)$$

There are two constraints applied in this analysis, one for the  $\alpha$  matrix and one for the  $\beta$  matrix.

They are:

$$\sum_{j=1}^k \alpha_{ij} = 1 \text{ with } \alpha_{ij} \geq 0 \text{ and } i = 1, \dots, n \quad (3.14)$$

and

$$\sum_{i=1}^n \beta_{ji} = 1 \text{ with } \beta_{ji} \geq 0 \text{ and } j = 1, \dots, k \quad (3.15)$$

In other words, these constraints imply that the sum of each individual's archetype coefficients

( $\alpha$  parameters) is equal to 1 with each probability equal to or greater than 0; the sum of all coefficients of the data set is also equal to 1 with each parameter greater than or equal to 0 (Eugster & Leisch 2009).

Once the probabilities of each individual consumer falling into all archetypes (defined by extreme consumers) are known and each archetype's attribute coefficient are known, the model can be estimated as several MNL models, with one MNL per archetype. That is, one can use “all or nothing” assignment to assign people into the archetypes, or one can use the  $\alpha$  coefficients as weights to estimate MNL models for each archetype.

#### Latent Class Model (LCM)

Louviere, Hensher and Swait (2000) give a clear description of how to model consumer choices using an LCM. Assume there are a discrete number of classes of consumers and consumers are homogeneous in each class; these classes are adequate to describe the joint discrete density of attribute parameters. In describing this model, two types of differences are involved; parameter differences in  $\beta$  vectors and scale differences in the  $\lambda$  vector. Scale differences represent differences of error variances among classes. Together, a member  $q$  of a latent class  $s$  in choosing choice alternative  $i$  can be denoted as:

$$U_{iq|s} = \lambda_s \alpha_{i|s} + \lambda_s \beta_s X_{iq} + \epsilon_{iq|s} \quad (3.16)$$

where  $U_{iq|s}$  is the utility of individual  $q$  of latent class  $s$  in choosing alternative  $i$ ;  $\lambda_s \alpha_{i|s}$  is the alternative-specific parameter of alternative  $i$  adjusted by a scale parameter  $\lambda_s$  for class  $s$ ;  $\lambda_s \beta_s X_{iq}$  is the attribute matrix  $X$  for alternative  $i$  multiplied by attribute parameters  $\beta$  for the class and further adjusted by the class scale parameter  $\lambda$ ; and  $\epsilon_{iq|s}$  is conditionally the IID extreme type I error within class  $s$ .

This model assumes each class model is a separate MNL following MNL assumptions. The probability of alternative  $i$  in class  $s$  being chosen over all alternatives  $j$  is shown as a standard MNL model form with the exception of class specific attribute parameters and the scale parameter:

$$P_{iq|s} = \frac{e^{\lambda_s \beta_s X_{iq}}}{\sum_{j \in C_q} e^{\lambda_s \beta_s X_{jq}}} \quad (3.17)$$

This means that, within each latent class  $s$ , the probability model is an independent MNL model with its own parameters and scale factor. Within each class, the probability of choosing alternative  $i$  is the product of two probabilities, one probability being the probability of alternative  $i$  being chosen, and the second probability being the probability of any individual falling into this class. The unconditional probability of  $i$  being chosen across all latent classes is therefore:

$$P_{iq} = \sum_{s=1}^S P_{iq|s} W_{qs} \quad (3.18)$$

In this form,  $W_{qs}$  is the probability that a person belongs to class  $s$ .

As noted in Louviere, Hensher and Swait (2000), a common approach is to constrain the scale parameter  $\lambda$  for one class to 1, and identify scale parameters for other classes relative to 1. Once this is done, parameters  $\beta$  can be estimated with the adjusted scale parameters. Since only utility matters, utility differences between classes in making choices are not affected by the value to which the first class scale parameter is constrained. Estimating  $\beta$  is more direct and differences of  $\beta$  parameters across latent classes are easier to interpret.

#### Comparing Class Approaches Statistically and Intuitively

In this research, what learners need to learn are the similarities and dissimilarities of classes to make predictions about their choices. Before deciding whether a class model under one method can be used as a target model for the learning experiment, it needs to be compared with other models. This thesis compares archetypal analysis and LCM in terms of statistical fit and model interpretation before deciding to adopt models from archetypal analysis.

The DCM literature commonly suggests the use of statistical criteria such as log-likelihood and Akaike Information Criterion (AIC) to compare and assess models (e.g. Ben-Akiva & Lerman 1985; Louviere, Hensher & Swait 2000; Train 2009). These criteria are objective and can be used to judge models statistically. However, fuzzier criteria also can be used. For example, in this thesis, to find a model to serve as a target model for learning, a more parsimonious model is preferred because it has a major benefit for intuitive learning. If learners can learn from a less complex model with fewer relationships, better performance should be expected. Another criterion is particularly related to the selection of classes. Regardless of which modelling approach is used, the behaviour of different classes should be clear and identifiably different. The clearer the definitions of classes, the easier it is for learners to understand and compare.

### **3.3.5 Estimating Learner Model for Real-Time Feedback**

Sections 3.3.3 and 3.3.4 discussed two types of target models for learning. In a learning experiment when learners are learning a target model, it is reasonable to believe that they will make better predictions if they better understand the target model. Better understanding of the target model can be considered as recognising the differences between their own models and the target model and being able to adjust for errors. In the Intelligent Tutoring System literature, a learner model or “student model” as it is often called, is the key component indicating what learners believe relative to the target knowledge/model (Woolf 2009). This section discusses the method for estimating the learner model in the context of this research.

There are two criteria that a learner model should aim to achieve in this research. First, it needs to be an individual and dynamically generated model during the experiment so each learner can receive immediate feedback before new tasks. This also means that this model may be an approximate model only due to limited data and the need for fast estimation methods to avoid delay. However, in post-experiment analysis, more robust but slower estimation methods can be

used to provide better statistical fits. Second, the learner model should be a valid and meaningful model to be compared with the target model. These requirements are essential in developing and computing each learner's own learner model based on his/her training data.

Models consistent with RUT are often computed using MLE, or a combination of MLE and simulation. Train (2009) provides detailed descriptions of MLE estimation procedures. MLE requires complex and iterative numerical procedures using algorithms such as the Newton-Raphson method, and there is no guarantee that convergence can be reached with limited training data within a limited time. Each iteration brings the estimation closer to convergence but the total number of iterations required to generate the best model will vary depending on the data. Doing this estimation dynamically during an experiment is computationally difficult, and the estimation time and whether the model will converge are uncertain. Therefore, it is not an ideal method for estimation during an experiment, but should be excellent for post-experiment analysis.

In widely cited research by Louviere and Woodworth (1983) and other statistical literature on categorical analysis (e.g. Agresti 2002; Greene 2003), Weighted Least Squares (WLS) was used as an alternative to MLE. WLS is a good alternative because it meets the two requirements discussed earlier. The computation simplicity of WLS allows one to produce estimates during an experiment. The general form of WLS is simple: to estimate model parameters relating to predictors ( $x_1$  to  $x_n$ ), the form of WLS is simply an extension of ordinary least square estimation method with a weight vector  $w$ :

$$\hat{\beta} = (X'WX)^{-1}X'WY \quad (3.19)$$

As stated in both Agresti (2002) and Greene (2003), weights applied in WLS are flexible and can be specified differently. In general, applications dealing with heteroscedasticity to account for variance in the error component of a model, weights are often specified as the inverse of the error



variance. In general regression models when the error component is assumed to be a normal distribution with variance of  $\sigma^2$ , weight is often specified as  $1/\sigma^2$ . In models dealing with the heteroscedasticity problem when variances are not constant and there is a scale  $\omega$  in the error variance, weights applied in WLS can be specified as  $1/\omega$  (Greene 2003). In other cases, weights can be specified as other relevant variables to improve ordinary least squares estimation. It is easy to see that computation for WLS is much simpler than models using MLE because estimation is a one step process without further iterations, and the estimation of parameter vectors can be solved directly in matrix algebra.

WLS is consistent with other random utility models asymptotically, therefore it can be applied as an alternative to MNL using an MLE approach. As pointed out by Agresti (2002), both WLS and MLE estimators are asymptotically equivalent and both estimators belong to the class of “best asymptotically normal” (BAN) estimators. Indeed, maximum likelihood estimators often start with WLS generated estimators and consist of iterative use of WLS. Louviere and Woodworth (1983) provide explanations as to why and how to use WLS in estimating choice data collected in discrete choice model experiments. According to them, WLS can be considered consistent with other RUT models. Indeed, the WLS method in estimating choices amongst alternatives can be considered as an MNL using the least squares method. It differs from other MNL models using MLE in its approach.

Furthermore, in the particular WLS example provided by Louviere and Woodworth (1983), the dependent variable is the logarithmic form of the predicted probability. This is in agreement with how probability predictions should be transformed to calculate a performance score in the theory of subjective probability assessment discussed in Section 3.2. Predictors are vectors of product

attributes and attribute levels similar to any other choice models. Weights applied in the WLS are the inverse of the variance, and equivalent here to the probabilities of the alternative being chosen.

The form of the WLS is articulated below to show how it is applied in this experiment to develop an individualised learner model based on discussions above. For example, in a training session consisting of 16 trial tasks, each with four options, the 16 sets of full probability predictions are collected, each set containing four probabilities that sum to 1. In total, there are 64 probabilities that can be used for WLS estimation. The dependent variable is a  $64 \times 1$  vector with the log form of 64 probabilities ( $\ln(p_1) \dots \ln(p_{64})$ ). Predictor ( $X$ ) is a  $64 \times n$  matrix with 64 combinations of  $n$  product attributes or levels. Weights are probabilities on the diagonal in a diagonal matrix:

$$\begin{bmatrix} p_1 & & & \\ & p_2 & & \\ & & \ddots & \\ & & & p_{64} \end{bmatrix}_{64 \times 64}$$

Using Equation 3.19, estimations can be generated with parameter vector  $\hat{\beta}$  of  $n \times 1$  size and one parameter for each product attribute or attribute level. As a generalisation, this approach can be applied to any number of data points greater or less than 64 as long as the number of parameters is less than the size of the data.

Considering the results of this model, if predictions need to be generated for any particular alternative  $i$  over other alternatives, then taking the exponentials of  $\hat{\beta}X$ , the ratio of  $\exp(\hat{\beta}X_i)$  for alternative  $i$  over other alternatives ( $\sum_{j \in C} \exp(\hat{\beta}X_j)$ ) yields the probability of  $i$  being chosen over other alternatives. This is exactly equivalent to the form of the MNL model in Equation 3.11.

### 3.4 Classification

In this section, the term “classification” refers to classifying inputs such as subjective probability predictions to classes or groups known *a priori*, such as consumer groups identified from a

consumer study. This is close to a standard definition for classification in statistical learning and decision theories (Cherkassky & Mulier 2007). When a person is learning a single target model to make predictions, a learner model for this person at a particular learning stage can be estimated by the system and used to inform the differences with the target model, as discussed in Section 3.3.5. When a person is learning a target model representing a particular class different from several other classes, this approach of comparing the learner model with several target models of different classes becomes difficult. Nonetheless, it is possible for a person to compare their own learning model with several class models to establish which class model is closest to the learner model. This method is neither efficient nor intuitive. First, one needs to process a learner model and provide comparisons of this model with several class models in computation which can cause a delay in the system. Second, the learner needs to receive much more information than required, and a lot would be only indirectly relevant. Instead of this approach, if learners can be informed directly about which class they belong to, based on their responses, and whether they are learning the right class model, the approach would be simpler. This method involves mapping features in response data to one class, which is a typical classification problem (Cherkassky & Mulier 2007; Hastie, Tibshirani & Friedman 2009).

### **3.4.1 Selecting the Appropriate Classification Method**

Classification is a main area of interest in statistical learning and a core feature in areas such as data mining. Many different methods are developed for classification (Hastie & Tibshirani 1994; Hastie, Tibshirani & Friedman 2009; Lippmann 1994; Mitchell 1997). A general principle for choosing a classification method is to apply a simple and direct method, especially with known classes or known probabilities of classes. As Cherkassky and Mulier (2007) pointed out, the traditional thinking of using regression models or density estimations for classifications is flawed conceptually. Among typical statistical learning problems, classification is the simplest and most direct compared to more complex problems such as regression or density estimation. The main

principle in solving the statistical learning problem is: “do not solve a specified problem by indirectly solving a harder problem as an intermediate step” (Cherkassky & Mulier 2007, p. 342). Generally speaking, regardless of a large number of classification methods in the literature, there are three common classification approaches. Lippmann (1994) referred to them as the probability density function (PDF) classifier, the posterior probabilities classifier, and the boundary forming classifier. Figure 3.2 illustrates these three kinds of classifiers.

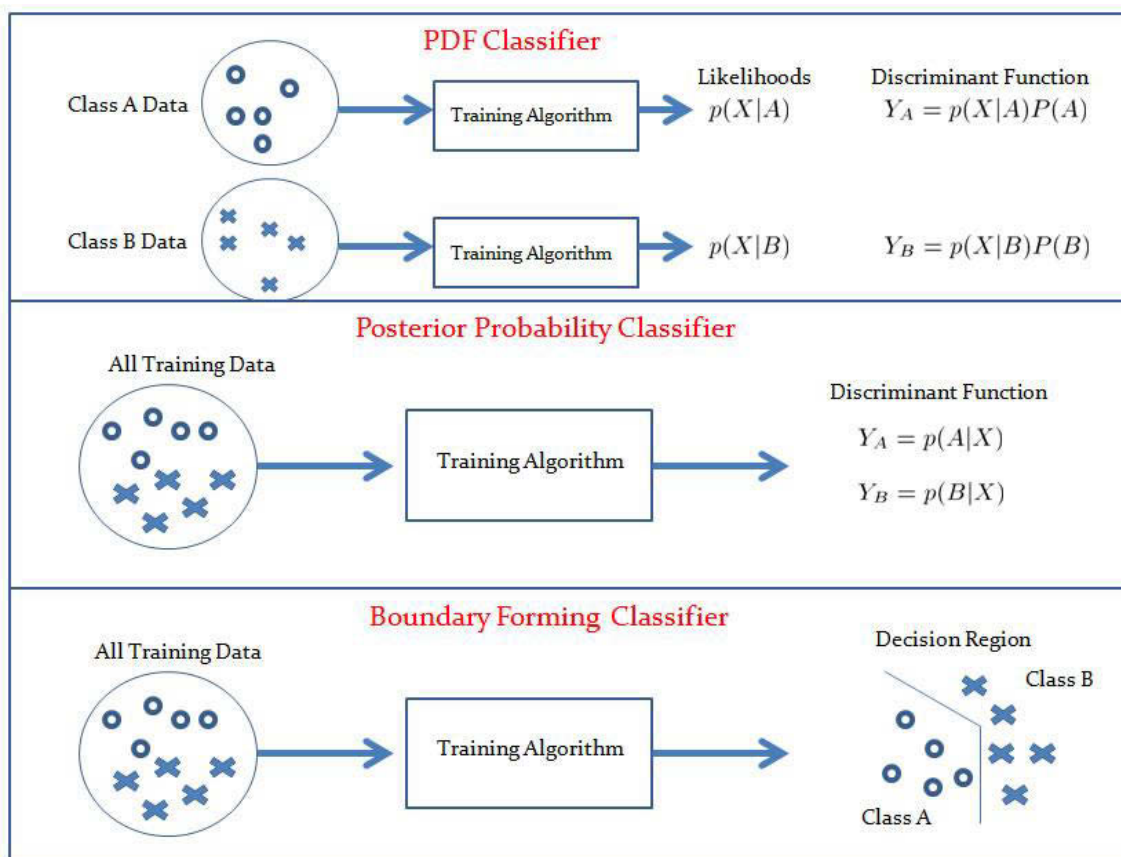


Figure 3.2 Three Classifiers (based on Figures 3 to 5 in Lippmann 1994, pp. 87–88)

Among the three classifiers, the PDF classifier requires some form of statistical analysis as input features of certain probability density functions in order to estimate the probabilities of responses belonging to different classes. As illustrated in Figure 3.2, in a simple two-class case, given training data set as  $X(x_1, x_2, \dots, x_k)$ , to estimate a likelihood function of  $X$  separately for class  $A$  and  $B$ , the

discriminant function for classification is the combination of the likelihood function with probabilities of classes in the sample. To estimate PDF, a large amount of data is required; hence, this approach is often applied in data mining where large amounts of data and input features are available. However, this is not the case, and not the purpose of the classification problem, in this research. There is a very limited amount of learning data available in the learning tasks. Second, as mentioned earlier, this method is more difficult and indirect compared to a simple classification problem. In the experiment for this research, classes are known *a priori* as a result of either the latent class model or the archetypal analysis using consumer choice data collected in Stage 1 of the consumer experiment. A simpler classification method should be possible.

A boundary forming classifier may sound convincing because it does not require a full statistical model as a PDF classifier but to map training data to pre-defined prototypes. The most common boundary forming method is the K-Nearest Neighbour (KNN) approach. The idea is, in a geometrical space, an existing class with a new instance (or response) is most similar to, and should also be closest, in distance geometrically. Therefore, it becomes a task to search for “neighbours” in geometrical space. In this classification task, obviously only one nearest neighbour is needed which is the target class that a new instance is closest to. There are several methods used in calculating geometrical distances, such as Euclidean distance, Euclidean squared, and city block. Without a specific reason, Euclidean distance is commonly used. If both training instances and classes can be defined by a feature vector with  $n$  features  $(a_1(x), a_2(x) \dots a_n(x))$ , then the Euclidean distance can be calculated as in Equation 3.20 (adapted from Mitchell 1997, p. 232) and the class with the smallest number is the class that the training instance belongs to:

$$d(x_{training}, x_{class}) = \sqrt{\sum_{r=1}^n (a_r(x_{training}) - a_r(x_{class}))^2} \quad (3.20)$$

Relating to this research, the feature vector should be attributes of the particular choice problem in considering training data. In post-experiment analysis of this research, a boundary method using a distance measure is used to calculate how close learner predicted probabilities are to true probabilities to identify which learning approach is more effective. This is also related to the discussion of scoring rules in Section 3.2. However, this is not the method chosen to be built into the training system to provide a real-time classifier and feedback for Learning Approach Four.

In the context of Learning Approach Four, thinking about the training data which can be used in classification, probability predictions are given by learners and the probabilities of classes (consumer groups) are known *a priori*. It is appropriate to consider a Bayesian classifier because it applies to prior and posterior probabilities directly.

### 3.4.2 Bayesian Classifier and Choosing Appropriate Likelihood Function

Given training data  $D$ , the posterior probability of a best class  $h$  (from all classes  $H$ ) according to Bayes' Theorem is:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (3.21)$$

$P(D|h)$  is the likelihood or conditional probability of data  $D$  given that  $h$  is true.  $P(h)$  is the prior probability of  $h$  in all classes  $H$ .  $P(D)$  is the prior probability of training data  $D$  being observed (Mitchell 1997).

Treating  $P(D)$  as a constant, the prior probability of any  $h$  in  $H$  can also be considered equal for all classes in learners' minds (this can be further reinforced if actual proportions of classes are indeed close). According to the maximum *a posterior* (MAP) approach, to identify the best class  $h$  for training data  $D$  is to maximise the likelihood or conditional probability  $P(D|h)$ , as denoted in Equation 3.22:

$$h_{ML} = \underset{h \in H}{argmax} P(D|h) \quad (3.22)$$

A Bayesian classifier applying the MAP approach without adding a loss function for misclassification error is called Naïve Bayesian classifier, and with a loss function added, is called an Optimal Bayesian Classifier in machine learning literature (Mitchell 1997). If a likelihood function for gaining conditional probability is unambiguously clear, a Naïve Bayesian classifier is simpler for computational purposes.

In earlier discussions in Section 3.2 on scoring rules, it is clear that scoring rules must be strictly proper to gain the accuracy of predictions between two probability distributions. In thinking about the problem in this thesis, the training data is the full probability prediction of discrete options summing to 1. Probabilities of options given in each training scenario are known for each class. The real problem underlying this classification is to calculate several scores using a chosen scoring rule and identify the best score. Depending on which scoring rule is chosen to select the closest class, the best score can be the minimum or the maximum score. As mentioned in Section 3.2, because scoring rules are strictly proper, the closest class is unambiguous because the optimal score is unique in any given scenario. In this case, each score calculated using predicted probabilities and each class' probabilities can be simplified as:

$$\sum_{k=1}^K r_k * \log p_k \quad (3.23)$$

In Equation 3.23,  $k$  signifies the number of options in a choice scenario for probability prediction,  $r$  is the known probabilities of options for a class and  $p$  is the probabilities predicted by learners. This score rule is also the log-likelihood function. Therefore, the class with the maximum score below zero is the target class.

The above classification approach is the way to classify in each separate training or choice scenario. If a summary needs to be given for multiple training scenarios to see how learners have performed over time in predicting a certain class, then a mean score can be produced using all scores for that particular class in all training scenarios. In some cases, a geometric mean can be used instead of an arithmetic mean to show the central tendency if scores are unbounded. The geometric mean of  $n$  number of scores is simply the  $n$ th root of the product of these scores.

In short, this section discussed several classification methods in machine and statistical learning. For this research, the most appropriate classifier is a Bayesian classifier using a logarithmic scoring rule as the likelihood function to calculate posterior probabilities.

### **3.5 Summary**

This section discussed theories and detailed methods for all evaluation, estimation and classification approaches which are used in supporting the test of learning approaches discussed in Chapter 2. Some of these approaches will be used to prepare the proposed learning experiment, such as the class models discussed in Section 3.3.4. Some approaches are used in both experiment and analysis stages, such as the use of scoring rules. Other approaches such as the idea of a discrete choice model is the foundation to develop a detailed method in testing model learning in analysing collected learning data. Together, these evaluations, estimation and classification approaches also construct the learner (student) model and evaluation agent components in the proposed intelligent tutoring system used in this research.

In Chapter 4, the discussion will be focused on the design and analysis of the actual empirical study in this research.



## Chapter 4 Design an Empirical Study

### 4.1 Overview

Chapter 2 proposed four learning approaches and discussed them from a theoretical perspective, and Chapter 3 considered methodologies of evaluation, estimation and classification to support these approaches in the experiment for this research. By integrating the foregoing discussions, this chapter discusses how the empirical study was designed and conducted to test these learning approaches and related hypotheses. The experiment was split into two stages. In Stage 1, an online consumer survey was conducted to provide data to develop target models used for learning. Stage 2 was the main learning experiment asking those who completed the Stage 1 survey to learn from the target models in the assigned learning approach. Learners were asked to make subjective probability predictions to particular scenarios in training tasks. Both stages of the study were conducted online without interviewers or tutors. In the Stage 2 learning experiment, an intelligent tutoring system incorporating information, techniques and feedback mechanisms was used to assist learning.

The reason for using the same respondents in both stages of the experiment is simple to explain. In Stage 1, the survey ensured that the respondents selected were interested and qualified to answer questions relating to a particular product category. Because of respondents' interests and relevance, they were more likely to be motivated to participate in learning activities. Besides, if all respondents completed both stages, a rich set of data could have been collected for each individual for analysis and hypotheses testing. To be more specific, the Stage 1 consumer survey collected respondents' own choices and their initial subjective predictions before learning the target models. The Stage 2 learning experiment collected predictions from respondents (or learners as they were called in Stage 2) during two training sessions consisting of 16 tasks each. In total, each learner who completed both the Stage 1 survey and Stage 2 experiment provided their own choices for one

session (Session 0 with 16 tasks) and predictions in three sessions (Sessions 0, 1 and 2, each with 16 tasks). These predictions are probabilities about how consumers on average would choose each option. To make it simpler for learners, probability predictions were asked as proportions. This is a common elicitation method in asking people to elicit probability distributions (O'Hagan et al. 2006).

This chapter discusses and reports the details of the data collection and experiment plan for each stage. The chapter commences by providing some background on the stated preference (SP) method and the experimental design method. This is followed by a discussion of the experiment plan for the Stage 1 consumer online survey. After Stage 1 is discussed, the chapter will discuss the experiment plan and procedure for the Stage 2 learning experiment.

For the convenience of readers, methodologies and procedures for analysis for both the Stage 1 consumer survey and Stage 2 learning experiment will be discussed in Chapter 5, before analysis results are presented and discussed.

## **4.2 Overview of Stated Preference and Experimental Design Methods**

The Stated Preference method (SP) is a common approach applied in studying consumer choice and is consistent with the RUT framework (Louviere, Hensher, Swait 2000). The SP method uses surveys controlled by experimental designs to collect stated preferences and choice data.

There are two main reasons to use the SP method in studying consumer choices. First, based on Lancaster's (1966) consumer theory, utilities of goods and services can be characterised by the properties of goods and services. Each unique property can be characterised by a number of levels. In DCM literature, the properties of goods and services are identified as "attribute" and "attribute level". Lancaster's theory established the foundation to allow the measurement of utility through

the preference of product attributes. Second, instead of using revealed preference (RP) data which contain unknown factors and noise, the SP method allows the use of experimental design to control the availability and allocation of attributes and their levels in preference elicitation tasks. This way, choices are made under well-defined choice scenarios to reduce statistical biases. These advantages of the SP data over RP data are well known (Ben-Akiva et al. 1999; Louviere, Hensher & Swait 2000; Swait & Andrews 2003). Another benefit of using the SP method is that some hypothetical products that do not exist in a real market can be studied. In an SP choice scenario, respondents need to make trade-offs based on what have been given so that demands for any potential new products can be observed.

The main advantages of the SP method also match the purpose of experimental design which is to effectively control and reduce estimation errors. Experimental designs allow comparisons of subjects' responses to different treatment conditions (Cox & Reid 2000). Not all treatment conditions need complex experimental designs. For example, if there is a single attribute with few levels to be studied, subjects can be randomly assigned to treatment conditions based on these levels. However, problems concerning consumer choices are rarely so simple. Usually combinations of two or more attributes and several attribute levels are involved. As best put by Louviere, Hensher and Swait (2000), "a design experiment is a way of manipulating attributes and their levels to permit testing of certain hypotheses of interest" (p. 84).

The field related to the design of experiment (DOE) is a highly specialised field. Providing a full review of this subject in this thesis is neither necessary nor possible. However, understanding the nature and implications of DOE is important for the purposes of modelling and hypothesis testing. As put by Bliemer and Rose (2010), while developing choice models is to identify and estimate an attribute parameter vector  $\theta$  using the collected attribute and their levels  $x$  and choice outcomes

$y$ , creating an experimental design is the inverse problem which is to determine how to gain an optimal  $x$  to maximise the efficiency to estimate parameter  $\theta$ .

In this research, a design is required to collect consumers' stated preferences for a chosen product category to develop target models for learning. As well, a second design is required for building training tasks when people are learning from target models and making probability predictions. In thinking about both designs, the focus is on modelling and learning the main effects of attributes. The reason for not considering attribute interactions in this research is simple; until people can demonstrate that they can learn the main effects of attributes from a target model and improve their probability predictions, adding interactivity terms can only make the design larger and introduce more difficulties during model learning and probability predictions. Therefore, the design for the Stage 1 consumer survey should allow the development of models of main effects only covering both aggregated and class models. On the other hand, the design for the Stage 2 learning experiment should also focus on main effects only, allowing the testing of how much understanding learners have regarding main effects.

A complete enumeration of all attributes and their levels is called full factorial design. This design has attractive statistical properties because all effects including main effects and all way interactions can be identified. However, this design is the largest design so it is unnecessary if only some effects are of interest (Louviere, Hensher & Swait 2000). Moreover the size of a full factorial design is often so large that it prohibits effective practice. If the purpose is to develop main effects models, full factorial designs are not a consideration. However, the key difference for many experimental design approaches is to identify an efficient or optimal fraction of the full factorial design.

Design strategies in selecting a fraction of the full factorial design can be based on anticipated models and statistical criteria. Several design strategies for generating fractional designs are widespread in literature and practice. For example, researchers can select orthogonal arrays (OA) as a starting design and apply certain replication methods to create an orthogonal fractional factorial design to eliminate attribute correlations in the design matrix (e.g. Louviere, Hensher & Swait 2000). There are also different optimal designs to satisfy certain optimal rules or statistical criteria without involving prior information of the model parameters (e.g. Huber & Zwerina 1996; Street & Burgess 2007). Recent studies in this field focus on using prior information of parameters to generate Bayesian efficient designs (e.g. Bliemer & Rose 2010; Bliemer & Rose 2011; Puckett & Rose 2010; Rose et al. 2008; Scarpa, Campbell & Hutchinson 2007). Most of these approaches use some starting design from an existing orthogonal array design library, or alternatively are generated using special design software.

Applying an orthogonal design is a common practice especially if the area of interest has not been previously studied or special software is lacking. This is because eliminating correlations between attributes is a desired feature for modelling. However, some have argued that using orthogonal arrays does not guarantee the removal of correlations between attributes because correlations between attributes do not simply come from attributes and levels ( $X$  matrix), but a combination of the underlying attribute parameters and attributes ( $\beta X$ ) (Puckett & Rose 2010). A widely cited paper by Huber and Zwerina (1996) proposed several desired features for an efficient design:

- 1) Level balance: all attributes and levels appear an equal number of times in all choice sets;
- 2) Orthogonality: any two levels of different attributes appear an equal number of times and are uncorrelated;
- 3) Minimum overlap: any attribute level should repeat as little as possible in each choice set;

- 4) Utility balance: all choice alternatives in each single choice set should be as equally attractive as possible to avoid a dominant alternative.

Some researchers disagree with one or more of these criteria. For example, a critique of “minimum overlap” is that it generates a design that can measure main effects but precludes measurement on interactions (Street & Burgess 2007). “Utility balance” is also not a desired feature because under utility balanced scenarios, all products are equally attractive, hence it will be impossible to observe trade-offs in people’s choices (Puckett & Rose 2010).

New ideas are continually being explored in this field. For example, Street and Burgess (2007) proposed a method to generate a D-Optimal design for generic choice alternatives. These designs cannot be used in alternative-specific SP studies when there are constants applicable to labelled choice alternatives (“alternative-specific constant”). This design uses an orthogonal array as a starting design for the first alternative and then generator vectors to create the full design; for example, the design for the second alternative is a fold-over design of the first alternative. The main optimality criterion it follows is to maximise the determinant of the information matrix for main effects and/or interactions. D-optimality is not the only statistical principle for optimal designs as there are other optimal principles which satisfy other desired statistical properties (Atkinson & Donev 1992).

These types of optimal designs have been criticised on the basis of empirical findings. As pointed out by Puckett and Rose (2010):

Whilst optimal designs can be a powerful tool in achieving statistically significant parameter estimates under small sample sizes, behavioural factors can outweigh statistical factors in determining an appropriate sample size ... one must ensure that stability in parameter

estimates has been reached before one can have confidence that the statistically significant parameter estimates obtained are also plausible estimates. (pp. 188–189)

A different type of D-efficient design was proposed. This design can be found when the determinant of a variance-covariance matrix of expected maximum likelihood estimator is minimised (Bliemer & Rose 2010; Rose et al. 2008).

The idea of applying prior information has become more popular in generating designs for SP studies. The benefit of applying prior information is that there should be “significant efficiency gains without loss of respondent efficiency” (Scarpa, Campbell & Hutchinson 2007, p. 617). As reported by Sándor and Wedel (2001), by applying prior parameter information in an SP study, standard errors of estimates can be reduced and expected predictive validity can be increased. In practice, it is noticeable that design software such as Ngene (<http://www.choice-metrics.com/features.html>) can generate Bayesian efficient designs using prior parameter information.

This researcher adopted a standard approach to create the main effects design for the Stage 1 consumer survey using a publicly available orthogonal array from a widely accepted online library of orthogonal arrays maintained by Sloane (<http://neilsloane.com/oadir>). This design is not as small as those efficient designs but it is a less arguable approach that may provide stability in parameter estimates. Stability in parameter estimates is important for developing reliable and intuitive target models for learners to comprehend. Based on some past research on the same product which will be discussed in Section 4.3, a list of the most important attributes and attribute levels were selected to develop the choice experiment.

Next, based on the modelling results from the Stage 1 analysis, the Stage 2 design was developed and modified. This design contains two smaller designs, each with half the set numbers of the Stage 1 design, giving training sets for two training sessions. The reason for being able to cut down the size of the design for each session is because minor correlations for some alternatives are allowed, reflecting the analysis results of Stage 1. The smaller design for each training session is a combination of two separate orthogonal arrays. These designs will be discussed further in relevant sections. The two designs used in Stage 1 and Stage 2 share some close connections which allow the possibility of combining data from two stages for analysis. More importantly, in the Stage 2 learning experiment, all learners complete identical training tasks but under different learning approaches. This allows not only direct comparisons of learning approaches, but also individual models and model comparisons.

## **4.3 Designing Stage 1 Consumer Survey**

### **4.3.1 Choosing the Product Category**

The product category chosen for this empirical study is long-haul, cross-country air travel in Australia. This researcher commissioned Pure Profile (one of the largest online panel suppliers in Australia) to invite their panel members living in Sydney to participate in this online consumer survey. Among all those who participated, 519 respondents were qualified for having past experience or future interest in travelling from Sydney to other Australian destinations by flying cross-country. To provide more background to these cross-country flights, each trip can take four or more hours from origin to destination. Some flights may also need to make an intermediate stop (i.e. is not a direct flight), which may take even longer.

A good reason for selecting this product category is that it is not a product category with a large number of providers and products, unlike some consumer goods, hence, it is likely to generate



simpler and more comprehensive target models which will be intuitive for learning. This is an advantage for this proof-of-concept study. Another reason for choosing this product category is that this category was studied by the Centre for the Study of Choice (CenSoC) at the University of Technology Sydney in 2010, using a small pilot study with 200 respondents. With CenSoC's permission, this researcher gained access to the collected data of that study. Analysis results of that study provided this researcher with a useful indication of the attributes that are important in people's choices. Although the previous study was conducted to test a different research problem, namely brand equity and pricing, the results of that study nonetheless provided useful information to refine the present study.

The reason for not using models from the previous study is simple. There have been some major changes in the market since the time that study was conducted. For example, competition is continuously driving fares lower. There were also several major incidents that might have changed the landscape of the domestic air travel market. For example, as widely reported, one of the four airlines in the previous study, Tiger Airways, was banned from operation for a period of time by the Civil Aviation Safety Authority (CASA) in 2011, due to safety issues. Even though it has resumed restricted operations with a small number of flights and limited routes, it is no longer considered a main competitor in this market. Other incidents such as the industrial action taken by Qantas to ground its entire fleet in October 2011 in response to failed negotiations with unions also caused major public debate and upset at the time. These factors make it necessary to re-investigate consumer choices in this market to have up-to-date consumer models which will eventually be used as target models for learning.

#### **4.3.2 Selecting Attributes**

In the above-mentioned study by CenSoC in 2010, an MNL model as shown in Table 4.1 was developed to show the effects of attributes on consumer choices. The results of the 2010 model

are shown in Table 4.1. Among all attributes, “fare” was the most important attribute contributing to the model followed by “airline” brands. In contrast, “in-flight alcohol” and “number of stops” contribute little to people’s choices. Therefore, it is safe to ignore and remove both attributes from the new study to make the new consumer choice model simpler and model learning easier. It is worth noting that Qantas was the most favoured airline in the previous CenSoc model with a positive coefficient more than twice the size of the next favoured airline, Virgin. This situation has changed dramatically in the present study based on the analysis results of the Stage 1 consumer survey which will be outlined and discussed in Chapter 5. In the new consumer model, Qantas and Virgin are almost identical in terms of consumer preference. This is suggestive of a negative impact resulting from Qantas’ industrial action or else increasing satisfaction with Virgin. In the new study, Tiger was removed from the list of airlines, so only Qantas, Virgin and Jetstar were included in the consumer survey as alternatives. In the consumer survey experiment, respondents were asked to select a cross-country flight offer from these three airlines, or select not flying at all as an indication of dissatisfaction with any of the offers.

Table 4.1 Model results of the study conducted by CenSoC in 2010

Conditional (fixed-effects) logistic		Number of obs	=	12800
		Wald chi2(11)	=	2806.75
		Prob > chi2	=	0
<b>Log likelihood = -3032.77</b>		<b>Pseudo R2</b>	<b>=</b>	<b>0.32</b>
<b>Most Preferred Options</b>	<i>Coef.</i>	<i>Std. Err.</i>	$\chi^2$	$P > \chi^2$
Qantas	0.46	0.04	12.29	<b>0.00</b>
Virgin	0.20	0.04	4.67	<b>0.00</b>
Jetstar	-0.17	0.04	-3.79	<b>0.00</b>
Tiger	-0.50			
\$450	1.35	0.04	35.85	<b>0.00</b>
\$550	0.44	0.04	10.94	<b>0.00</b>
\$650	-0.86	0.06	-15.26	<b>0.00</b>
\$750	-0.94			
5 hours	0.24	0.02	10.22	<b>0.00</b>
7 hours	-0.24			
Free ticket changes	0.18	0.02	7.53	<b>0.00</b>
Pay for ticket changes	-0.18			
In-flight food & beverages - Free	0.29	0.02	12.49	<b>0.00</b>

In-flight food & beverages - Not free	-0.29			
In-flight alcohol - Free	0.07	0.03	2.75	<b>0.01</b>
In-flight alcohol - Not free	-0.07			
No stop	0.08	0.23	3.39	<b>0.00</b>
1 stop	-0.08			

### 4.3.3 Attributes, Experimental Design and Choice Tasks for Stage 1 Survey

As discussed earlier, this researcher chose an orthogonal main effects (OMEPE) design suitable for the alternative-specific experiment in this consumer choice study. There are three alternatives Qantas, Virgin and Jetstar. Four attributes are common to all alternatives. They are “return airfare”, “flying time”, “booking change” and “in-flight food and beverages”. There are four levels for “return airfare”: \$400, \$460, \$520, and \$580. Decreased fares compared to those applied in the previous study by CenSoC reflect reduced fares overall in the current market. There are two levels for “flying time”: four hours and six hours. For “booking change”, instead of using “free change” or “pay for change” as the levels used in the previous study, it was considered more realistic to ask whether a booking change is allowed. This is more relevant in the market especially for the low fare budget flight market because a booking change is not allowed for many offers. There are two levels “free” and “not free” for the attribute “in-flight food and beverages”, the same as for the previous study.

An OMEPE design of 32 sets was chosen for this experiment. Respondents were randomly assigned to half of the design to complete 16 sets of choice tasks. This design is shown in Appendix 1. The main properties of this design are: 1) main effects of all attributes can be independently estimated either as generic attributes across all alternatives or as alternative specific attributes (they are orthogonal across and within alternatives); and 2) all attribute levels appear an equal number of times for each alternative and across all three alternatives (level balanced). Moreover, for both blocks (versions) of 16 sets, a level balance is also maintained within each block.

To ensure the sample size for this study was adequate, a validation check was effected in the Ngene software. The test informed that if using an efficient and orthogonal design of 16 sets, a minimum sample size of 55 is required to estimate all the main effects. Therefore, a sample size of over 500 respondents is more than adequate to identify and estimate attribute effects to give rise to a well-established target model, and further segment consumers into clusters of reasonable size to account for preference heterogeneity among respondents. Since this study uses the same respondents for both stages, having over 500 respondents for the Stage 1 study, and assuming a drop-out rate of up to 60%, realistically ensures having 200 or more learners for the Stage 2 study. Depending on the use of incentives and other factors, and a longitudinal study using the same sample, the response rate will no doubt vary over time.

The screenshot version of the questionnaire for the Stage 1 online consumer survey is shown in Appendix 3. An example choice task is shown in the following Figure 4.1 (a similar example task used in the Stage 2 experiment has been shown in Figure 1.2 in Chapter 1). For each respondent, besides asking for their “most preferred” and “least preferred” choices, a prediction task similar to that used in the Stage 2 learning experiment, was also included.

	Qantas	Virgin Australia	Jetstar
Return Airfare	\$580	\$520	\$400
Flying Time	6 hours	4 hours	6 hours
Change Booking	Allowed	Not allowed	Not allowed
In-Flight Food and Beverages	Not Free	Free	Not Free

Which option would you **MOST LIKELY** choose?

☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Which option would you **LEAST LIKELY** choose?

☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Assume there are 100 other people also answering this question, how many of them do you think would choose each of the following options?

*Please enter a number in every box and make sure the sum is 100. Assume all four boxes should have a number in it, excluding 0 and 100.*

people would fly Qantas  
 people would fly Virgin Australia  
 people would fly Jetstar  
 people would not fly

= 0

Figure 4.1 Example choice and prediction task in Stage 1 online consumer survey

There are three reasons for including the prediction task in the Stage 1 survey. First, respondents can become familiar with the forthcoming learning tasks used in the Stage 2 experiment. Second, one extra session of prediction responses (Session 0) are available which can be considered as prior subjective probability predictions before learners are exposed to target models. This is the learning session before Sessions 1 and 2 in the Stage 2 learning experiment. Third, extra prediction responses provide extra validation information to eliminate respondents who gave little thought to their answers in these tasks. For example, if a respondent chose identical responses on most and least preferred questions in all choice sets and gave inconsistent or conflict prediction responses, it is highly likely the respondent was not thinking when answering questions. With duration also recorded for each choice set, it is easier to eliminate these respondents. In this study, 34 of the 519 respondents for the Stage 1 survey were considered “bad” respondents and removed from the data analysis and the Stage 2 invitation. A combination of criteria as discussed above was used for data cleaning. As a result, analysis for the Stage 1 survey was conducted using a cleaned data file with 485 respondents.

This survey was conducted online using a randomly selected online panel list from Pure Profile. The demographics of selected respondents invited to participate, match population statistics from the Australian Bureau of Statistics (ABS) closely on key demographic indicators such as age and gender to ensure that a good representation of the Australian population was surveyed.

## **4.4 Summary of Stage 2 Training and Learning Experiment**

### **4.4.1 Overview**

Using data collected from the Stage 1 consumer survey, this researcher developed two types of models, an aggregated model describing total consumers, and separate models describing three different consumer groups (classes). These models were used as target models in the Stage 2 training and learning experiment. These models and four proposed learning approaches were constructed in an online intelligent tutoring system. This system played several roles. First, it provided learners with information and feedback required for each learning approach. Second, it performed evaluations, estimations and classifications using learners' prediction responses in real-time and provided personalised feedback. Third, a common set of training tasks was controlled by the system to show to learners at an appropriate time. This tutoring system, including training tasks, was built in the form of an online survey. Learners received information and feedback in real-time, following instructions specific to each learning approach that they were randomly assigned. A document with screenshots and explanatory notes describing the Stage 2 learning experiment is shown in Appendix 4.

Respondents who had completed the Stage 1 survey were invited to participate in the Stage 2 learning experiment. Every learner received 32 identical training tasks which were grouped into two sessions. Each task required predictions to be made with knowledge the learners had gained from target models under the particular environment designed for each learning approach. In each

training task, learners were asked to predict the consumers' most preferred flight offer and allocate 100% to four options to represent their predictions of consumer choices for all offers. Responses in the allocation task can be easily converted to predictions of probabilities of each offer being chosen. There are four options to choose from in each scenario. They are "fly Qantas", "fly Virgin Australia", "fly Jetstar" and "not fly". Predictions that learners made were compared with probabilities from the target models to generate feedback that learners received depending on the particular learning approach. For example, for Learning Approach Two, probabilities from target models were simply shown to learners after they made their predictions, but no other information was given.

To avoid confusing learners, the system simply refers to probabilities predicted by target models as "actual" choices made by consumers. This is reasonable considering predictions made by target models are the most accurate predictions of actual choices for any prediction task scenarios. It is clear from the Stage 1 analysis that the best performing respondent is still less accurate than target models in making predictions of consumer choices (this finding will be covered in detail in Chapter 5). This is not a surprise finding given the volume of evidence from research comparing model and intuitive predictions (e.g. Grove et al. 2000).

The objective of the Stage 2 experiment is to test whether learners can improve subjective probability predictions if they also improve their understanding of the relationships in the target models; for example, the relationship between return airfare and choice probabilities. The assumption is that learners can improve progressively on prediction performance through the process of learning target models, but the key objective for the experiment which is also the research problem is to find out which approach or approaches are more effective in improving learners in making predictions.

It is important that all learners are assigned randomly to any of the four learning approaches. The four learning approaches proposed in Chapter 2 form the four experimental conditions in which different information, or the same information in different forms, different feedback mechanisms and procedures, are used following the theoretical ideas behind each approach. All learners completed 32 identical training tasks broken down into two sessions. This way, results were directly compared. Assigning learners randomly to different learning approaches also helps reduce bias due to differences in learners' characteristics such as their understanding of the problem and socio-demographics, because each learner had an equal chance of being assigned to any experimental conditions. Results were unlikely to be largely influenced by exogenous variables such as differences in experience and interest. In other words, the procedures were aimed to isolate prediction performance from factors that were not a part of the learning approaches. The following section describes the experimental design approach and the four experimental conditions in turn.

#### **4.4.2 Experimental Design for Learner Experiment**

As discussed in Section 4.3.3, the experimental design for the Stage 1 survey is an OMEP design with 32 sets, organised in two blocks of 16 tasks (Appendix 1). Each respondent only completed one of the two blocks and answered other related questions (such as screening and socio-demographic questions). For the Stage 2 experiment, because the maximum requirement of the intelligent tutoring system is to estimate a model in real-time using individual prediction responses for the whole training session (Learning Approach Three), a complete design without blocking was desired. To estimate a reliable learner model using responses in training tasks from a session, all tasks should be completed. There were two sessions in this experiment so two independent designs were needed. Besides this requirement, the elapsed time for each session needed to be reasonable to maintain learner interest while they are learning remotely online without help and supervision from trainers in person. Therefore, having 32 tasks in a session and 64 tasks in total



sounds too long and exhausting for learners. It may cause learners to quit the experiment completely or complete the experiment on different days. The latter case introduces extraneous factors into the results such as long-term memory. Having too many tasks increases this possibility. Controlling each session at 16 tasks and the full survey at 32 tasks can avoid the danger of high incompleteness rate, multiple online sessions and other extraneous factors. As shown in the data collected, which will be discussed in Chapter 5, taking these precautions in designing the experiment was the right decision, because all learners were shown to complete the experiment in one run on the same day and the final completion rate was about 50% (240 learners after data cleaning achieved from 485 invitations). This was achieved without extra incentives given by Pure Profile other than normal compensation for a standard survey. The experimental design for the two sessions is shown in Appendix 2.

As shown in the screenshot version of the Stage 2 experiment in Appendix 4, there was some information inserted between the two training sessions to clearly separate them for learners. Information inserted ranged from a simple question on a single screen to feedback on multiple screens depending upon the experimental condition to which a learner was assigned.

In finalising the experimental design, and instead of selecting the larger orthogonal design of 32 sets used in the Stage 1 survey, a smaller design with 16 sets was found to be available in the public library of orthogonal arrays maintained by Sloane allowing attributes for up to two alternatives. It became possible to use this design to create new designs because the results of the Stage 1 survey showed there were no cross-effects in choices made for Qantas and Virgin offers, but minor cross-effects were shown in choices made for Jetstar offers. It implies that consumers choose Qantas and Virgin offers because their offers were considered satisfactory, but they can choose Jetstar offers not because Jetstar offers are satisfactory, but Qantas and Virgin offers are poor. This

finding suggests two smaller orthogonal arrays of 16 sets can be used instead of one larger orthogonal array of 32 sets allowing all three alternatives to be orthogonal. One orthogonal array can be used for Qantas and Virgin to ensure their attributes are orthogonal, and the other array ensures attributes for Jetstar are orthogonal within itself. By concatenating the two smaller designs, Qantas and Virgin are still orthogonal to each other, but some correlations are allowed with Jetstar, matching results of the Stage 1 analysis. The design for Session Two is just a variation of the Session One design, with some attribute columns reverse ordered.

#### 4.4.3 Four Experimental Conditions Matching Four Learning Approaches

Learners were randomly assigned to four experimental conditions with an equal quota target. Of all invited 485 respondents who completed the Stage 1 survey, 252 respondents were collected. There were exactly 63 learners kept for each experimental condition for analysis and hypothesis testing. As mentioned, the four experimental conditions matched the four learning approaches proposed in Chapter 2. Table 4.2 provides a summary of these four experimental conditions. Sections 4.4.3.1 to 4.4.3.4 briefly summarise each experimental condition. Screenshots of the whole experiment and each condition are included in Appendix 4.

Table 4.2 Four Experimental Conditions

Experimental Conditions	1	2	3	4
Learning Approach	One	Two	Three	Four
Training Sessions	2	2	2	2
Training Tasks	32 (16 per session)	32 (16 per session)	32 (16 per session)	32 (16 per session)
Self-Regulated Learning	Yes	No	No	No
Type of Feedback	No	Outcome	Cognitive & feed-forward	Cognitive & feed-forward
KR Approach in Feedback	Not Relevant	Not Relevant	Attributes stated explicitly	Relationships of consumer groups shown in chart/texts
Categorisation Learning	No	No	No	Yes
Feedback about the Stage 1	Yes	Yes	Yes	Yes
Feed-Forward Information before	General summary of Stage 1 predictions +	General summary of Stage 1 predictions	General summary of Stage 1 predictions +	General summary of Stage 1 predictions +

Session One Training Tasks	Self-regulated training using DSS		Attribute-based Comparison of learner model from the Stage 1 survey with target model	Details of three consumer groups including their spatial positions related to attributes, similarities, and dissimilarities
Prediction Tasks	Predicting all consumers	Predicting all consumers	Predicting all consumers	Predicting one consumer group
Real-time Analysis	No	No	Estimating Learner Model	Bayesian Classifier
Target Model for Training	Aggregated MNL	Aggregated MNL	Aggregated MNL	MNL model for each consumer group

#### 4.4.3.1 Experimental Condition 1 (EC1)

All learners in this condition received identical information and the tool so no individualised evaluation was required. Learners controlled the number of learning scenarios they wanted to go through in an online DSS. This DSS offers “what-if” scenarios so learners can observe outcome probabilities to the choice combinations they constructed. In this experiment condition, learners could determine how long they wanted to use the DSS before they started training tasks. Learners also had total control of scenarios they wanted to construct. A minimum duration of two minutes was required by the system during which the DSS was shown to learners the first time. Two minutes was not adequate for learning the target model, but it had the effect of bringing it to the attention of learners so as to avoid them accidentally skipping the training phase.

There were four steps in the procedure once learners passed the initial introduction. Step One allowed learners to have the first practice using the DSS. Step Two was for learners to complete the first session of 16 training tasks. Step Three was to allow learners to have the second, also the last, practice using the DSS. This time there was no constraint on the minimum time required. Learners could decide how long they needed to spend time training on the DSS. Step Four asked learners to complete the second session of 16 training tasks. There was no other feedback for learners during or after training tasks. Please note, the actual DSS used for training was shown in

Figure 1.1 in Section 1.2. It is also included in Appendix 4 as a part of the screenshot copy of the experiment. An example training task asking people to make a prediction was shown in Figure 1.2 in Section 1.2. This information will not be repeated here.

#### **4.4.3.2 Experimental Condition 2 (EC2)**

This condition aimed to replicate the traditional learning method in training people to make probability predictions in MCPL studies (e.g. Cooksey 1996). This approach offered the simplest form of feedback. After showing the summary of the Stage 1 predictions (common to all other experimental conditions), learners were asked to make predictions in 16 training tasks. They received correct answers after each task. After a short break with a separate single response question, learners continued with the second session with another 16 training tasks. Again, they received correct answers after completing each task. No other type of feedback was used in this experimental condition. Outcome feedback was common to all learners because training tasks were common to all learners. In this condition, no evaluation, estimation or classification was required. Learners needed to think about where to improve and which attributes had caused discrepancies in the prediction results. Outcome feedback was prepared beforehand based on predictions generated by the aggregated MNL model from the Stage 1 data analysis. The following Figure 4.2 is an example showing outcome feedback.

Task 1 (Tasks 1 to 16 in Session One):

	Qantas 	Virgin Australia 	Jetstar 
Return Airfare	\$580	\$580	\$400
Flying Time	4 hours	4 hours	4 hours
Change Booking	Not allowed	Not allowed	Allowed
In-Flight Food and Beverages	Free	Free	Not Free

Correct Answers vs. Your Predictions

Options	Correct Choices	Your Predictions
Fly Qantas	8	15
Fly Virgin Australia	8	10
Fly Jetstar	83	60
Not Fly	1	15



Figure 4.2 EC2 - an example of outcome feedback

#### 4.4.3.3 Experimental Condition 3 (EC3)

The key characteristic for this experimental condition was the use of attribute level statistical information from learner models in feedback. After each session of 16 tasks, prediction responses were used for model estimation. Using the WLS method as discussed in Chapter 3, the tutoring system estimated an individual model for each learner. This real-time learner model reflected how learners made predictions in the newly completed session. Each learner model was compared against the target model on all attributes one by one. The results of these comparisons were presented to learners in feedback they received. This type of feedback matched the definition of cognitive feedback, containing attribute level statistical and behavioural information specific to each learner. Each learner's model was estimated twice by the system, one after each session. Because the feedback after Session 2 was not followed by another session of training tasks, it is unknown what learners had learned from the second round of feedback. However, this issue was compensated by using an extra round of prediction responses, which was the prediction session that learners completed in the Stage 1 survey (this researcher names it as Session 0). Learner models

and related feedback were presented to learners as feed-forward feedback before they started the Session 1 tasks in the Stage 2 experiment. In this way, learners had two opportunities to adjust their predictions.

As discussed in Chapter 2, cognitive feedback in the MCPL literature refers to feedback covering task information, statistical information about task characteristics, information about learners' characteristics, and in particular, information about functional relationships between attributes and outcomes (e.g. Cooksey 1996). From a knowledge representation perspective, the focus of the particular feedback was on “features”, or the product attributes. In this condition, only the aggregated consumer MNL model was applied for comparison. Figure 4.3 demonstrates an example of such feedback. It is based on the attribute “fare” and its relationship with choice probabilities. More examples including screenshots of feedback on every attribute are included in Appendix 4. To help learners understand how much weight they should give to each attribute in making predictions, a summary of attributes' relative importance was also presented to learners and repeated in every round of feedback. Attributes' relative importance was derived from target model's attribute coefficients. Figure 4.4 shows how this information was presented to learners.

#### Fare

Using your predictions in the last survey, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., flying time) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Figure 4.3 An example feedback comparing “fare” in a learner model and the target model

#### Importance of Features in Consumer Decision-Making

In making decisions, consumers think “airline” is the most important factor, accounting for 39% of total importance. This is followed by 35% for “fare”, 12% for “flying time”, 9% on “free in-flight food and beverages”, and 5% for “allowing ticket change”.

Importance of features in decision-making	Consumer Choices
Airline	39
Fare	35
Flying time	12
Ticket change	5
In-flight food & beverages	9

Please click on “>>” to continue.



Figure 4.4 EC3 - relative importance of attributes

#### 4.4.3.4 Experimental Condition 4 (EC4)

The key difference in this condition relative to the previous three conditions is that learners were asked to predict choices of one of three randomly selected consumer groups instead of total consumers. Therefore, the target model for each learner was one of the three consumer class models instead of the aggregated MNL model for all consumers. To ensure learners understood

their target groups, similarities and dissimilarities were illustrated in visual format and the features of each class were explained in text format. Together they were shown as feed-forward information before training tasks (see Appendix 4). After each training task, learners were informed whether they were predicting the group the system had asked them to predict. If not, they were told which other groups they were actually predicting. After the whole session, learners were given feedback showing a summary of their performance over the whole session. This was shown in a line chart format demonstrating moving averages of the chosen logarithmic score rule comparing predicted probabilities and actual probabilities of the target consumer group (see Section 3.4.2 for more background). Figures 4.5 and 4.6 demonstrate the feedback after each task and each session.

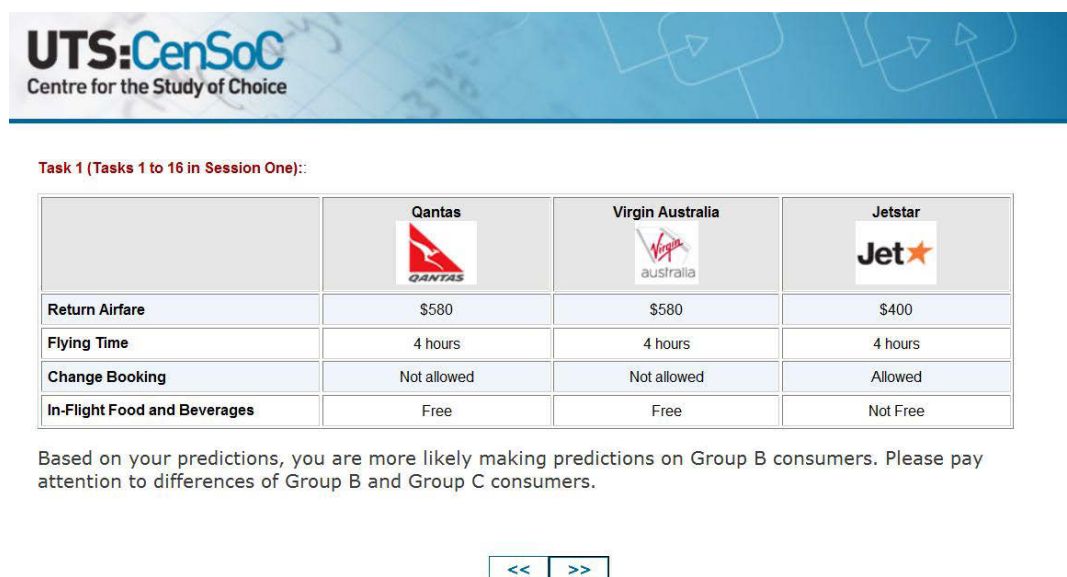


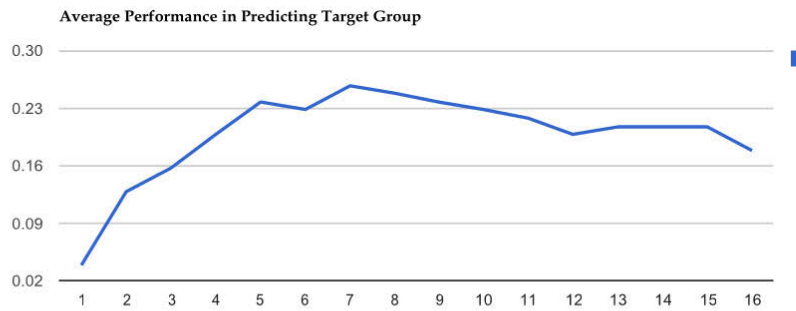
Figure 4.5 EC4 - an example feedback after each task



### Summary of Your Performance in Session One

In total, you were predicting Group C consumers correctly 5 of 16 time(s) in Session One.

The chart below shows how your prediction accuracy increases or decreases over the 16 tasks. Movement of the line shows what your average performance is at a particular task with a score. For example, the position on Point 5 shows your average performance from choice scenario 1 to choice scenario 5.



Please click on ">>" to continue.



Figure 4.6 EC4 - an example feedback after each Session

The method for performing the Bayesian classification task using a logarithmic scoring rule (log-likelihood) was given in Section 3.4.2. To explain briefly, three log-likelihood numbers were calculated by the system in real-time using predicted probabilities in each task. These numbers were compared by the system and one of the three consumer groups with the highest score (the maximum of the three log-likelihood numbers) was denominated as the group being predicted. That is, the predicted probabilities are the closest to the actual probabilities of that class. This answers the question whether a learner is predicting the right group and serves as key information in the after task feedback (Figure 4.6).

Shown in line chart format in Figure 4.6, the key information in the session feedback is a moving average of a “performance” indicator. This indicator is actually the average of all likelihood

numbers for the correct target group. Behaviourally, the likelihood of the target group may be considered as learners' affinity to the target group. Technically, the system calculates averages of likelihood up to 16 tasks of each session for the target group. The likelihood numbers were directly calculated first by converting all log-likelihood scores for every task into the exponential form. To calculate the average likelihood for  $n$  number of tasks starting from task 1, the system calculates the geometric mean, which is the  $n$ th root of the products of all likelihood numbers, as shown in Equation 4.1 below:

$$l_1^n = \sqrt[n]{\prod l_1 l_2 \dots l_n} \quad (4.1)$$

In the exemplified case in Figure 4.6, a learner was observed to improve quickly in the first few tasks, then remain relatively stable in the following tasks.

## 4.5 Summary

Chapter 4 discussed topics related to experimental design and data collection procedures for the empirical study for this research. This study was conducted during August 2012 to October 2012 in the following four steps. First, the researcher collected data for 485 cleaned respondents on cross-country air travel from Sydney to any destinations. In each choice set respondents provided both their own choices and predicted the choices of all consumers. In this way, an extra session of predictions was collected. Second, this researcher conducted an analysis to develop four target models which was used for training in the Stage 2 experiment. There was an aggregated model for all consumers and three consumer group models. Third, all 485 respondents were invited to participate in the Stage 2 learning experiment, though completed by only 252 learners. They had been evenly and randomly split into four experimental conditions based on the four learning approaches discussed in Chapter 2. Fourth, using prediction data collected in both the Stage 2 learning experiment and the Stage 1 consumer survey, data analysis was conducted to test the

research problem and research hypotheses H1a to H3b and explore other findings. Methodologies and results of the analysis will be discussed in detail in Chapter 5.

## Chapter 5 Methodologies and Results of Analysis

### 5.1 Introduction

Chapter 4 discussed the experimental design and fieldwork plans. This chapter will discuss methodologies for data analysis and present the analysis results. For each main type of analysis such as testing prediction accuracy and testing target model parameter learning, methodologies and results will be discussed and presented together. Answers to the research hypotheses will be summarised after the results are presented. Discussions in this chapter will focus on the results only and not extended to implications in the broader context of marketing theories and practices. That task will be kept for Chapter 6 when the conclusions of this research are drawn.

Background information about respondents such as age, gender, and education are given in Appendix 5. This information is not directly relevant to the research problem and hypotheses. This researcher was only interested in the findings about learning approaches that could be generalised more broadly. However, the background information about the sample is important to make sure there is a good coverage of the population and to ensure the findings are not limited to an unacceptably small subsection of the population. In this study, respondents were selected randomly by Pure Profile to match the population statistics of the Australian Bureau of Statistics before invitations were issued to participate in the research. Since this research does not aim to make connections between learning and exogenous factors such as age or gender, individual differences that might have contributed to certain variations in responses are treated as random errors in analysis. As will be discussed later in the chapter, analysis for the Stage 2 learning experiment started with individual level estimation to gain key indicators to further test prediction accuracy and model attribute learning. In this way individual differences have been largely accounted for by individual estimated results.

## 5.2 Analysis Methods and Results for Stage 1 Consumer Survey

The objective for the Stage 1 consumer survey data analysis is clear; analysis results should provide target choice models for the Stage 2 learning experiment. As mentioned before, among the four learning approaches, Learning Approaches One, Two and Three adopt an identical target model representing choices for all consumers. Learning Approach Four uses different target models to be learned; one model for each consumer group.

### 5.2.1 Analysis Methods

A theoretical review of relevant estimation approaches was discussed in Chapter 3. Using choice data collected in the Stage 1 survey, two types of analyses can be conducted. The first type of analysis was to develop an aggregated, fixed effects MNL model. This model represents choice behaviour for all consumers. From a modelling perspective, utility functions in this model for the three alternatives, Qantas, Virgin and Jetstar, can be simplified as Equations 5.1 to 5.3.

$$U_{qantas} = \beta_{ASCqantas} + \beta_{att1fare} + \beta_{att2time} + \beta_{att3change} + \beta_{att4food} \quad (5.1)$$

$$U_{virgin} = \beta_{ASCvirgin} + \beta_{att1fare} + \beta_{att2time} + \beta_{att3change} + \beta_{att4food} \quad (5.2)$$

$$U_{jetstar} = \beta_{ASCjetstar} + \beta_{att1fare} + \beta_{att2time} + \beta_{att3change} + \beta_{att4food} \quad (5.3)$$

These are the main effects utility functions for each alternative without cross effects. There are good reasons for selecting a simple model of learning. Attributes and attribute levels in this study are generic and common for all three airlines, so identical coefficients can be used for generic attributes not varied by alternatives. This model is more parsimonious compared to models treating attributes as alternative specific. A simple model is an advantage for the Stage 2 experiment because it is easier to learn than a complex model. For an unsupervised online experiment that gives learners total control of the time they are willing to spend on the experiment, a complex model with many parameters can cause a high drop-out rate if the number of training sessions increases, or result in insufficient learning if the number of training sessions remains. Therefore,

it is unnecessary to use a complex model for this proof of concept study. Once the results of this study demonstrate that the learning model is a potentially valid approach to improve subjective predictions, more complex models can be tested in future studies.

A simple model can be selected for learning under one condition. That is, the simple model does not neglect any important cross effects between alternatives that are well known to learners and have a major impact on prediction outcomes. Such importance does not mean these cross-effects are statistically significant in a model, but the size of the effect is enough to alter outcomes. A cross effect included is also known as the “Mother Logit” model (McFadden, Train & Tye 1977). Equation 5.4 provides the utility function for one alternative, Qantas as an example.

$$\begin{aligned}
 U_{qantas} = & \beta_{ASCqantas} + \beta_{att1qantas} + \beta_{att2qantas} + \beta_{att3qantas} + \beta_{att4qantas} \\
 & + \beta_{att1virgin} + \beta_{att2virgin} + \beta_{att3virgin} + \beta_{att4virgin} \\
 & + \beta_{att1jetstar} + \beta_{att2jetstar} + \beta_{att3jetstar} + \beta_{att4jetstar}
 \end{aligned} \tag{5.4}$$

The second type of analysis conducted aimed to identify latent classes (consumer groups). On preferences and choices, consumers are considered homogeneous within each class but heterogeneous between classes. Using the two estimation approaches discussed in Chapter 3, Latent Class Modelling (LCM) and archetypal analysis, this analysis was conducted. Results were compared on both goodness of fit statistics and explanations of output classes. This will be discussed after the results are presented.

### 5.2.2 Target Model One – Fixed-Effects MNL Model for All Consumers’ Choices

Table 5.1 shows the results of a conditional logit model using the choice data of 485 respondents of the Stage 1 survey after data cleaning. There are nine independent estimated parameters, three for ASCs (Qantas, Virgin and Jetstar), three for return airfare (\$400, \$460, and \$520), one for flying time (four hours), one for ticket change (allowed) and one for in-flight food and beverages (free).

For each attribute, the effect of the missing level is the negative sum of other independent levels.

This is due to the effects coding method that was used as a default in preparing choice data.

Table 5.1 Conditional logit model for cross-country flight offer choices

Conditional (fixed-effects) logistic		Number of obs	=	31040
		Wald chi2(9)	=	1199
		Prob > chi2	=	0
Log likelihood = -5503.0962		Pseudo $\rho^2$	=	0.4884
Most Preferred	Coef.	Std. Err.	z	P>z
Qantas (ASC)	0.93	0.06	15.31	0.00
Virgin (ASC)	0.89	0.06	15.30	0.00
Jetstar (ASC)	0.04	0.06	0.70	0.48
Not fly (ASC)	-1.85			
\$400	1.98	0.07	26.82	0.00
\$460	0.51	0.03	15.91	0.00
\$520	-0.81	0.04	-18.89	0.00
\$580	-1.68			
4 hours	0.56	0.03	19.93	0.00
6 hours	-0.56			
Ticket changes allowed	0.22	0.02	10.25	0.00
Ticket changes not allowed	-0.22			
In-flight food & beverages - Free	0.43	0.03	15.83	0.00
In-flight food & beverages - Not free	-0.43			

From the results, it can be said that alternative specific constants and attribute level coefficients are all significant (i.e. although Jetstar ASC is not significant, the alternative specific constants are significant as a group. It is also clear from McFadden Pseudo  $\rho^2$  ( $1 - L/L_0$ ) with a value of 0.4884, that this model has a very good statistical fit. According to McFadden (1979, p. 307), “values of 0.2 to 0.4 for  $\rho^2$  represents an excellent fit”. This shows that pre-experimental analysis using past study data helped to identify important attributes that mattered most in choices made and, updated attribute levels from recent market information greatly increased the relevance of the model to consumer choices. It is worth noting that brand effect for Qantas (0.93) dropped significantly relative to Virgin (0.89) compared to the last study conducted by the CenSoC study (see Table 4.1). In the previous study, the effect for Qantas was more than twice that for Virgin. In this model, the two airlines are almost identical. Effects for all attributes match common expectations.

Basically, consumers prefer cheaper fares, shorter flying time, flexibility of changing ticket and free in-flight food/beverages. The effect size for the cheapest fare \$400 (1.98) is almost four times the size of the next level \$460 (0.51). Fares at \$520 and \$580 have negative effects on choices. The effects of three other attributes are statistically significant but less important to choices in the order of flying time, in-flight food/beverages and ticket change.

A mother logit model was also conducted to check whether there were any statistically significant cross-effects that were important with large effect sizes. Table 5.2 shows the results of this model. From the results, the statistical fit of this model is slightly better than the simple MNL model with McFadden Pseudo  $\rho^2$  at 0.5035 compared to 0.4884 for the previous model. However, there are 48 more independent parameters in the model. Compared to the simple model, this model only improves marginally with many more terms. By examining cross-effects, it becomes clear that there are six statistically significant cross-effects (highlighted in bold in Table 5.2), one each for Qantas and Virgin and four for Jetstar. The attributes of Qantas and Virgin offers have more impact on Jetstar than the reverse. These cross-effects are not as important as main effects by looking at their sizes. In this case, the simpler model in Table 5.1 is better suited to be the target learning model and no important effects will be missed that have a major impact on prediction results.

Table 5.2 Mother Logit model including all cross-effects (CEs)

<b>Conditional (fixed-effects)</b>		Number of obs	=	31040
		Wald chi2(57)	=	10832.84
		Prob > chi2	=	0
Log likelihood = -5341.2225		Pseudo R2	=	0.5035
Most Preferred	Coef.	Std. Err.	z	P>z
Qantas (ASC)	0.97	0.04	24.94	0.00
Virgin (ASC)	0.89	0.04	22.16	0.00
Jetstar (ASC)	-0.12	0.05	-2.24	0.03
Qantas fare - \$400	1.63	0.20	8.15	0.00
Qantas fare - \$460	0.62	0.17	3.60	0.00
Qantas fare - \$520	-0.97	0.16	-6.22	0.00
Virgin fare on Qantas (CE) - \$400	0.11	0.22	0.50	0.62
Virgin fare on Qantas (CE) - \$460	-0.07	0.18	-0.40	0.69
Virgin fare on Qantas (CE) - \$520	0.06	0.15	0.42	0.67



Jetstar fare on Qantas (CE) - \$400	0.02	0.20	0.08	0.94
Jetstar fare on Qantas (CE) - \$460	-0.02	0.17	-0.11	0.91
Jetstar fare on Qantas (CE) - \$520	0.10	0.16	0.59	0.56
Qantas time – 4 hours	0.20	0.09	2.18	0.03
Virgin time on Qantas (CE) – 4 hours	0.02	0.09	0.22	0.83
Jetstar time on Qantas (CE) – 4 hours	-0.10	0.11	-0.91	0.36
Qantas booking change – allowed	0.25	0.09	2.75	0.01
Virgin booking change on Qantas (CE) - allowed	-0.02	0.09	-0.20	0.84
Jetstar booking change on Qantas (CE) - allowed	-0.12	0.10	-1.20	0.23
Qantas food and beverages – free	0.06	0.09	0.63	0.53
Virgin food/beverages on Qantas (CE) – free	<b>-0.19</b>	<b>0.10</b>	<b>-1.96</b>	<b>0.05</b>
Jetstar food and beverages on Qantas (CE) – free	-0.11	0.10	-1.20	0.23
Virgin fare - \$400	1.79	0.21	8.42	0.00
Virgin fare - \$460	0.51	0.18	2.84	0.01
Virgin fare - \$520	-0.75	0.15	-4.98	0.00
Qantas fare on Virgin (CE) - \$400	-0.28	0.21	-1.35	0.18
Qantas fare on Virgin (CE) - \$460	0.14	0.18	0.77	0.44
Qantas fare on Virgin (CE) - \$520	-0.18	0.15	-1.18	0.24
Jetstar fare on Virgin (CE) - \$400	-0.05	0.20	-0.25	0.81
Jetstar fare on Virgin (CE) - \$460	0.05	0.17	0.27	0.79
Jetstar fare on Virgin (CE) - \$520	-0.04	0.16	-0.27	0.79
Virgin time – 4 hours	0.61	0.09	6.41	0.00
Qantas time on Virgin (CE) – 4 hours	<b>-0.19</b>	<b>0.09</b>	<b>-2.13</b>	<b>0.03</b>
Jetstar time on Virgin (CE) – 4 hours	-0.14	0.11	-1.29	0.20
Virgin booking change – allowed	0.35	0.09	3.72	0.00
Qantas booking change on Virgin (CE) - allowed	0.03	0.09	0.35	0.72
Jetstar booking change on Virgin (CE) - allowed	0.03	0.10	0.34	0.73
Virgin food and beverages – free	0.34	0.10	3.43	0.00
Qantas food and beverages on Virgin (CE) – free	-0.15	0.09	-1.64	0.10
Jetstar food and beverages on Virgin (CE) – free	-0.01	0.10	-0.13	0.90
Jetstar fare - \$400	2.10	0.20	10.42	0.00
Jetstar fare - \$460	0.35	0.18	1.99	0.05
Jetstar fare - \$520	-0.51	0.17	-2.95	0.00
Qantas fare on Jetstar (CE) - \$400	-0.32	0.23	-1.41	0.16
Qantas fare on Jetstar (CE) - \$460	-0.14	0.19	-0.73	0.46
Qantas fare on Jetstar (CE) - \$520	-0.08	0.16	-0.52	0.60
Virgin fare on Jetstar (CE) - \$400	0.14	0.23	0.60	0.55
Virgin fare on Jetstar (CE) - \$460	0.15	0.19	0.79	0.43
Virgin fare on Jetstar (CE) - \$520	-0.16	0.16	-0.97	0.33
Jetstar time – 4 hours	0.49	0.12	4.16	0.00
Qantas time on Jetstar (CE) – 4 hours	<b>-0.37</b>	<b>0.09</b>	<b>-3.89</b>	<b>0.00</b>
Virgin time on Jetstar (CE) – 4 hours	<b>-0.23</b>	<b>0.10</b>	<b>-2.28</b>	<b>0.02</b>
Jetstar booking change – allowed	0.02	0.10	0.21	0.83
Qantas booking change on Jetstar (CE) - allowed	-0.14	0.10	-1.46	0.15
Virgin booking change on Jetstar (CE) - allowed	-0.08	0.10	-0.86	0.39
Jetstar food and beverages – free	0.39	0.11	3.67	0.00
Qantas food/beverages on Jetstar (CE) – free	<b>-0.32</b>	<b>0.10</b>	<b>-3.39</b>	<b>0.00</b>
<b>Virgin food/beverages on Jetstar (CE) – free</b>	<b>-0.26</b>	<b>0.10</b>	<b>-2.52</b>	<b>0.01</b>

The results also show that choices for Qantas and Virgin are quite independent of cross-effects (with two minor exceptions as highlighted in bold). This means that choices made with regard to the Qantas or Virgin offers are largely due to the utilities of these offers and not influenced by offers from other airlines. Choices made on Jetstar are more influenced by effects from Qantas

and Virgin. This finding did help to reduce the orthogonal design size for the Stage 2 experiment by combining two orthogonal designs, one covering Qantas and Virgin and one for Jetstar. This helps to cut the size of the design by half (16 sets instead of 32 sets, see Appendix 2). However, some correlations are allowed between the Qantas/Virgin design and the Jetstar design.

In the Stage 1 survey, after respondents provided their own choices, they were also asked to predict consumers' choice probabilities for each set. Without using a more complex method to examine how much better the model in Table 5.1 performs over respondents' subjective predictions, a summary table of variations from actual choice probabilities is summarised in Table 5.3. The mean differences from actual probabilities were calculated over all respondents, all sets over all alternatives in sets. It is clear that the model performs much better than subjective probabilities in predicting choices. For example, the mean difference of model predicted probabilities is 0.04 above or below the actual choice probabilities. Meanwhile, the mean difference of respondents' predicted probabilities is 0.16 above or below the actual probabilities. In fact, the model performs better than the best performed respondent in making predictions ( $\pm 0.07$ ). This is not a surprising finding given all the evidence from research in the area of Judgment and Decision Making literature (e.g. Dawes 1971; Grove et al. 2000). In the analysis conducted in Stage 2, more rigorous measures were used to calculate differences of probability distributions (i.e. strictly proper scoring rules as discussed in Chapter 3). For a simple summary of Stage 1, Table 5.3 provides adequate data for ease of interpretation.

Table 5.3 Model predictions and learner predictions vs. actual choice probabilities

Average differences of probabilities	Model vs. Actual	Learner Prediction vs. Actual
All Options	$\pm 0.04$	$\pm 0.16$
Fly Qantas	$\pm 0.05$	$\pm 0.19$
Fly Virgin	$\pm 0.05$	$\pm 0.19$
Fly Jetstar	$\pm 0.04$	$\pm 0.16$

Not fly	$\pm 0.01$	$\pm 0.08$
"most preferred" option in all sets	$\pm 0.05$	$\pm 0.25$
"least preferred" option in all sets	$\pm 0.03$	$\pm 0.12$

### 5.2.2 Target Model Two – Fixed-Effects MNL Models of Consumer Classes

Class models were prepared in anticipation of the target models used in Learning Approach Four. As discussed in Chapter 3, two methods were used for this analysis, archetypal analysis and Latent Class Model (LCM). With consideration of the sample size of learners in this learning approach (initially estimated to be around 50 learners), models for two, three, four, and five classes were tested, and a decision was made that the three classes approach was the most appropriate one for the Stage 2 experiment. First, results for two classes always show one class is more dominant with a much higher proportion than the other, and this may trigger different behaviours by learners, which is not a feature of this learning approach. Four or more classes introduce too many small differences that are not salient enough to differentiate for interpretation. This may be a problem for limited sample size and limited online learning time (initially estimated to be about 30 minutes maximum online).

A comparison of results for two methods of reaching three classes shows not only archetypal analysis has a better fit overall, but also identified more balanced proportions and intuitively more meaningful classes (consumer groups). Choosing the three groups identified by the archetypal analysis, fixed-effects MNL models were developed and used as target models for Learning Approach Four in the Stage 2 experiment. Having a clear definition of each consumer group certainly helps explain consumer behaviour to learners. Also, having near evenly split classes can suggest to learners that there is a near equal opportunity that a consumer may fall into any of the three groups. This helps eliminate possibly irrelevant and unintended heuristics developed by

learners using this learning approach. It certainly helps to simplify the classification algorithm by ignoring prior class probabilities (see Equation 3.25 in Chapter 3).

As shown in Table 5.4, average Pseudo  $\rho^2$  indicates the archetypal analysis produced class models with high statistical fits. The three classes produced by archetypal analysis also have evenly split proportions at 38%, 32% and 30%.

Table 5.4 Results of archetypal analysis

	Archetypal
Log-Likelihood (Sum of 3 Classes)	-4670.87
Pseudo R2 (Average)	0.57
Pseudo R2 (Class 1)	0.61
Pseudo R2 (Class 2)	0.45
Pseudo R2 (Class 3)	0.64
Proportion % (Class 1)	38%
Proportion % (Class 2)	32%
Proportion % (Class 3)	30%

Table 5.5 shows the results of the three fixed-effects models generated by archetypal analysis. From the effects of attributes, it is also clear that differences between the three classes are salient for interpretation if described in words. Among the three classes, the key characteristic for Class 1 is the choice of cheapest fares, the key characteristic for Class 2 is the choice of Qantas and Virgin, and Class 3 is between the other two classes and selected more on other features such as flying time and in-flight food.

Table 5.5 Results of three fixed-effects models for the three classes by archetypal analysis

	Class 1	Class 2	Class 3
Log-Likelihood	-1615.76	-1916.33	-1138.79
Pseudo R2	0.61	0.45	0.64
Most Preferred	Coef.	Coef.	Coef.
Qantas (ASC)	0.81	1.54	1.03
Virgin (ASC)	0.71	1.33	1.25
Jetstar (ASC)	-0.22	-0.21	0.80
Not fly (ASC)	-1.30	-2.66	-3.09
\$400	3.37	0.98	2.17
\$460	0.91	0.52	0.66
\$520	-1.88	-0.40	-0.89
\$580	-2.39	-1.10	-1.94
4 hours	0.56	0.49	0.92
6 hours	-0.56	-0.49	-0.92
Ticket changes allowed	0.36	0.32	0.26

Ticket changes not allowed	-0.36	-0.32	-0.26
In-flight food & beverages - Free	0.40	0.20	0.92
In-flight food & beverages - Not free	-0.40	-0.20	-0.92

### 5.3 Analysis Methods and Results for Stage 2 Learning Experiment

The objective of the analysis for the Stage 2 learning experiment is to test hypotheses H1a to H3b regarding prediction accuracy and model learning, especially the comparison between the results of Learning Approaches Two and Three with Learning Approach One. Learning Approach Four is a separate condition more for exploratory purposes to see whether the idea of categorisation learning works. Experimental designs used for all learning approaches are identical and questions asked in training tasks are the same for Approaches One, Two and Three. Prediction accuracy and target model parameter learning are treated as two distinct types of analysis and they will be discussed in separate sections with methods and results.

Before going into details of the analysis, it is worth summarising some response rates, durations of the learning experiment and feedback from learners, to gain a better understanding of how learners reacted when they were engaged in the experiment.

#### 5.3.1 Response Rate, Duration and Learners' Feedback

Of the 485 respondents from Stage 1 who were invited to complete the learning experiment, 365 entered the online link for the experiment. Of those, 258 respondents completed the learning experiment. To balance four experimental conditions and remove those respondents who were most likely not applying effort to learn, six respondents with the shortest duration (all below 10 minutes) were deleted from data for analysis. In total, 252 respondents (hereafter “learners”) were retained in the data set for analysis. The above figures yield a completion rate of 53% of total respondents invited to participate. No reminders were sent to invite those 107 respondents who appeared to have quit the experiment because this researcher only wanted to retain those who

completed the experiment at one sitting. Including learners who completed the experiment in more than one sitting, or on separate days, can introduce other unknown exogenous factors such as the veracity of participants' memory capacity.

As show in Table 5.6, the time spent by learners in the experiment varied greatly, and overall the learners spent much more time in this learning experiment than for standard online surveys. In this case, the mean duration may not be an accurate indicator given the number of outliers representing much more than 60 minutes in total. Some participants may have taken a break during learning. It is equally possible that the outliers may reflect respondents of more deliberative nature who diligently took time to "get it right", then perhaps the nature of the content of each Approach gave rise to the variations. By examining the medians and percentiles, it becomes clear that learners under each learning approach spent a similar amount time in completing the experiment. Among the four approaches, learners under Approach Four spent the longest time in the experiment, whilst learners under Approach Three spent the shortest time in the experiment.

Table 5.6 Duration for Stage 2 Learning Experiment

<b>Experimental Conditions</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>
<b>Approach</b>	<b>One</b>	<b>Two</b>	<b>Three</b>	<b>Four</b>
Time Spent by	Median: 31.4	Median: 32.2	Median: 28.0	Median: 34.2
Learners (in	25 Percentile: 20	25 Percentile: 20	25 Percentile: 21	25 Percentile: 24
minutes)	50 Percentile: 31	50 Percentile: 32	50 Percentile: 28	50 Percentile: 34
	75 Percentile: 61	75 Percentile: 65	75 Percentile: 55	75 Percentile: 67

Table 5.7 shows learners' opinions regarding the four learning approaches. Among the four learning approaches, Approach One (online DSS) induced the greatest confidence among learners with 69.8% learners saying they would improve using this method. This is followed by Approach Two (outcome feedback) at 66.7%. People's opinions regarding Approach Three and Approach Four were more or less evenly divided with around half saying they would improve and the other half saying they would not improve. Interestingly, results to be shown in later sections about

learners' actual improvements in their prediction accuracy and model learning, will reveal that learners' own feelings are not necessarily a reliable indicator in this regard.

Table 5.7 Respondents' feelings about the four learning approaches after Session 2

My predictions will improve using this learning approach	Learning Approach One	Learning Approach Two	Learning Approach Three	Learning Approach Four	Total
Yes	69.8%	66.7%	50.8%	55.6%	60.7%
No	30.2%	33.3%	49.2%	44.4%	39.3%
Total	100.0%	100.0%	100.0%	100.0%	100.0%

The overall willingness to participate in future learning experiments similar to this one was very high. As shown in Table 5.8, up to 96.8% of all 252 learners said they would participate in another training program like this one. This high proportion is not varied by learning approaches.

Table 5.8 Respondents' willingness to participate in future training after Session 2

If there is another survey or training program in future, are you interested in participating?	Learning Approach One	Learning Approach Two	Learning Approach Three	Learning Approach Four	Total
Yes	98.4%	96.8%	96.8%	95.2%	96.8%
No	1.6%	3.2%	3.2%	4.8%	3.2%
Total	100.0%	100.0%	100.0%	100.0%	100.0%

From open-ended feedback, most learners thought the experiment was interesting and helped them to improve. For Approach One, the most common feeling in feedback was regarding the easy to use feature of the online DSS. For Approach Two, learners mentioned that the training method with correct answers given immediately after their responses was like playing a game and they had to “gauge” which attributes caused the differences between their answers and actual answers. For Approach Three, many mentioned that the comparison of their own models with actual consumer models gave them helpful information. However, some did complain that the information given was too much and became difficult and confusing for them to process. For Approach Four, it was interesting to see the contrast in opinions on the approach. Some said the approach worked for them and they improved a lot in later tasks. Others said they did not improve

and did not think the method worked for them. Interestingly, some participants who were assigned to predict a consumer group, and who were also a member of that group (all learners were informed in the beginning which groups they belonged to, see Appendix 4) queried whether the feedback they received was correct. They believed they were predicting the right consumer group but were told otherwise by the feedback. This researcher speculates that in telling learners which groups they belonged to might have provided some learners with a false assumption that they could predict independently of the experiment according to their own ideas.

### **5.3.2 Testing Prediction Accuracy**

To test hypotheses of prediction accuracy across learning approaches, it is worth summarising what data was collected for analysis. All 252 learners were randomly and evenly divided into four experimental conditions matching the four learning approaches. Each learner completed 32 identical prediction tasks in the Stage 2 experiment. The 32 tasks were arranged in two training sessions with 16 tasks per session. For Approaches One, Two and Three, learners were asked to predict consumers' choices of the "most preferred" option and the probability for each option. For Approach Four, learners were asked to answer the same types of questions but for an assigned consumer group. These two sessions were called Session 1 and Session 2 (hereinafter referred to as "S1" and "S2"). Besides these two sessions, all learners also had an extra session of prediction tasks in the Stage 1 survey immediately following their own choices in each task. This session is called Session 0 (hereinafter referred to as "S0"), a prior learning session. For learners trained under Approaches One, Two and Three, data for all sessions can be used for analysis because both predicted and true probabilities (approximated by the target model) are available for the same target consumers. For learners trained under Approach Four, we can only use data in Sessions 1 and 2 because only in these two sessions were predictions made of assigned consumer groups.



Thinking about the data, learners actually made more than one kind of prediction in the training tasks. First, learners predicted the “most preferred” option. They chose options directly in S1 and S2 and indirectly in S0 by giving the chosen option(s) the highest probability. To test the accuracy of this prediction, a simple method is to cross-tabulate the correct answers with predicted answers. The results should provide the performance under each learning approach. By aggregating the number of correct predictions per learner in each session, a respondent level cross-tabulation can also help to answer how many learners improved their prediction accuracy in later sessions. For example, if a learner correctly predicted his/her preferred options on 8, 12 and 12 occasions of the 16 possible in Sessions 0, 1 and 2 respectively, we can conclude that this learner improved in Session 1, did not improve further in Session 2, but overall improved during the whole experiment. This individual level analysis can provide indicators as to whether improvements are widely spread among learners, or only contributed by a few learners whilst the large bulk of learners performed poorly.

For Approach Four, learners made a direct prediction of the probabilities and an indirect prediction of the target consumer group they were asked to predict. Only when a learner’s prediction was closest to the target group, would the built-in Bayesian classifier inform the learner that a correct prediction had been made. To analyse this prediction accuracy, cross-tabulations should be adequate because both the learner predicted consumer group and the target consumer group are known for each learner and each task. The analysis should inform with regard to performance in both S1 and S2.

The above two types of predictions can be analysed by a simple cross-tabular method. The most important type of prediction to test prediction accuracy is learners’ predictions of probabilities for all four options in each set. These subjective predictions need to be analysed to see how well they

match with actual probabilities behind target learning models. For 252 learners, each learner made probability predictions covering four options in each of 48 choice sets in three sessions (S0, S1, and S2). Excepting S0 for Approach Four, we can compare predicted probabilities with actual probabilities in every set across all learners. As discussed in Chapter 3, this analysis should use strictly proper scoring rules to test how close the two probability distributions match.

Two discrete probability distributions with four options can be denoted as  $P(p_1, p_2, p_3, p_4)$  and  $Q(q_1, q_2, q_3, q_4)$ , with  $P$  representing learners' predicted probability vector and  $Q$  representing actual probability vector. To test prediction accuracy is to analyse how “close” are the two distributions and to further compare the results of this analysis by learning approaches and sessions. In comparing the results for the learning approach, an effective method should demonstrate a close match between two distributions for both the response and respondent levels. For the response level, the results should indicate whether the approach has triggered better prediction accuracy overall covering all tasks and all learners. For the respondent level, results should indicate whether learners have improved their accuracy and reduced prediction errors in later sessions. For this analysis, a scoring rule is required as a key measure to directly show the closeness of  $P$  and  $Q$ .

According to the literature on scoring rules and general probability theory, an effective way to find out how close are two probability distributions with discrete outcomes, is to use a distance function (metric) to calculate a positive distance between the two distributions (e.g. Friedman 1983; Gneiting & Raftery 2007; Nau 1985; Pollard 2002). Distance functions are strictly proper and have properties that other scoring rules do not have such as non-negativity, symmetry and transitivity. Moreover, it is also convenient to find a distance function with clear lower and upper bounds and the preferred target known *a priori*.

Hellinger distance (hereinafter referred to as “HD”) is often chosen as a distance measure between probability densities. It is by nature a Euclidean distance applied to square roots of the probability vectors. It is categorised as a generalised spherical scoring rule. Pollard (2002) discussed it as a measurement tool in the context of probability and measurement theory. More recently Jose, Nau and Winkler (2008) discussed it and its properties, such as symmetry, in the context of scoring rules. For two discrete probability vectors with  $k$  elements  $P(p_1, p_2, \dots, p_k)$  and  $Q(q_1, q_2, \dots, q_k)$ , HD represents how close or similar are the two distributions. Its formula is shown below in Equation 5.5.

$$HD(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{p_i} - \sqrt{q_i})^2}, \quad 0 \leq H(P, Q) \leq 1 \quad (5.5)$$

If all matching probabilities in the two probability distributions are identical, HD equals zero. If for every non-zero probability in one distribution, the matching probability in the other distribution is zero, then HD equals one. Both are extreme cases and unlikely to occur in real predictions. Real HDs are most likely to situate between the two numbers. A lower HD learning towards zero represents that a learner is making less errors and is more accurate in prediction.

In analysing prediction accuracy in this experiment, once HDs are calculated for every training set, the following two actions can be taken. First, for the response level, all HDs can be used for response level analysis. To use common statistical methods for linear and normally distributed data, Tukey’s ladder transformations can be applied to find the best transformation to normalise HDs. Once HDs are normalised, common statistical methods such as mean comparisons by learning approaches and sessions can be conducted easily. The results provide an overall comparison of the prediction accuracy of the four learning approaches relative to the three sessions for the response level. The alternative method is to use HD or its transformation as the dependent variable, using either learning approaches combined with sessions as the independent variable to

perform a regression analysis. By comparing coefficients, it should also inform which learning approach yields better prediction accuracy and whether the accuracy also improves with successive sessions.

The foregoing response level analysis can show which approach performs best by treating each HD as an independent data unit, without considering that every 48 HDs came from one learner (i.e. as discussed, every learner has 48 HDs calculated, one for each training task). For the respondent level, if a learner did improve his/her accuracy in training, the HD should decrease in later sessions. For each learner, there are three mean HDs, one for each session (for Approach Four, only two mean HDs, one for each of S1 and S2, will be used for analysis). Using this data, the model discussed below can be run. The results should inform which learning approach is more effective for individual learners, causing them to improve their prediction accuracy more quickly in latter sessions.

Taking two sessions for example (S0 and S1), for each learner, we can calculate the difference of two mean HDs for two sessions. The difference  $H_1 - H_0$  can be denoted as  $\Delta H^{1-0}$ . If  $H_1$  is lower than  $H_0$ , it means the prediction accuracy has been improved in S1, so  $\Delta H^{1-0}$  should be negative and vice versa. Thinking about  $\Delta H^{1-0}$  as a function of  $H_0$ , the model is expressed as:

$$\Delta H^{1-0} = \alpha + \beta H_0 + \epsilon \quad (5.6)$$

The intercept  $\alpha$  represents a systematic prediction error component. When it is 0, it means a learner does not have errors that are systematic over or under prediction errors that will not be corrected in S1. The term  $\beta$  determines the relationship of  $\Delta H^{1-0}$  and  $H_0$ . It informs how much error is corrected in S1. The term  $\epsilon$  represents an unknown random error component.

In estimating  $\alpha$  and  $\beta$ , assume there is no systematic error ( $\hat{\alpha} = 0$ ), and that Equations 5.7 to 5.9 inclusive, present three representative changes to the prediction accuracy from S0 to S1. In Equation 5.7, a learner is able to eliminate all prediction errors in S0 and make perfect predictions in S1. In Equation 5.8, a learner is able to reduce errors made in S0 by one half. In Equation 5.9, a learner does not improve from S0 to S1. In Equation 5.10, the HD in S1 is the negative of HD in S0. This is not possible because HD is always positive.

$$\text{If } H_1 = 0 \text{ then } \hat{\beta} = -1 \text{ because } \Delta H^{1-0} = H_1 - H_0 = -H_0 \quad (5.7)$$

$$\text{If } H_1 = \frac{1}{2}H_0 \text{ then } \hat{\beta} = -\frac{1}{2} \text{ because } \Delta H^{1-0} = H_1 - H_0 = -\frac{1}{2}H_0 \quad (5.8)$$

$$\text{If } H_1 = H_0 \text{ then } \hat{\beta} = 0 \text{ because } \Delta H^{1-0} = H_1 - H_0 = 0 \quad (5.9)$$

$$\text{Can } H_1 = -H_0? \text{ It is not possible because Hellinger Distance is } \geq 0 \quad (5.10)$$

The above cases suggest that  $\beta$  is a coefficient greater than or equal to -1 and less than or equal to 0. If  $\beta$  is equal to -1, it means that all prediction errors are removed completely in S1 and predictions are perfect in S1. In comparing two learning approaches, an approach with a lower  $\beta$  closer to -1 is preferred. Following this principle, testing prediction accuracy can be conducted by comparing values of  $\hat{\beta}$  for different learning approaches. We can conclude that a learning approach with lower negative  $\hat{\beta}$  is more effective in improving prediction accuracy, however, it is also important to consider the relative size of  $\alpha$ . An approach with a higher  $\alpha$  suggests that the approach produces more systematic over- or under-predictions among learners. Therefore, if an approach has a low value of  $\beta$  but a higher  $\alpha$ , it suggests that learners under this approach are adjusting their predictions rapidly, but unfortunately with more numerous systematic errors. If this is the case, a competing approach with a higher  $\beta$  but lower  $\alpha$  may be preferred. It means that although learners are improving their prediction accuracy slowly, they are doing so with a lesser number of systematic errors.

In summary, to test the prediction accuracy for the two categorical predictions of “most preferred” option and consumer group, cross tabulating learners’ predictions by correct answers can be used. To test the prediction accuracy of probability predictions, a distance measure as a strictly proper scoring rule can be chosen. This researcher selected Hellinger distance (HD) as the scoring rule to test the relative proximity of learner predictions and actual probability distributions. Moreover, this can be tested for both a response level using all HDs and a respondent level using the mean HDs of each learner in each session. The former analysis provides an overall picture of prediction accuracy by learning approaches and the latter analysis reveals which approach reflects improvement in individual learners in latter sessions.

### **5.3.3 Results - Prediction Accuracy**

#### **5.3.3.1 Prediction Accuracy for “Preferred” Option**

The first prediction type checked is the prediction for the “preferred” option in each set. For the response level, Table 5.9 shows the proportions of correctly predicted responses in each learning approach, separated by sessions. Approach Two has the highest proportion of correct predictions at 80.8% representing the mean of sessions S1 and S2 in the Stage 2 learning experiment, 13% more than the starting session S0 at 67.8%. Although the mean proportions for Approach One and Approach Three at 76.6% and 76.5% respectively, are lower than Approach Two, both approaches see learners improve their accuracy in both sessions S1 and S2, this being especially the case for Approach Three. For Approach Four, S0 is not relevant because tasks involved in sessions S1 and S2 are about a particular consumer segment, so the effective starting session for Approach Four is S1. From the results for Approach Four, it can be simply said that learners did not improve in S2 over S1. The overall performance of Approach Four in this type of prediction is lower than the other three approaches. However, a direct comparison with the other approaches would be meaningless because of different prediction tasks involved.

Table 5.9 Correct predictions of preferred options for total predictions

	(n = 1008 responses per cell)			
	Approach One	Approach Two	Approach Three	Approach Four
<b>Session 0 (Stage 1)</b>	68.4%	67.8%	66.5%	
<b>Session 1</b>	75.9%	80.6%	74.9%	70.2%
<b>Session 2</b>	77.4%	81.0%	78.2%	65.3%
<b>Average (S1 &amp; S2)</b>	76.6%	80.8%	76.5%	67.8%

For the respondent level, each respondent was categorised into three groups: those who predicted correctly more often in S2 than in S1 (better), those who predicted correctly an equivalent number of times in S1 and S2 (same), and those who predicted correctly less often in S2 than in S1 (worse). As shown in Table 5.10, Approach Three had the highest proportion of better performers at 54%, with Approach Four having the second highest proportion at 50.8%. These were followed by 44.4% and 38.1% respectively for Approaches One and Two. It is no surprise that more learners improved under Approach Three, but it is surprising that half of the learners actually improved in Approach Four. Looking at the Table 5.10 results in totality, it seemed the case that those who did not improve in S2 became worse and contributed more to the overall responses of incorrect predictions. Learners in Approach Two did not improve much in S2 over S1. However, this is partly due to the fact that in S1, the learners had already achieved the highest level of accuracy compared to all other approaches, so there were less room for improvement.

Table 5.10 Respondents' performance in predicting preferred options

	<b>Individuals who predicted better/same/worse in Session 2 than in Session 1 (base = 63 respondents per approach)</b>			
	Approach One	Approach Two	Approach Three	Approach Four
<b>S2 &gt; S1 (better)</b>	44.4%	38.1%	54.0%	50.8%
<b>S2 = S1 (same)</b>	17.5%	33.3%	22.2%	14.3%
<b>S2 &lt; S1 (worse)</b>	38.1%	28.6%	23.8%	34.9%

Respondents do not always perform consistently across all tasks, as the difficulty level of task sets is unequal. Figure 5.1 shows the proportions of respondents who predicted correctly in the 32 sets

in S1 and S2. When tasks are easier to predict, learners in every approach performed well in general (such as Task sets 7 and 13 in S1). When tasks are difficult, learners in every approach performed relatively poorly (such as Task sets 2 and 13 in S2). An interesting finding is that although learners in Approach Two performed well overall, in those more difficult tasks (such as Tasks sets 15 and 16, and Task sets 2 and 13), a lesser number of learners performed well. This may be the unique characteristic for “trial-’n-error” learners. Perhaps if similar tasks had not been seen before without references for learners to apply, performance could have been worse than in other learning approaches. Surprisingly, although learners in Approach Four did not perform well overall, they had a more consistent and stable performance in S1 in particular, performing better than learners in the other approaches when tasks were difficult in general.

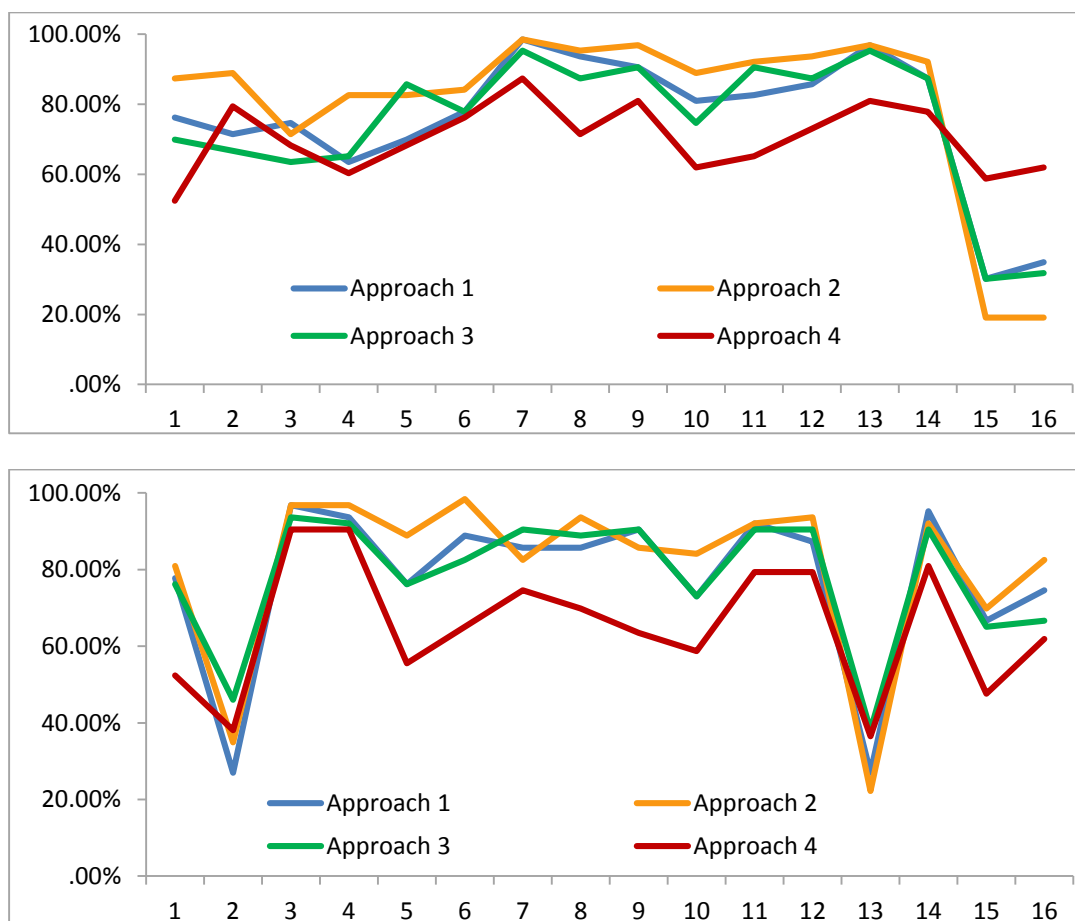


Figure 5.1 Proportions of respondents who predicted correctly in S1 and S2 by set



### 5.3.3.2 Prediction Accuracy on Target Consumer Group (Learning Approach Four)

The second type of predictions relate to Learning Approach Four only. Each of the 63 learners was asked to predict probabilities of one of the three consumer groups. Based on the Bayesian classifier using posterior probabilities, the intelligent tutoring system can determine which group each learner was predicting. The automatic feedback then informed learners whether they had predicted the correct group, and if not, which group had been predicted. In Table 5.11, it is clear that learner prediction success varies greatly. The worst performing learner in each of sessions S1 and S2 only predicted correctly on one occasion, whilst the best performer achieved 15 correct predictions of a possible 16. The mean number of correct predictions by learners in both sessions S1 and S2 were 6.98 and 7.27 respectively, of 16 predictions.

Table 5.11 Learners' performance in correctly predicting target consumer group

	Learners	Minimum	Maximum	Mean	Std. Deviation
Session 1	N=63	1 (of 16)	15 (of 16)	6.98	4.99
Session 2	N=63	1 (of 16)	15 (of 16)	7.27	4.25

It is also interesting to note that people performed differently in predicting the three groups. Since learners were all randomly assigned to the three groups, the differences shown in Table 5.12 are most likely due to the degree of difficulty arising from the specific nature of a group's task composition.

Table 5.12 Prediction success achieved by learners (% of total responses)

Target groups to predict	Session 1	Session 2	Total
Group A	37.2%	36.6%	36.9%
Group B	75.3%	73.8%	74.6%
Group C	18.5%	25.9%	22.2%
Total	43.7%	45.4%	44.5%

Learners predicted Group B most accurately with about 75% success rate for both S1 and S2. Group B consumers looked for qualities and cheap fares were less attractive to them than to other groups. They preferred Qantas and Virgin over Jetstar clearly. These characteristics were strong

signals for learners to assign higher probabilities to the Qantas and Virgin. On the other hand, even though Group A consumers strongly preferred the cheapest fare, learners could still encounter difficulty in judging what probabilities they should assign to options with cheaper fares in a training set. This situation prevails because all groups preferred cheaper fares to a certain degree. It is not surprising that learners performed worst in predicting Group C, as Group C has a profile which fits between Groups A and B, and the Group C characteristics are not as salient as the other two groups. There is therefore a lack of strong signals indicating levels of preference for the options. These are valuable insights for the training of people to understand the various segments. In other words, there need to be strong signals demonstrating a group's main characteristics, and especially differentiating those characteristics from other groups.

Table 5.13 shows data which gives rise to an interesting finding. That is, learners who belong to the target group may not achieve the best prediction performance compared to those who do not belong to the target group.

Table 5.13 Proportion of correct predictions (groups belonged to by groups to predict)

Learners	Target group to predict			Total
	Group A	Group B	Group C	
<b>Group A</b>	<b>32.6%</b>	78.7%	29.2%	51.3%
<b>Group B</b>	21.1%	<b>76.6%</b>	18.8%	32.1%
<b>Group C</b>	49.6%	65.6%	<b>20.6%</b>	47.0%
<b>Total</b>	36.9%	74.6%	22.2%	44.5%

This finding is common for all three groups. For example, learners who belonged to Group C were more successful in predicting Group A than learners who belonged to Group A (49.6% vs. 32.6%). Learners who belonged to Group A were more successful in predicting Group B than learners who belonged to Group B (78.7% vs. 76.6%). Learners who belonged to Group A were more successful in predicting Group C than learners who belonged to Group C (29.2% vs. 20.6%).

By informing learners which segments they belonged to, as was done in this experiment, it is possible that learners may be induced to adopt false assumption with regard to predictions about their own group insofar as they need only follow their own instincts to make predictions about their own group. This phenomenon may also explain why, in some open-ended feedback for Approach Four, a few learners argued that they were right and the system was in error.

### 5.3.3.3 Prediction Accuracy on Probabilities of Options

The most important type of prediction is probability prediction. In each training set, all learners were asked to assign 100% across the four options “fly Qantas”, “fly Virgin”, “fly Jetstar” and “not fly” in the set. Percentages can easily be converted to probabilities summing to 1. As discussed in Section 5.3.2, Hellinger Distance (HD) was chosen as the scoring rule for this analysis. In preparing data, the HD was calculated for each training set using learner predicted probabilities and target model approximated probabilities. For each learner, there were 16 HDs for each session. For Approaches One, Two and Three, there were 48 HDs in total, 16 for each session. For Approach Four, there were 32 HDs in total, 16 for S1 and 16 for S2. This data represented the response level data, the main data for analysis. For individual level analysis, three mean HDs were calculated for each learner, one for each session. For Approach Four, only two of three mean HDs were used for individual level analysis.

As shown in Figure 5.2, the right transformation to normalise response level HD data is to use its square root, namely  $\sqrt{HD}$ . This transformation is a monotone transformation and does not change orders of data in the original HDs. To double check it, Figure 5.3 shows that  $\sqrt{HD}$  aligns well with the normal distribution but not so for the original HD. As discussed in Section 5.3.2, the mean comparison and linear regression model were run. Using transformed HD data for regression analysis yields higher statistical fits than from using the original HD data (e.g.  $R^2$  at 0.92, instead of 0.60).

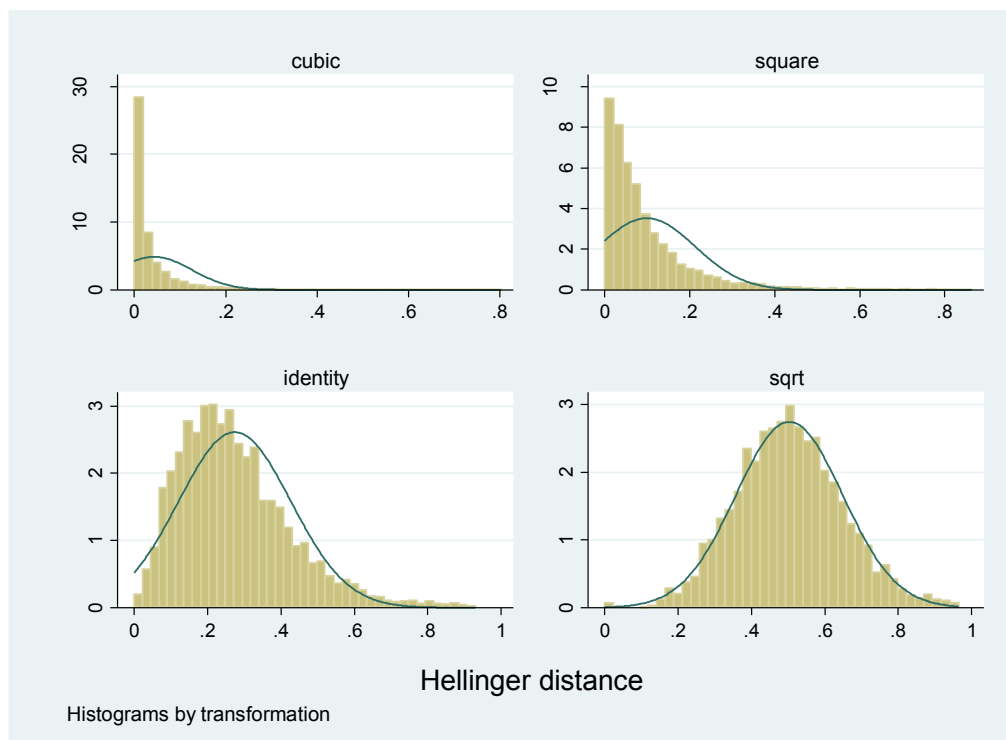


Figure 5.2 Transformations of Hellinger Distances (HDs)

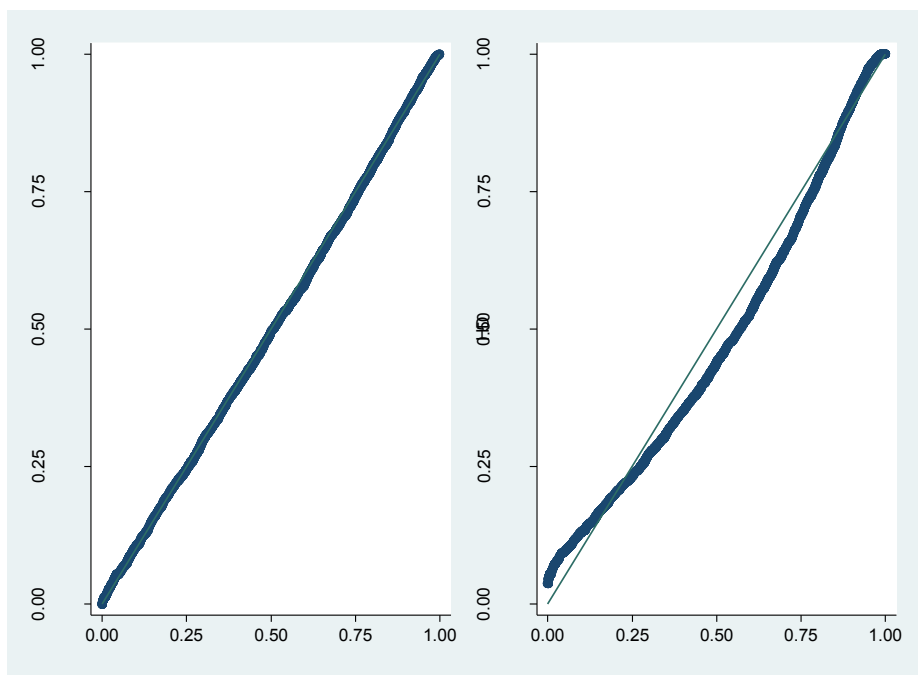


Figure 5.3 Checking square root and original HD against normal distribution

Tables 5.14 and 5.15 provide both mean comparisons of transformed HDs and regression results with transformed HDs as the dependent variable, and a combination of learning approaches and

sessions as the independent variable. Means and coefficients are identical for each combination of learning approaches and training sessions. It should be noted that the results of S0 for Approach Four are not relevant due to the different tasks required in S1 and S2, and are consequently shown as struck out in Table 5.14 and not included in Table 5.15. From these results, one key observation is worth noting; that is, excepting S1 and S2 for Approach Four, in which variances appear to be a little higher than the rest, all other combinations, levels of variances and standard deviations are quite close and highly significant at the  $p = 0.001$  level. With these results, in comparing combinations of learning approaches and sessions, the means or regression coefficients can be compared directly with little concern due to the variances ranging in a narrow band of values.

Table 5.14 Means of transformed HD by learning approaches & sessions

Session & Approach	Mean Sqrt(HD)	N	Std. Deviation	Variance
session 0 (approach 1)	0.525	1008	0.132	0.017
session 1 (approach 1)	0.500	1008	0.137	0.019
session 2 (approach 1)	0.500	1008	0.141	0.020
session 0 (approach 2)	0.529	1008	0.134	0.018
session 1 (approach 2)	0.452	1008	0.139	0.019
session 2 (approach 2)	0.436	1008	0.135	0.018
session 0 (approach 3)	0.533	1008	0.130	0.017
session 1 (approach 3)	0.504	1008	0.146	0.021
session 2 (approach 3)	0.489	1008	0.135	0.018
session 1 (approach 4)	0.525	1008	0.173	0.030
session 2 (approach 4)	0.537	1008	0.167	0.028
Total	0.503	12096	0.145	0.021

Table 5.15 Results of regression analysis with transformed HD as dependent variable

Source	SS	df	MS	Number of obs	11088	
				F(11, 11077)	12464.03	
<b>Model</b>	2811.511	11	255.592	Prob > F	0	
<b>Residual</b>	227.149	11077	0.020506	R-squared	0.9252	
				Adj R-squared	0.9252	
<b>Total</b>	3038.661	11088	0.274049	Root MSE	0.1432	
<b>SQRT(HD)</b>	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
session 0 (approach 1)	0.525	0.005	116.390	0.000	0.516	0.534
session 1 (approach 1)	0.500	0.005	110.810	0.000	0.491	0.509
session 2 (approach 1)	0.500	0.005	110.750	0.000	0.491	0.508
session 0 (approach 2)	0.529	0.005	117.270	0.000	0.520	0.538
session 1 (approach 2)	0.452	0.005	100.160	0.000	0.443	0.461
session 2 (approach 2)	0.436	0.005	96.700	0.000	0.427	0.445
session 0 (approach 3)	0.533	0.005	118.190	0.000	0.524	0.542
session 1 (approach 3)	0.504	0.005	111.720	0.000	0.495	0.513
session 2 (approach 3)	0.488	0.005	108.300	0.000	0.480	0.497
session 1 (approach 4)	0.525	0.005	116.310	0.000	0.516	0.533
session 2 (approach 4)	0.537	0.005	119.040	0.000	0.528	0.546

As shown in Table 5.16, two points can be stated. First, all approaches have the same  $\sqrt{HD}$  in S0 which means learners under all approaches are not statistically different in terms of prediction accuracy before training. Second, the reduction of prediction errors in S1 and S2 are statistically significant. Among all approaches, Approach Two yielded the best outcomes in terms of prediction accuracy. First, the magnitude of improvement in prediction accuracy is the largest, with decremental  $\sqrt{HD}$  at 0.093 (i.e. from 0.529 to 0.436), being more than twice that of the second best performer, Approach Three with decremental  $\sqrt{HD}$  as 0.044 (i.e. from 0.533 to 0.489). Second, improvements in prediction accuracy were continuous from S0 through S1 to S2. Both Approaches Two and Three performed better than Approach One. Results for Approach One showed that although learners improved in S1 compared to S0, there were no further improvements in S2. Results for Approach Four showed a contrary tendency to the other approaches, with prediction accuracy decreasing from S1 to S2.

Table 5.16 Comparing prediction accuracy by learning approaches &amp; sessions

Learning Approach	Session	Sqrt(HD)	Accuracy (HD)
1	0	0.525	-
1	1	0.500	Improved
1	2	0.500	Unchanged
2	0	0.529	-
2	1	0.452	Improved
2	2	0.436	Improved
3	0	0.533	-
3	1	0.504	Improved
3	2	0.489	Improved
4	1	0.525	-
4	2	0.537	Decreased

Alternatively, using the means of HDs from starting and finishing sessions, an analysis was conducted with a model similar to Equation 5.6 ( $\Delta H^{\text{end-start}} = \alpha + \beta H_{\text{start}} + \epsilon$ ). The starting session for Learning Approaches One, Two and Three is S0, and for Approach Four is S1. The end session is S2 for all approaches. This analysis can provide answers as to whether learners improved in the last session compared to the beginning session. It is worth noting that when the means of HDs were calculated for each learner by averaging 16 HDs from 16 tasks, it can be the case that even two HD means are identical, but the actual situations may be quite different. For example, in one case, all individual HDs are close to the mean. In the other case, all individual HDs are dispersed away from, and to either side of the mean. Therefore, the former analysis of response level HDs is a better indicator when ranking learning approaches because it covers all scenarios, treating each prediction task separately. However, analysis for the respondent level is also necessary because it summarises learners' performance. Table 5.17 and Figure 5.4 display regression results and scatter plots based on the regression function.

Table 5.17 A summary of model fits for regression analysis using mean HDs

Source	SS	df	MS	Number of obs	=	252
				<b>F( 8, 244)</b>	=	26.59
<b>Model</b>	0.995	8	0.124	<b>Prob &gt; F</b>	=	0
<b>Residual</b>	1.141	244	0.005	<b>R-squared</b>	=	0.466
				<b>Adj R-squared</b>	=	0.448
<b>Total</b>	2.137	252	0.008	<b>Root MSE</b>	=	0.068
HD_end session	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
approach 1_alpha	0.072	0.035	2.030	0.043	0.002	0.142
approach 2_alpha	0.071	0.036	1.990	0.048	0.001	0.141
approach 3_alpha	0.128	0.036	3.600	0.000	0.058	0.198
approach 4_alpha	0.112	0.027	4.140	0.000	0.059	0.165
approach 1_beta	-0.326	0.117	-2.780	0.006	-0.557	-0.095
approach 2_beta	-0.537	0.116	-4.630	0.000	-0.766	-0.309
approach 3_beta	-0.572	0.114	-5.000	0.000	-0.798	-0.347
approach 4_beta	-0.330	0.084	-3.950	0.000	-0.495	-0.165

As discussed in Section 5.3.2, the ideal value for intercept  $\alpha$  is zero which indicates there are no over or under system predictions. The ideal value for slope  $\beta$  is -1 which indicates that all prediction differences can be corrected. An approach with smaller  $\beta$  and smaller  $\alpha$  is preferred. Table 5.17 shows estimated  $\hat{\alpha}$  and  $\hat{\beta}$  for each approach. Figure 5.4 shows a scatter plot for each approach with regression line using estimated  $\hat{\alpha}$  and  $\hat{\beta}$ . To make comparison easier, the reference line with ideal  $\alpha$  and  $\beta$  is also shown in each scatter plot.



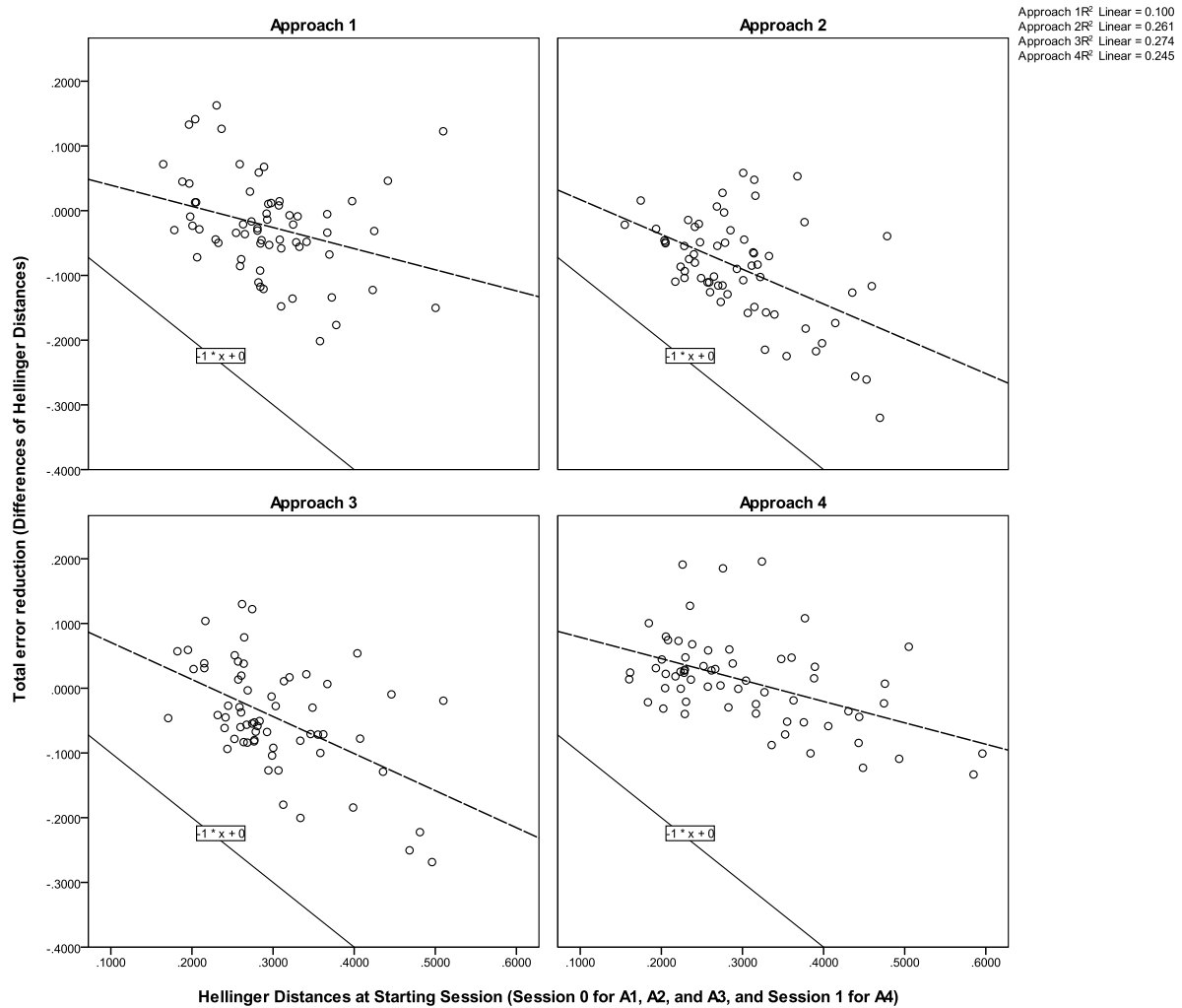


Figure 5.4 Prediction accuracy improvements from starting session to end session by approach

Although Approach Three shows learners' prediction accuracy is improving the fastest, the systematic error captured by  $\hat{\alpha}$  is much larger than with Approaches Two and One. This is clear according to coefficients in Table 5.17 although not so clear pictorially in Figure 5.4. Overall, Approach Two is the most effective approach to improving prediction accuracy fastest with the least systematic error in prediction. The advantage of Approach Three over Approach One is not obvious, because even though learners under Approach Three are improving faster the systematic error component is also largest. One could argue that the large systematic error for Approach Three may be caused by outliers (in Figure 5.4), nevertheless, we cannot conclude decisively without consulting other data because the same point can be made with regard to Approach One.

Approach Four was examined separately because the tasks were different. From the results, it appears that learners under Approach Four were improving but the systematic error component was large.

Table 5.18 demonstrates the proportions of learners who improved their prediction accuracy over the sessions. This was based on smaller mean HDs in a later session than an earlier session because smaller HDs represent more accurate predictions. Three comparisons were included: being S1 over S0 (excluding Approach Four), S2 over S1 and the last session (S2) over the beginning session (either S0 or S1). Examining overall improvement from S0 to S2 (or S1 to S2 for Approach Four), these results show that Approach Two is the best in improving prediction accuracy effectively, because the percentage of learners who improved in the last session over the first session was the highest at 89% among all approaches (i.e., 89% of learners had lower mean HDs in S2 than S0). Approach Three is slightly better than Approach One (70% versus 67%). Approach Four is less effective in improving the prediction accuracy of learners.

Table 5.18 Proportions of learners who improved their prediction accuracy

	Learners	Session 1 over Session 0	Session 2 over Session 1	Overall
Learning Approach 1	63	67%	51%	67%
Learning Approach 2	63	89%	67%	89%
Learning Approach 3	63	62%	71%	70%
Learning Approach 4	63	NA (different tasks)	40%	40%

Another finding regarding Approach Four is that levels of prediction accuracy for probabilities are different depending on which target segment is predicted. This finding is similar to the previous findings in predicting consumer groups as discussed in Section 5.3.3.2. Table 5.19 shows that average transformed HDs for target consumer groups are different. Prediction accuracy is the highest among learners who predicted Group B. Group B is followed by Group A and Group C

with regard to accuracy. This order is the same as the order discovered in predicting consumer groups as shown in Table 5.12.

Table 5.19 Average transformed HDs by target segments for Approach Four

		N	Mean	Std. Deviation	Minimum	Maximum
<b>Group A</b>	1 Session 1	21	0.538	0.083	0.430	0.702
	2 Session 2	21	0.567	0.072	0.452	0.721
<b>Group B</b>	1 Session 1	21	0.505	0.078	0.400	0.711
	2 Session 2	21	0.505	0.077	0.402	0.754
<b>Group C</b>	1 Session 1	21	0.592	0.092	0.456	0.772
	2 Session 2	21	0.598	0.064	0.499	0.703

### 5.3.4 Summary of Hypotheses H1a, H2a and H3a

This section summarises the results and provides answers to research hypotheses H1a, H2a and H3a. Hypothesis H1a assumes that Approach Two is more effective than Approach One in improving prediction accuracy. Hypothesis H2a assumes that Approach Three is more effective than Approach One using the same measure. Hypothesis H3a treats Approach Four separately and assumes it helps learners improve their prediction accuracy without comparing it to other learning approaches.

H1a states that:

Learners who receive outcome feedback after each training task (Approach Two) make more accurate probability predictions than those who perform self-regulated learning using a “plain vanilla” MDSS (Approach One).

Based on the results discussed in Section 5.3.3, learners using Approach Two were more accurate predictors than learners using Approach One, and they reached the highest level of accuracy among all four approaches in session S1 and maintained that same level of accuracy in session S2. These results were also shown to be statistically significant in results such as Table 5.15. Based on the full probabilities of all options, which is the focus of this hypothesis, Approach Two is more effective than Approach One in two respects: it is more accurate in both S1 and S2 based on all

predictions made, and there are more learners who improved their prediction accuracy in later sessions than in earlier sessions. Overall, Approach Two is a more effective approach than Approach One in every respect with regard to improving prediction accuracy.

H2a states that:

Learners who receive diagnosis of their own model after training tasks (Approach Three) make more accurate predictions than those who perform self-regulated learning using a “plain vanilla” MDSS (Approach One).

The results clearly show that Approach Two is the most effective approach among all approaches in helping learners improve their prediction accuracy. It is not so clear with regard to Approach Three. Regarding the first type of prediction for the preferred option, Approach Three performed at a similar level as Approach One based on total numbers of correct predictions made. Approach Three performed better than Approach One with more learners improving their prediction accuracy in S2 than in S1 (see Table 5.10). For the full probability predictions, Approach Three is more effective in improving prediction accuracy based on all predictions made (see Table 5.16). However, in terms of the total number of learners who improved their accuracy in the last session over the first session, Approach Three is close to Approach One. Overall, Approach Three is more effective but its advantage over Approach One is much less than over Approach Two.

H3a states that:

Learners who receive class information and classification feedback after training tasks (Approach Four) improve their predictions of probabilities matching a particular class with more tasks and feedback given in training.

Due to task differences, Approach Four was not compared with other learning approaches. This hypothesis is more exploratory in nature to ascertain whether by giving learners class information

as feed-forward information and providing classification feedback after each task can induce learners to improve their prediction accuracy. As discussed, besides full probability predictions, two other types of predictions were also made; a direct prediction for the “preferred” option and an implicit prediction for the consumer group. One observation is that learners’ predictions vary greatly under this learning approach. When a result is based on the response level, it shows prediction accuracy was not improved in session S2 compared to S1, excepting the prediction accuracy for the consumer group improved slightly (see Table 5.11). When a result is based on the respondent level, it shows a good number of learners actually improved. For example, 50.8% of learners improved their prediction accuracy for the “preferred” option in S2 (Table 5.10) and 40% of learners improved their prediction accuracy for full probabilities in S2 (Table 5.18). Depending on the consumer groups assigned for prediction purposes, and whether the learners belonged to the same group, the learners’ implicit predictions for target consumer groups can vary greatly. Learners’ accuracy in predicting full probabilities also vary greatly depending on the target segment. To conclude, under Approach Four the prediction accuracy varied based on target segments and individual differences. Some learners improved their accuracy and others did not.

### **5.3.5 Testing Target Model Parameter Learning**

#### **5.3.5.1 The Nature of Testing Model Parameter Learning**

Testing whether learners understand the attribute parameters in a model from predictions they made in tasks, is more challenging than testing prediction accuracy from the same data. Unlike the testing of prediction accuracy, prediction data alone does not tell us whether learners understand the underlying target parameters. To make the task even more difficult in this experiment, it is not a static problem because learners completed multiple sessions, so the actual question is not only about testing the model understanding for a particular session, but testing whether and how fast learners improved their understanding over more than one session, from the point where they

started before the learning sessions. Moreover, several learning approaches need to be compared in this process. There are no existing methods or models, enabling us to solve this problem, in either the learning or the DCM literature. In this respect, the analysis is unique in terms of using the limited data in hand, which comprises two probabilities, learner predicted and target model approximated probabilities, and the experimental design matrix covering all attributes of options for prediction.

One can argue that people cannot articulate the exact values of attribute parameters they have in mind when making predictions, even though they may be able to describe them in general terms. For example, some learners provided feedback that consumers always chose the cheapest fare. Knowledge of attribute parameters is implicit and it takes effect jointly when in the process of making predictions. When learners are learning target model parameters, the knowledge of attributes keeps changing with the more information received. At any particular time, a learner's knowledge of an attribute parameter is a mixture of the learner's existing knowledge plus updates from newly received information.

In the session prior to learning the target model ( $S_0$ ), learners made their own choices with regard to flight options but also made predictions of what they believed consumers would choose. In this process, they relied on their own knowledge to make predictions and make choices. It is reasonable to believe that their judgements in these two processes should be quite similar whether the task is to predict or to choose. Assume the learners' parameters of an attribute form a distribution of some type. To make it simpler for discussion, we can assume this distribution is normal. In this context, a key location parameter is a mean which can be used to represent the estimated parameter in an aggregated model. Please note, this example is only used to explain the idea of learners converging to a fixed parameter. It is not suggesting that a coefficient in a target model has to be

the mean of individual parameters. Depending on responses, experimental designs, distributions of individual parameters and other factors, the coefficient can be any value relating to the parameter distribution.

The distribution for attribute parameters in the starting state is shown in red in Figure 5.5. In the learning experiment, the estimated fixed-effect model was used as the target model to train learners. For each independent attribute term, the target model coefficient was the mean (or other value) identified from the prior learning parameter distribution. Using different learning approaches, the experiment was designed to make learners converge to this fixed parameter. This is shown in Figure 5.5.

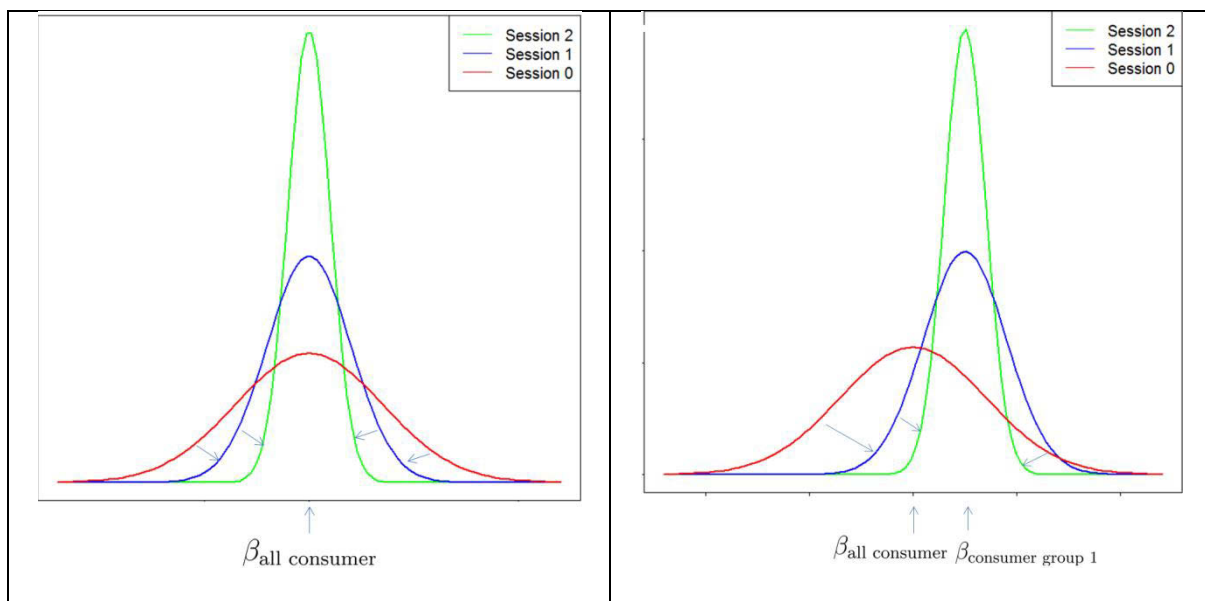


Figure 5.5 Model learning process – converging to a fixed parameter

New distributions of learners' individual parameters in this converging process are shown in blue and green for S1 and S2. This figure demonstrates the expected behaviour of learners if converging to the target fixed parameter. For some learning approaches or attribute term, learning may not occur, in which case this convergence process will not be observed. The second graph in Figure 5.5 shows the situation if the target model is not a model covering all individuals but a segment of individuals. In summary, informing learners to learn from a fixed-effects model is to train learners

to essentially converge to fixed parameters. If learning is effective, we should observe that the differences between the learners' own parameters and target model parameters for attributes keep reducing, eventually approaching zero for no differences.

The forgoing process is not observable because neither changes of learner parameters nor differences of learner and target model parameters are directly observable. Inferences must be made from the data and some initial analysis results. The focus of the initial analysis is to identify differences between the learner and target model parameters for each of the attribute terms. For each attribute term, initial analysis should provide a coefficient showing how large or small is the difference between learner and target model parameters. The estimated coefficients from the initial analysis can then be used as the new data points for follow-up analysis summarising and comparing overall differences between learning approaches and sessions, to make inferences about whether learners have improved convergence to the target parameter with more training tasks.

The initial analysis is basically a regression analysis with attributes as independent variables and differences of learner and model as the dependent variable. This analysis was conducted on an individual level, separated by sessions. For each learner, there should be three sets of attribute coefficients, one for each session. The dynamic problem of whether learners have improved by reducing differences in later sessions is not answered by initial analysis. In the follow-up analysis, the objective is to use coefficients from the initial analysis to operate a model and ascertain whether these coefficients fluctuate from one session to another. This analysis solves the abovementioned dynamic problem and answers research hypotheses H1b to H3b for the target model parameter learning. In the initial analysis, coefficients of attribute terms represent the contributions of attributes to the differences between learner and model parameters. Therefore, the smaller the size of an attribute coefficient, the lesser is the role it plays in causing differences between learners and



the target model. In the following section, models for the initial analysis and follow-up analysis will be given and discussed.

### 5.3.5.2 Models and Analysis Methods in Testing Model Parameter Learning

In the initial analysis, differences between learner and model probabilities are regressed against independent attribute terms in the designed matrix. This analysis provides key information needed for follow-up analysis; that is, how much each attribute term contributes to differences between subjective predictions and the target model.

In calculating differences between the two, instead of calculating raw probabilities, differences were calculated based on log-odds. Odd and log-odd are basic concepts in categorical analysis models such as MNL and log-linear model (Agresti 2002). For example, following IIA principle for MNL, the odd of choosing A over choosing B is a function of the odd of utility A over utility B as shown in Equation 5.11.

$$\frac{P_A}{P_B} = \frac{U_A}{U_B} = \frac{E^{V_A}}{E^{V_B}} \quad (5.11)$$

In psychological learning literature, the generalised matching law model developed by Baum (1974) also adopts log-odds to identify the relationship between responses and reinforcements. The model basically states that the odd of observing A over B is proportional to the odd of reinforcing A over B:

$$\frac{A}{B} = k \left( \frac{R_A}{R_B} \right)^a \text{ or } \log \frac{A}{B} = \log K + a * \log \frac{R_A}{R_B} \quad (5.12)$$

In this model, if slope term  $a$  and constant  $K$  are both equal to one, then there is a perfect match between responses and reinforcements.

Supported by ideas from models such as the above, the initial analysis model can be developed. Starting with the following notations, the basic terms used in this model can be defined as:

$P_{ic}^t$  - Probability of alternative  $i$  predicted by a learner in choice sets  $c$  for training session  $t$

$P_{jc}^t$  - Probability of alternative  $j$  predicted by a learner in choice sets  $c$  for training session  $t$

$P_{ic}^{tm}$  - Probability of alternative  $i$  given by target model in choice sets  $c$  for training session  $t$

$P_{jc}^{tm}$  - Probability of alternative  $j$  given by target model in choice sets  $c$  for training session  $t$

“Fly” options in experimental design (i.e. fly Qantas, fly Virgin and fly Jetstar) can be regarded as alternative  $i$  and the “not fly” option can be regarded as alternative  $j$ . In each set, the log-odd of probability of any  $i$  over probability of  $j$  can be calculated for both the learner predicted probabilities and the target model approximated probabilities separately, as shown in Equations 5.13 and 5.14.

$$\log \frac{P_{ic}^t}{P_{jc}^t} = (X_{ic}^t - X_{jc}^t)\beta_c^t + \epsilon_{ic}^t \quad (5.13)$$

$$\log \frac{P_{ic}^{tm}}{P_{jc}^{tm}} = (X_{icm}^t - X_{jcm}^t)\beta_c^{tm} + \epsilon_{ic}^{tm} \quad (5.14)$$

In training sets, the attribute levels for any alternative  $i$  are known *a priori* and they are indifferent to learner probabilities or model probabilities, so  $X_{ic}^t = X_{icm}^t$  is true. Attributes for the “not fly” option are always coded as 0, so  $X_{jc}^t = X_{jcm}^t = 0$  is true. If Equation 5.14 is subtracted from Equation 5.13, we have a function with differences of learner and model on “fly” options as the dependent variable, and the experimental design matrix for attributes as independent variables:

$$\log \frac{P_{ic}^t}{P_{jc}^t} - \log \frac{P_{ic}^{tm}}{P_{jc}^{tm}} = X_{ic}^t(\beta_c^t - \beta_c^{tm}) + (\epsilon_{ic}^t - \epsilon_{ic}^{tm}) \quad (5.15)$$

Log-odd differences, or the first component, may be simplified as  $\widetilde{Y}_{ic}^t$ . The term in the first bracket  $\beta_c^t - \beta_c^{tm}$  can be simplified as  $\Delta\beta_c^t$ , and  $\epsilon_{ic}^t - \epsilon_{ic}^{tm}$  can be simplified as  $\epsilon_{ic}^t$ . This can be further simplified as Equation 5.16:

$$\widetilde{Y}_{ic}^t = X_{ic}^t\Delta\beta_c^t + \epsilon_{ic}^t \quad (5.16)$$

In other words, it states that the difference between learner and model is a function of attribute design matrix, and the coefficients of attributes represent the differences between learners' values of attributes and the target model's value of attributes. In estimating this model, analysis shall give  $\widehat{\Delta\beta^t}$  consisting of true parameter  $\Delta\beta^t$  and errors. It is important to process this analysis for each individual learner, separately for each of the three training sessions. The results provide individual levels estimated as  $\widehat{\Delta\beta^t}$  for each of the three sessions:  $\widehat{\Delta\beta^0}$  for S0,  $\widehat{\Delta\beta^1}$  for S1 and  $\widehat{\Delta\beta^2}$  for S2. Since there are nine independent attribute terms in the target models (e.g. Qantas and \$400 fare), for any session  $t$ , there should be nine coefficients in  $\widehat{\Delta\beta^t}$  or  $\widehat{\Delta\beta_x^t}(\widehat{\Delta\beta_1^t}, \widehat{\Delta\beta_2^t}, \widehat{\Delta\beta_3^t}, \dots, \widehat{\Delta\beta_9^t})$ .

After this initial analysis, follow-up analysis should focus on testing whether learners have an improved understanding of attribute parameters from the beginning session to the end session, or find a function to define the relationship between  $\widehat{\Delta\beta_x^0}$  of S0,  $\widehat{\Delta\beta_x^1}$  of S1 and  $\widehat{\Delta\beta_x^2}$  of S2. If learning of a model parameter is improving, the contribution or effect size of this attribute on the learner and model differences should decrease, which means  $\widehat{\Delta\beta_x^2}$  is smaller than  $\widehat{\Delta\beta_x^1}$  and  $\widehat{\Delta\beta_x^1}$  is smaller than  $\widehat{\Delta\beta_x^0}$ . Generally speaking,  $\widehat{\Delta\beta_x^t}$  for an attribute term  $x$  is expected to decrease if training continues indefinitely, and eventually it should approach zero. Approaching zero means that learners' understanding of a model parameter is almost perfect so this attribute term has no further impact on the differences of learners and the target model.

It is easier to demonstrate the follow-up analysis using just two hypothetical sessions,  $t1$  and  $t2$ . For an attribute term  $x$  in a target model for learning, the difference of coefficients for two sessions  $\Delta\beta_x^{t1}$  and  $\Delta\beta_x^{t2}$  can be denoted as  $\Delta\beta_x^{t2-t1}$ . We consider  $\Delta\beta_x^{t2-t1}$  as a function of  $\Delta\beta_x^{t1}$ , as shown in Equation 5.17:

$$\Delta\beta_x^{t2-t1} = \Delta\beta_x^{t2} - \Delta\beta_x^{t1} = \alpha + \gamma\Delta\beta_x^{t1} + \epsilon \quad (5.17)$$

Using individual level estimated  $\Delta\beta_x^{t1}$  and  $\Delta\beta_x^{t2-t1}$  for analysis, regression analysis following Equation 5.17 should generate an estimated intercept  $\hat{\alpha}$  and estimated slope  $\hat{\gamma}$ . The remaining problems are to know the ideal values and possible ranges for these two terms to enable us to compare values of these two terms across learning approaches. It is not difficult to see that the best value for  $\Delta\beta_x^{t2}$  is zero; intuitively, it means in session  $t2$ , there is no difference between learners and the target model regarding this attribute. If  $\Delta\beta_x^{t1}$  is not zero and  $\Delta\beta_x^{t2}$  is zero, it means learners are able to correct any differences regarding the target attribute in one single step (session  $t2$  over session  $t1$ ). Continuing Equation 5.17, we have:

$$\Delta\beta_x^{t2} = \alpha + (\gamma + 1)\Delta\beta_x^{t1} + \epsilon$$

$$\text{if } \Delta\beta_x^{t2} = 0, \text{ then } \alpha + (\gamma + 1)\Delta\beta_x^{t1} + \epsilon = 0, \alpha = 0 \text{ and } \gamma = -1 \quad (5.18)$$

Clearly, the ideal  $\alpha$  should be zero and ideal  $\gamma$  should be -1. This researcher terms this ideal scenario as a “one step clearance”. It means that virtually all differences between the learner and model regarding this attribute are cleared. Although this is highly unlikely, it gives ideal values for  $\alpha$  and  $\gamma$ . Please note, the above discussion was made in expected values, and  $\epsilon$  terms in Equations 5.17 and 5.18 were assumed to be 0.

If an estimated  $\hat{\alpha}$  is not zero, it suggests that there is a systematic misunderstanding of the target attribute parameter. If this is the case, even if we have an ideal  $\hat{\gamma}$  close to -1, it means learners are learning quickly but unfortunately converging to a wrong value for the attribute with systematic misunderstanding of the target parameter. Rapid speed learning should see  $\hat{\gamma}$  close to -1. If learning is ongoing, the value range should be  $|\gamma + 1| < 1$  or  $-2 < \gamma < 0$ . If  $\gamma$  is outside this range, there is no learning. In comparing two estimated  $\gamma$  for two learning approaches, the better learning approach should always have an estimated  $\gamma$  closer to -1.

In estimating the follow-up model as denoted in Equation 5.17, it is worth noting that the independent variable is not a directly observed variable but individual coefficients estimated from the initial analysis. As Hausman (2001) pointed out, in estimating the linear regression model in which independent variables are not direct observed variables but indirect variables with errors, often a downward phenomenon can be observed. This means that estimated coefficients may have smaller magnitudes than true parameters. This researcher understands the limitation that using estimated parameters  $\widehat{\Delta\beta^t}$  from the initial analysis contains estimation errors that may influence the precision of estimated  $\alpha$  and  $\gamma$  in the follow-up analysis. However, since the key interest is to determine overall which approach is more effective in helping learners to learn target model parameters, errors in data may reduce the efficiency of the follow-up estimation but should not alter the order of approaches in the results. In other words, estimated  $\hat{\alpha}$  and  $\hat{\gamma}$  may not be the most precise measures but arguably errors in data (initial analysis parameters) should influence all approaches equally, therefore findings related to the order of approaches in terms of effectiveness should hold. To further validate the assumption that the chosen analysis method gives consistent estimators of  $\alpha$  and  $\gamma$ , this researcher compared the results when all attributes were estimated together in one model and separately in different models. Both approaches give almost identical  $\hat{\alpha}$  and  $\hat{\gamma}$  for every attribute. Therefore, the method given in Equation 5.17 should yield consistent estimation results empirically.

To further validate the results regarding orders of approaches, all nine attribute terms in the  $\widehat{\Delta\beta^t}$  vector may be combined to yield a single indicator for overall comparison between learning approaches and sessions. This indicator may be calculated by using a Euclidean norm to obtain a single distance measure covering all nine attributes as shown in Equation 5.19:

$$\|\Delta\beta_{all}^t\| = \sqrt{\sum_{x=1}^9 (\Delta\beta_x^t)^2} \quad (5.19)$$

For each respondent, distance measures can be calculated this way covering all three sessions. They are  $\|\Delta\beta_{all}^{s0}\|$ ,  $\|\Delta\beta_{all}^{s1}\|$ ,  $\|\Delta\beta_{all}^{s2}\|$ . Regression analysis can be processed with this overall indicator as dependent variable, and learning approaches and training sessions as independent variables as show in Equation 5.20:

$$\|\Delta\beta_{all}^t\| = \beta * X_{approach*session} + \epsilon \quad (5.20)$$

The results of this analysis may be used to validate the findings from the attribute by attribute analysis following the model in Equation 5.17, to ascertain whether a conclusion as to effectiveness, drawn from summarising the separate orders, agrees with the conclusion from this analysis.

### 5.3.6 Results – Target Model Parameter Learning

#### 5.3.6.1 Initial Analysis - Individual Level Coefficients $\widehat{\Delta\beta^t}$

As discussed in Section 5.3.5, the first step in the whole process was to conduct an analysis based on Equation 5.16 to obtain three sets of coefficients,  $\widehat{\Delta\beta^t}$  for S0, S1 and S2. Each set of  $\widehat{\Delta\beta^t}$  includes nine coefficients, one for each independent attribute term. Figure 5.6 employs nine histograms to portray the distributions of individual coefficients for each attribute term for each starting session (S0 for Approaches One, Two and Three, and S1 for Approach Four). Figure 5.7 shows the distributions for each end session (S2 for all approaches).

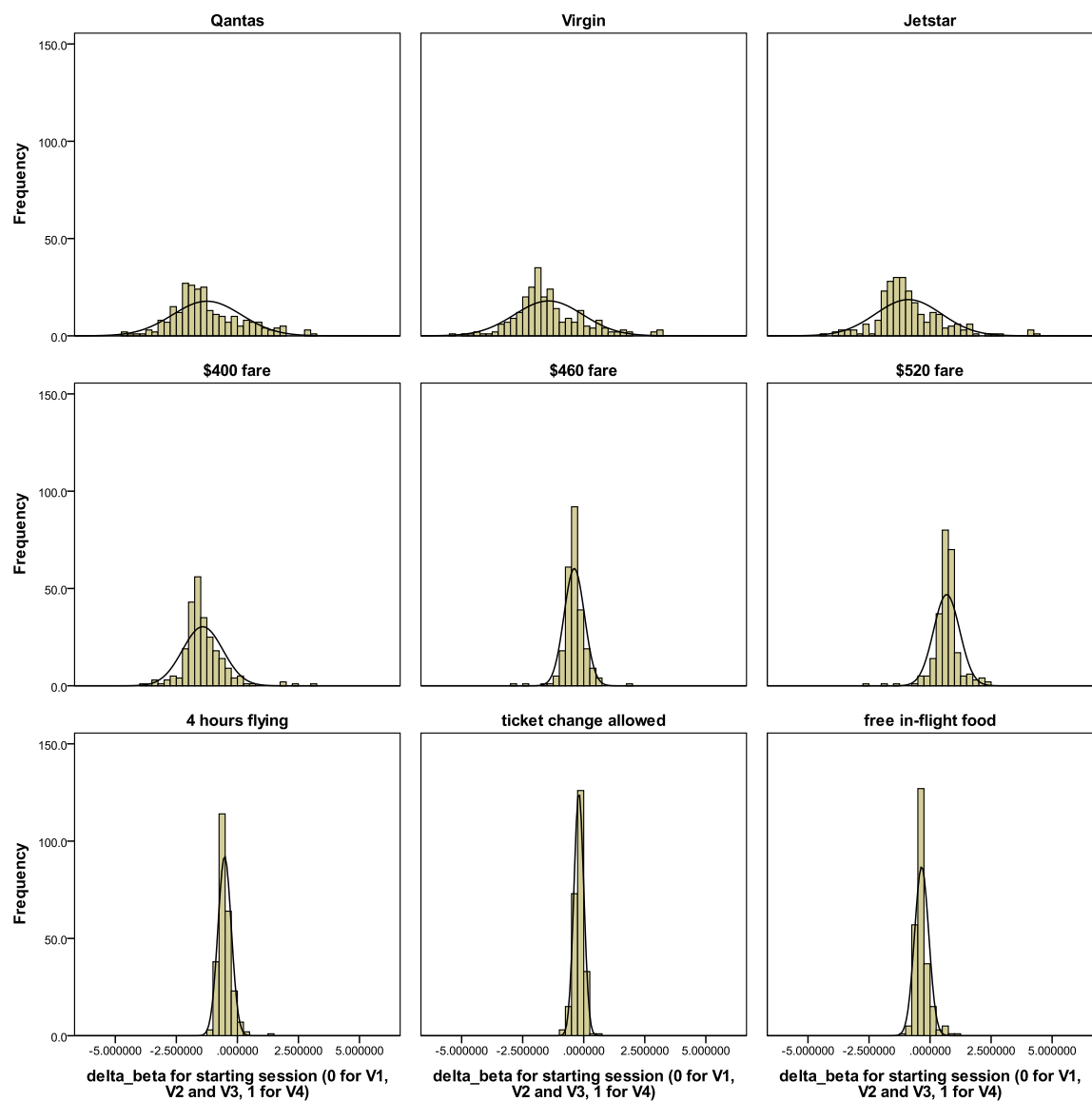


Figure 5.6 Distributions of individual coefficients of attributes for starting session

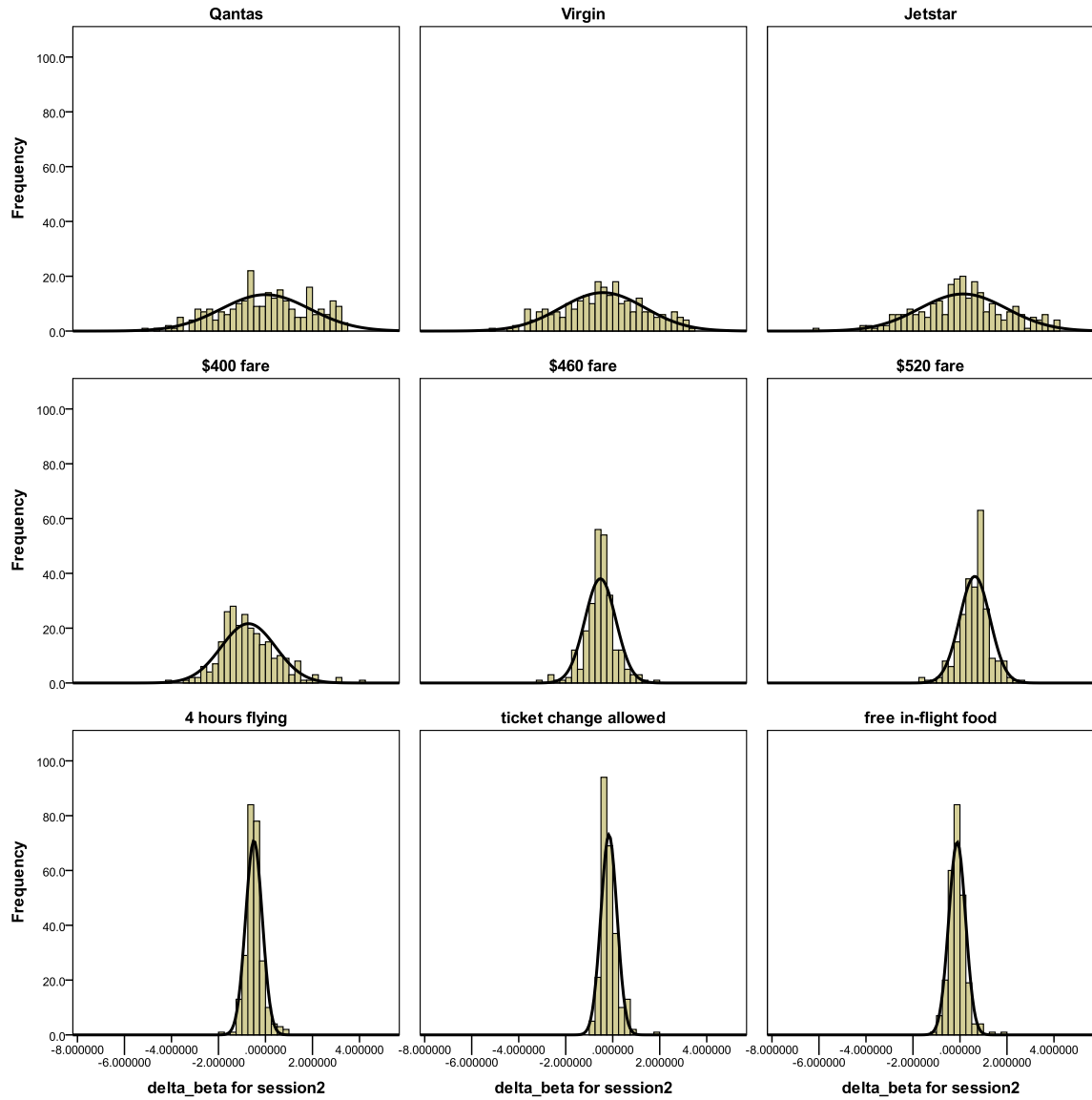


Figure 5.7 Distributions of individual coefficients of attributes for end session

Two observations can be made by inspection of these histograms: first, although the coefficients are not exactly normally distributed, in general, there is a clear central tendency for the individual coefficients to concentrate around the distribution mean. Therefore by comparing means, it is possible to initially obtain coarse comparisons between starting and end sessions to ascertain whether the differences in attributes of learner models and target model, have decreased in the last session following training. Second, the variances of coefficients for attribute terms are not the same. This means, the extent of differences for learner models and the target model are different



for attributes. For example, the coefficients of three airlines have much larger ranges than coefficients of flying time, ticket change and in-flight food, suggesting learners differ more in judging some attributes than other attributes.

Table 5.20 Means and standard deviations of individual coefficients for starting and end sessions

		N	Mean	Std. Deviation
<b>Qantas</b>	Starting Session	252	<b>-1.247</b>	1.413
	End Session	252	<b>0.002</b>	1.892
<b>Virgin</b>	Starting Session	252	<b>-1.438</b>	1.402
	End Session	252	<b>-0.419</b>	1.792
<b>Jetstar</b>	Starting Session	252	<b>-0.882</b>	1.350
	End Session	252	<b>0.136</b>	1.857
<b>\$400 fare</b>	Starting Session	252	<b>-1.420</b>	0.828
	End Session	252	<b>-0.726</b>	1.161
<b>\$460 fare</b>	Starting Session	252	<b>-0.388</b>	0.417
	End Session	252	<b>-0.526</b>	0.661
<b>\$520 fare</b>	Starting Session	252	<b>0.678</b>	0.536
	End Session	252	<b>0.631</b>	0.646
<b>4 hours flying</b>	Starting Session	252	<b>-0.522</b>	0.273
	End Session	252	<b>-0.491</b>	0.355
<b>ticket change allowed</b>	Starting Session	252	<b>-0.199</b>	0.201
	End Session	252	<b>-0.160</b>	0.344
<b>free in-flight food</b>	Starting Session	252	<b>-0.344</b>	0.289
	End Session	252	<b>-0.119</b>	0.357

Table 5.20 provides a summary of means and standard deviations of individual coefficients from the initial analysis. Although it may be observed that differences between learner models and target models are decreasing, as shown by the reduced means of coefficients, to test whether learners improved their understanding of target model parameters over the sessions and to compare the effectiveness of learning across approaches, more rigorous follow-up analysis was conducted using these individual coefficients.

### 5.3.6.2 Follow-up Analysis - Using $\widehat{\Delta\beta^t}$ to Test Model Parameter Learning

Using sets of individual coefficients,  $\widehat{\Delta\beta_x^0}$  for S0,  $\widehat{\Delta\beta_x^1}$  for S1 and  $\widehat{\Delta\beta_x^2}$  for S2, regression analysis using Equation 5.17 was conducted. On the following pages, each of the nine attribute terms are discussed with the results portrayed using graphs and tabular data. In the analysis discussed below, the “starting” session refers to S0 for Approaches One, Two and Three and S1 for Approach Four. The “end” session refers to S2 for all four approaches. The model uses the differences between coefficients of the end session and starting sessions as the dependent variable and coefficients of the starting session as the independent variable (end session over starting session). The model summarises the model parameter learning over the whole period of learning therefore it is a key indicator for comparisons between approaches. In the following sections, any terms with “\*\*\*” are significant at the  $p = 0.01$  level, any terms with “\*\*” are significant at the  $p = 0.05$  level and any terms with “\*” are significant at the  $p = 0.1$  level. To make comparisons easier to observe in the graphs, a reference line of the ideal formula with  $\hat{\alpha}$  equals 0 and  $\hat{\gamma}$  equals -1 is shown in each graph as the solid line. The actual regression line of data (individual coefficients) is shown by dotted line.

#### Qantas

Figure 5.8 demonstrates four scatter plots with the  $x$  axis showing individual coefficients of Qantas in the starting session and the  $y$  axis showing differences of individual coefficients between the end and starting sessions. The same results are tabulated in Table 5.21. From the results, it is clear that Approach Three performs best in terms of model parameter learning with low  $\hat{\alpha}$  and  $\hat{\gamma}$  close to -1. Approach Two is the second best performer which is less effective than Approach Three but much better than the other two approaches. Approach Four reflects very slow improvement in model learning but at least the improvement is in the right direction (between -1 to 0) and the systematic error is also low. The worst performer is Approach One with a large and significant systematic error component and slope which is in the wrong direction showing declining

performance with positive  $\hat{\gamma}$  (between 0 and 1). The ranks for the four approaches are provided in Table 5.21.

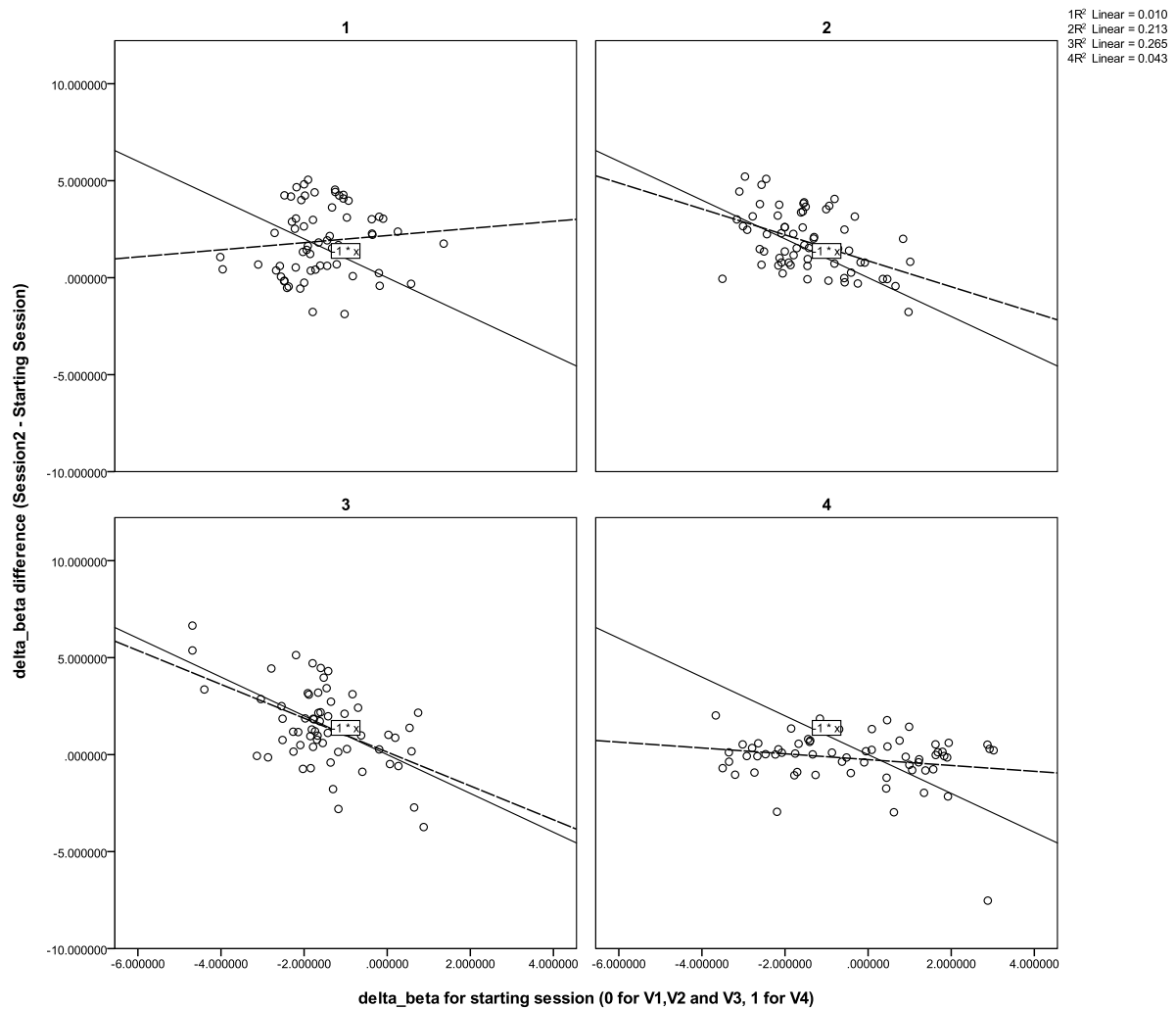


Figure 5.8 Difference between end and starting coefficients, over starting coefficient (Qantas)

Table 5.21 “Qantas” parameter learning by learning approach

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	0.184	2.173***	4
Learning Approach 2	-0.669***	0.870***	2
Learning Approach 3	-0.873***	0.131	1
Learning Approach 4	-0.150*	-0.258	3

## Virgin

Figure 5.9 and Table 5.22 demonstrate the results for Virgin. Approach Two is clearly the most effective approach to improved learning of the target model parameter of Virgin, with low systematic error and close to -1 slope. It is followed by Approach Three with low systematic error in  $\hat{\alpha}$  and  $\hat{\gamma}$  a bit further from -1 than was the case in Approach Two. These two approaches are followed by Approach Four, with higher systematic error and slower pace of improvement. Again, Approach One is the worst of the four with high systematic error in  $\hat{\alpha}$  and a slope  $\hat{\gamma}$  very distant from -1 (it is positive therefore in reverse direction).

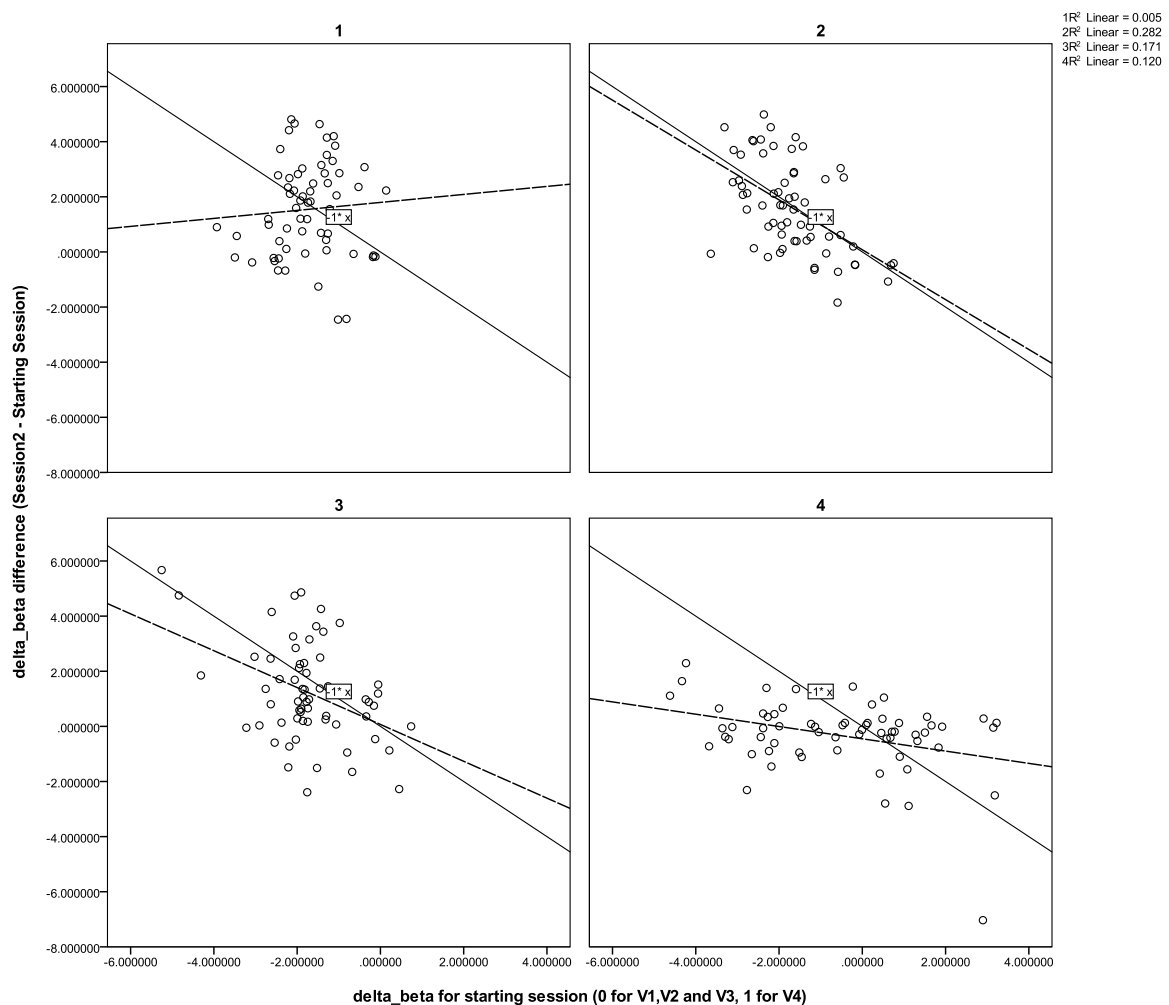


Figure 5.9 Difference between end and starting coefficients, over starting coefficient (Virgin)

Table 5.22 “Virgin Australia” parameter learning by learning approach

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	0.145	1.796***	4
Learning Approach 2	-0.905***	0.079	1
Learning Approach 3	-0.669***	0.070	2
Learning Approach 4	-0.223*	-0.450***	3

### Jetstar

Figure 5.10 and Table 5.23 demonstrate the results for Jetstar. Although systematic error is slightly higher than Approach Three, parameter learning speed is much faster in Approach Two. Again, Approach Four ranks third with a slow learning speed occasioned by lower systematic errors. Approach One performs the worst with high systematic error and a slope quite distant from -1.

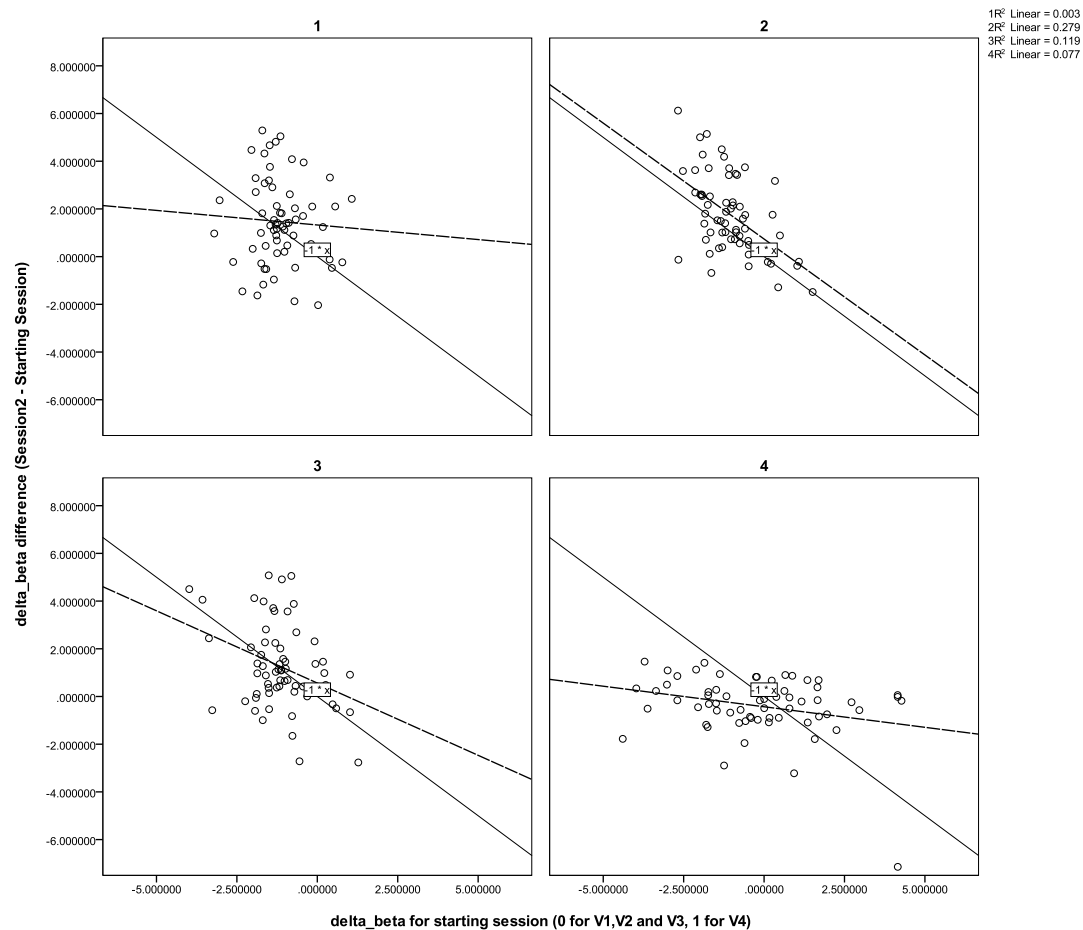


Figure 5.10 Difference between end and starting coefficients, over starting coefficient (Jetstar)

Table 5.23 “Jetstar” parameter learning by learning approach

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	0.122	1.327***	4
Learning Approach 2	-0.972***	0.738***	1
Learning Approach 3	-0.606***	0.563*	2
Learning Approach 4	-0.172**	-0.431***	3

### \$400 fare

Figure 5.11 and Table 5.24 demonstrate the results for the \$400 fare. Approach Two remains the best performer with low systematic error and a slope closest to -1. Approach Three performs poorly for this attribute with the highest system error and a slope quite distant from -1. It ranks last in this case. For Approaches One and Four, although the slopes are similar at -0.266 and -0.201 respectively, Approach Four has the lowest systematic error for this attribute, thus Approach Four ranks second and Approach One ranks third.

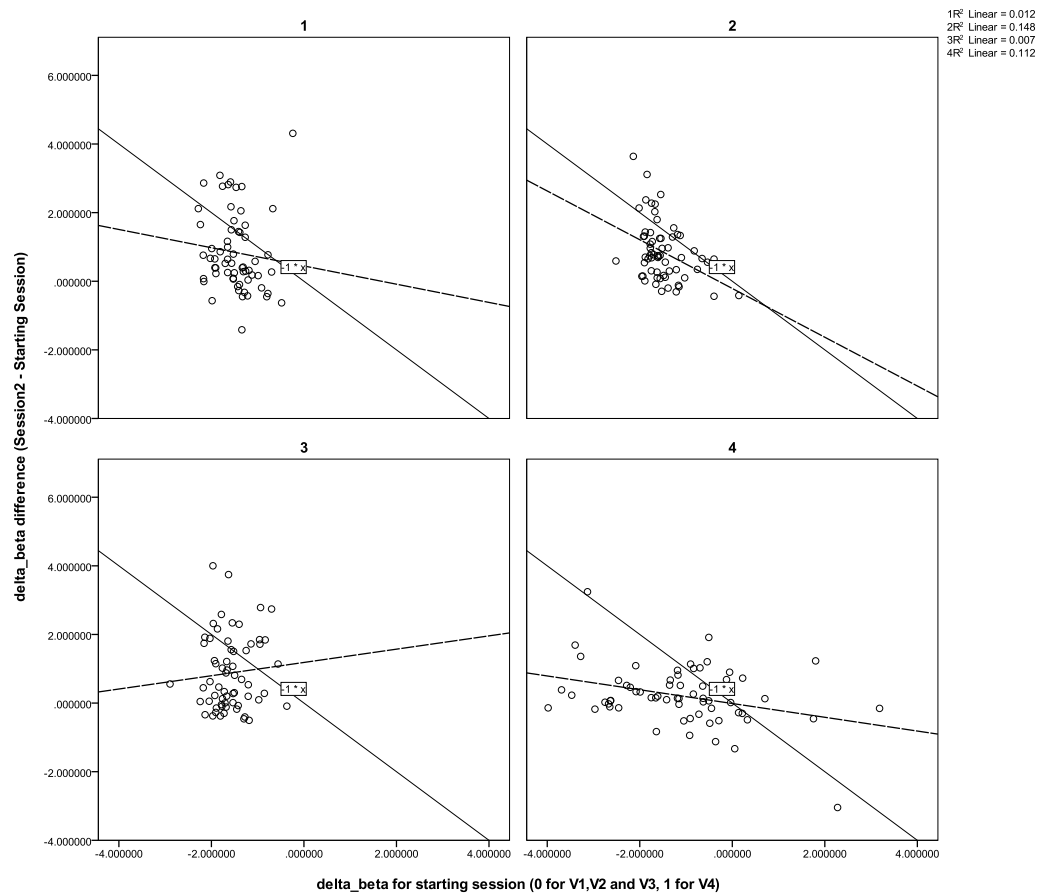


Figure 5.11 Difference between end and starting coefficients, over starting coefficient (\$400 fare)

Table 5.24 “\$400 fare” parameter learning by learning approach

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	-0.266	0.446	3
Learning Approach 2	-0.710***	-0.211	1
Learning Approach 3	0.194	1.184**	4
Learning Approach 4	-0.201***	-0.013	2

### \$460 fare

Figure 5.12 and Table 5.25 demonstrate the results for the \$460 fare. There are no significant differences among the approaches regarding systematic errors therefore we can compare slopes directly. Approach Two remains the best performer with a slope closest to -1. Approach Three is the second best performer with a slope next closest to -1. Approach One ranks third and Approach Four ranks fourth.

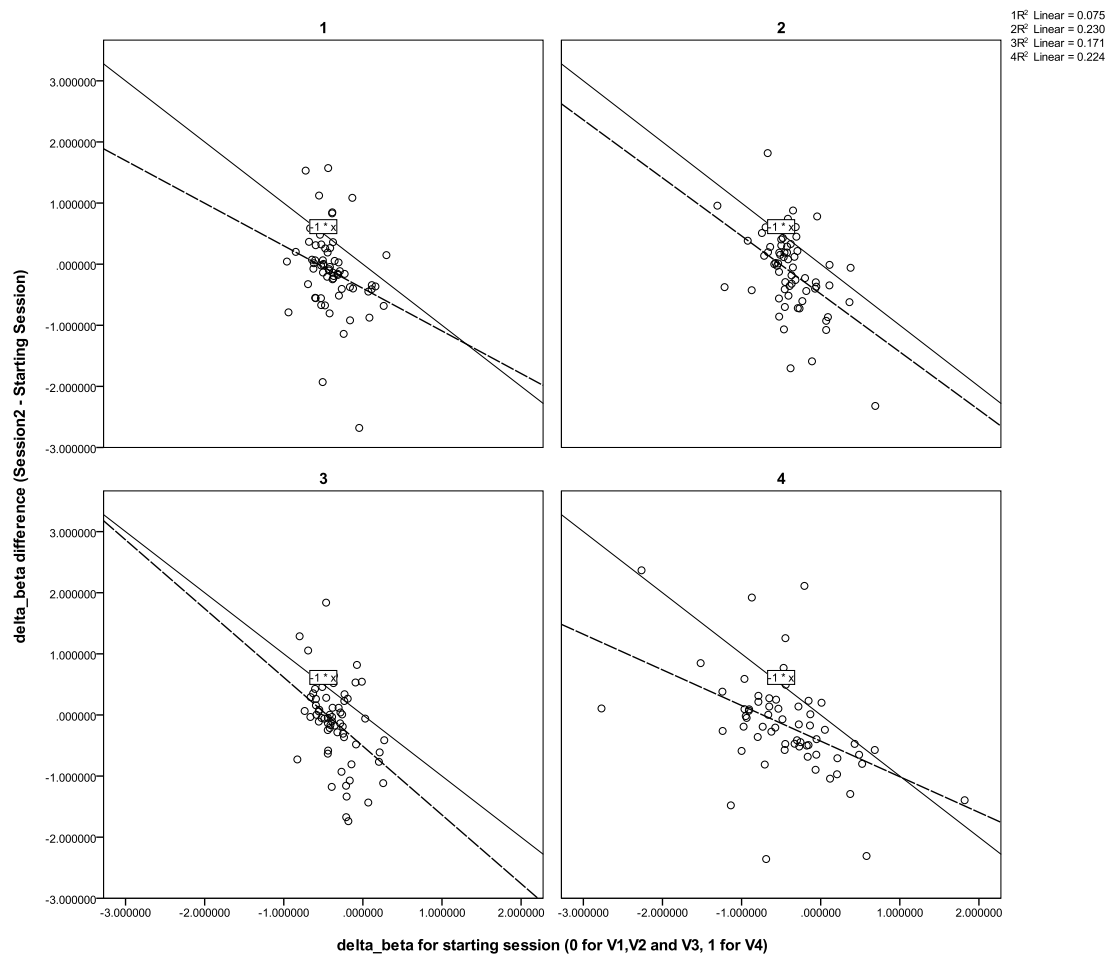


Figure 5.12 Difference between end and starting coefficients, over starting coefficient (\$460 fare)

Table 5.25 “\$460 fare” parameter learning by learning approaches

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	-0.697**	-0.396***	3
Learning Approach 2	-0.949***	-0.487***	1
Learning Approach 3	-1.125***	-0.511***	2
Learning Approach 4	-0.583***	-0.428***	4

### \$520 fare

Figure 5.13 and Table 5.26 demonstrate results for the \$520 fare. Although the systematic error component for Approach Two is a marginally higher than for Approach One, by comparing slopes and scatter plots against the reference line, Approach Two is arguably still the most effective among the four approaches. This is closely followed by Approach One as the second most effective approach.

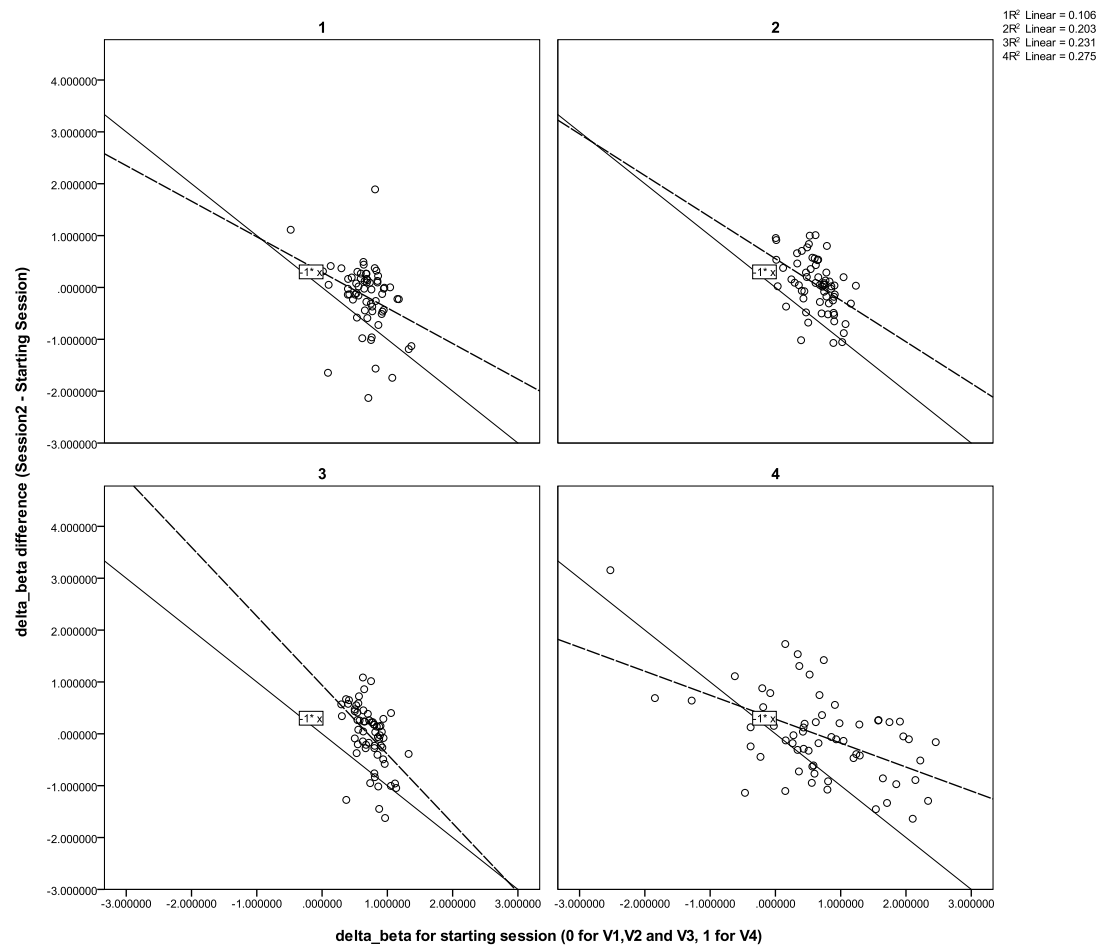


Figure 5.13 Difference between end and starting coefficients, over starting coefficient (\$520 fare)



Approach Three is ranked third. Its slope is similar to Approach One but its systematic error is higher. From observation of both scatter plot and slope, Approach Four ranks the last in terms of the effectiveness of model parameter learning. It is also obvious that individual coefficients vary more than with other approaches, which means there were more individual differences in this approach.

Table 5.26 “\$520 fare” parameter learning by learning approach

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	-0.686***	0.291	2
Learning Approach 2	-0.801***	0.555***	1
Learning Approach 3	-1.331***	0.937***	3
Learning Approach 4	-0.462***	0.283**	4

#### flying time 4 hours

Figure 5.14 and Table 5.27 demonstrate the results for “flying time 4 hours”. Systematic errors for Approaches One, Two and Three are very close, so are their slopes. Approach Two is the best performer followed by Approach One and Approach Three. Approach Four ranks fourth with a slope much further from -1.

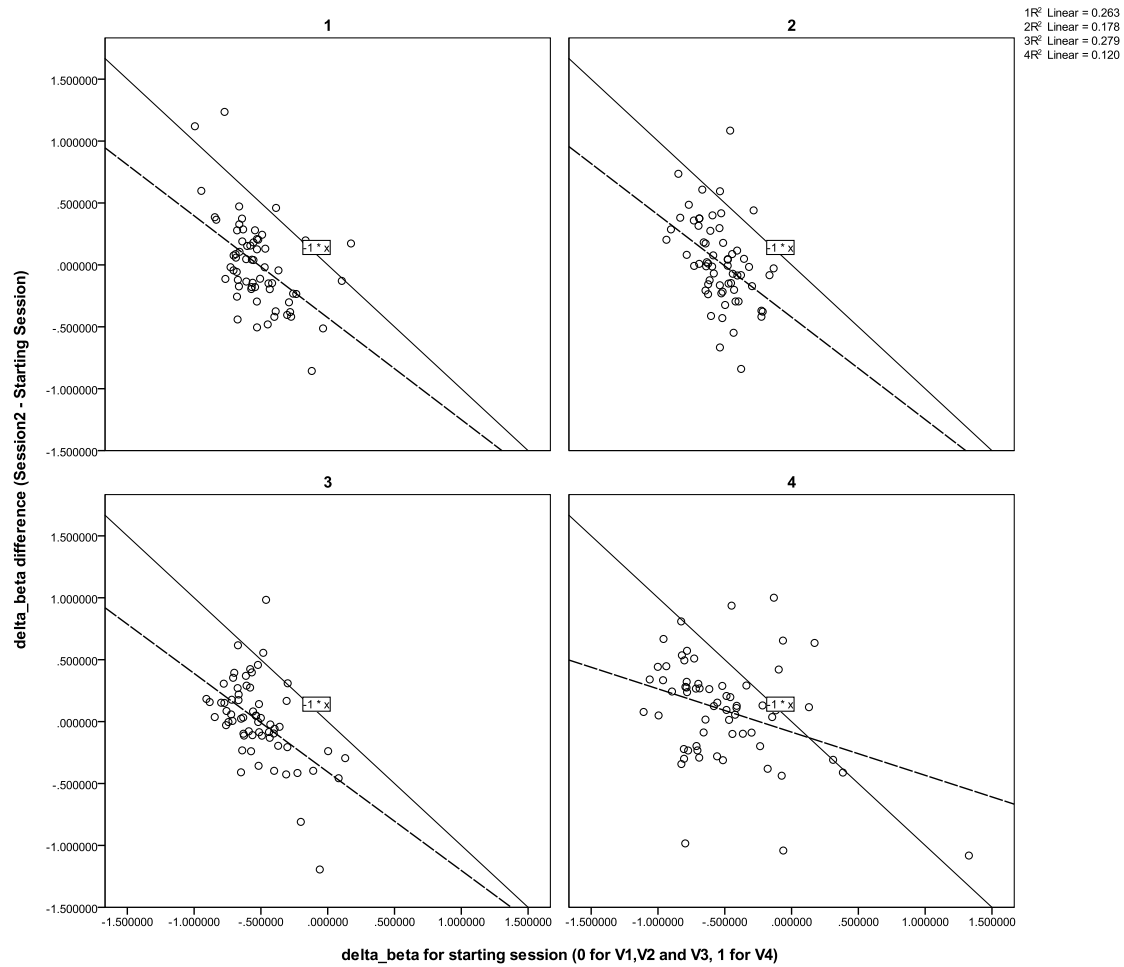


Figure 5.14 Difference between end and starting coefficients, over starting coefficient (4 hours)

Table 5.27 “Flying time 4 hours” parameter learning by learning approach

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	-0.823***	-0.427***	2
Learning Approach 2	-0.827***	-0.422***	1
Learning Approach 3	-0.796***	-0.407***	3
Learning Approach 4	-0.349***	-0.085	4

### Ticket change allowed

Figure 5.15 and Table 5.28 demonstrate the results for the attribute “ticket change allowed”. Systematic errors for all four approaches are similar and lower compared to the systematic errors in other attributes. Among the four approaches, Approach Two once again ranks first in improving learning of model parameters. It is followed by Approach Three and Approach One. Approach

Four ranks the last, however, compared to other attributes, Approach Four performs better for this attribute.

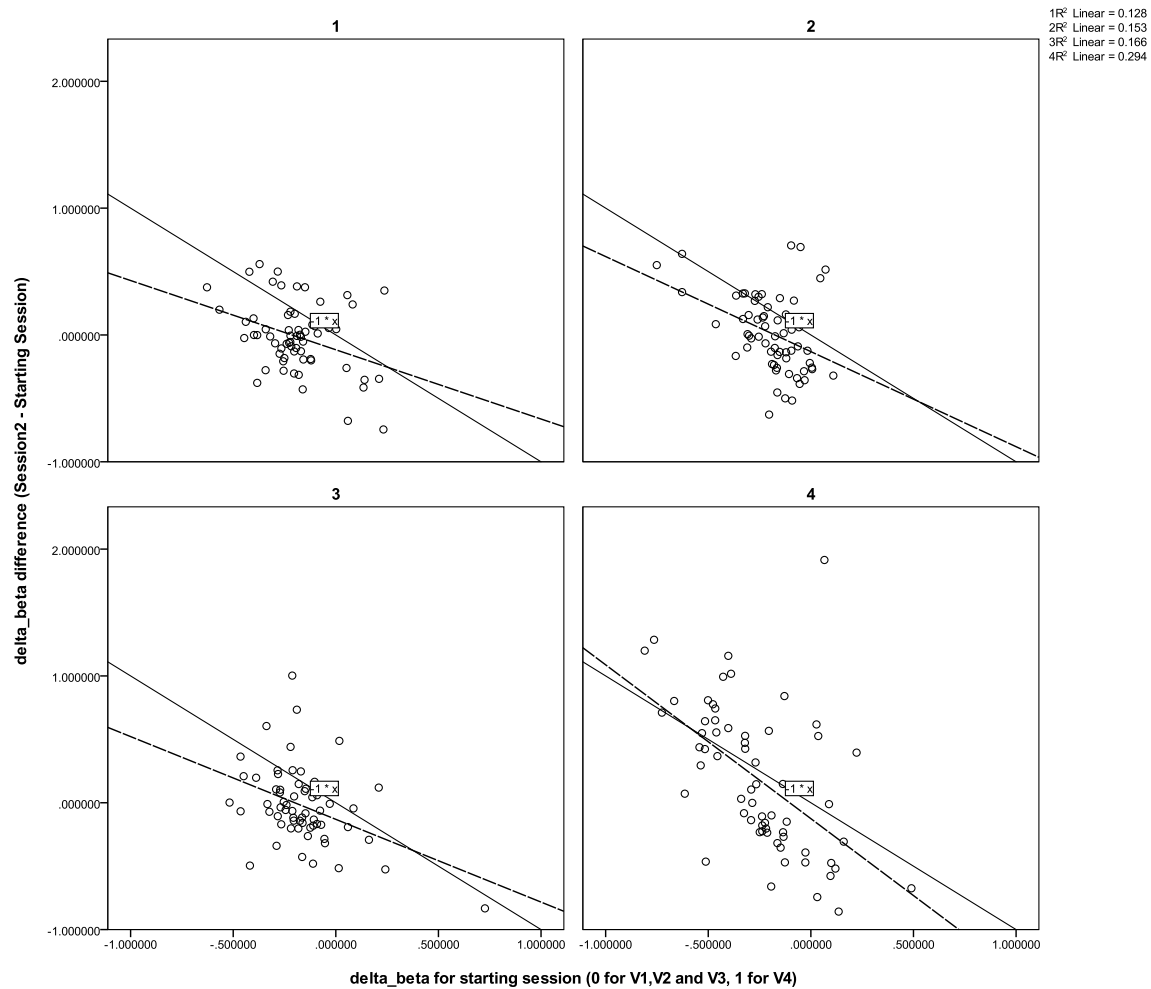


Figure 5.15 Difference between end and starting coefficients, over starting coefficient (change allowed)

Table 5.28 “Ticket change allowed” parameter learning by learning approaches

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	-0.546***	-0.117**	3
Learning Approach 2	-0.750***	-0.131**	1
Learning Approach 3	-0.652***	-0.131**	2
Learning Approach 4	-1.213***	-0.125	4

### Free in-flight food & beverages

Figure 5.16 and Table 5.29 demonstrate the results for the attribute “free in-flight food & beverages”. Systematic errors for all four approaches are not high which means that learners learnt this attribute quite well. Although differences among Approaches One, Two and Three are quite small, by comparing slopes, the rank order of approaches in terms of effectiveness is Approach One, Approach Three and Approach Two. Although learners under Approach Four were learning, the rate of learning is much slower. Again, it ranks fourth as in most other attributes.

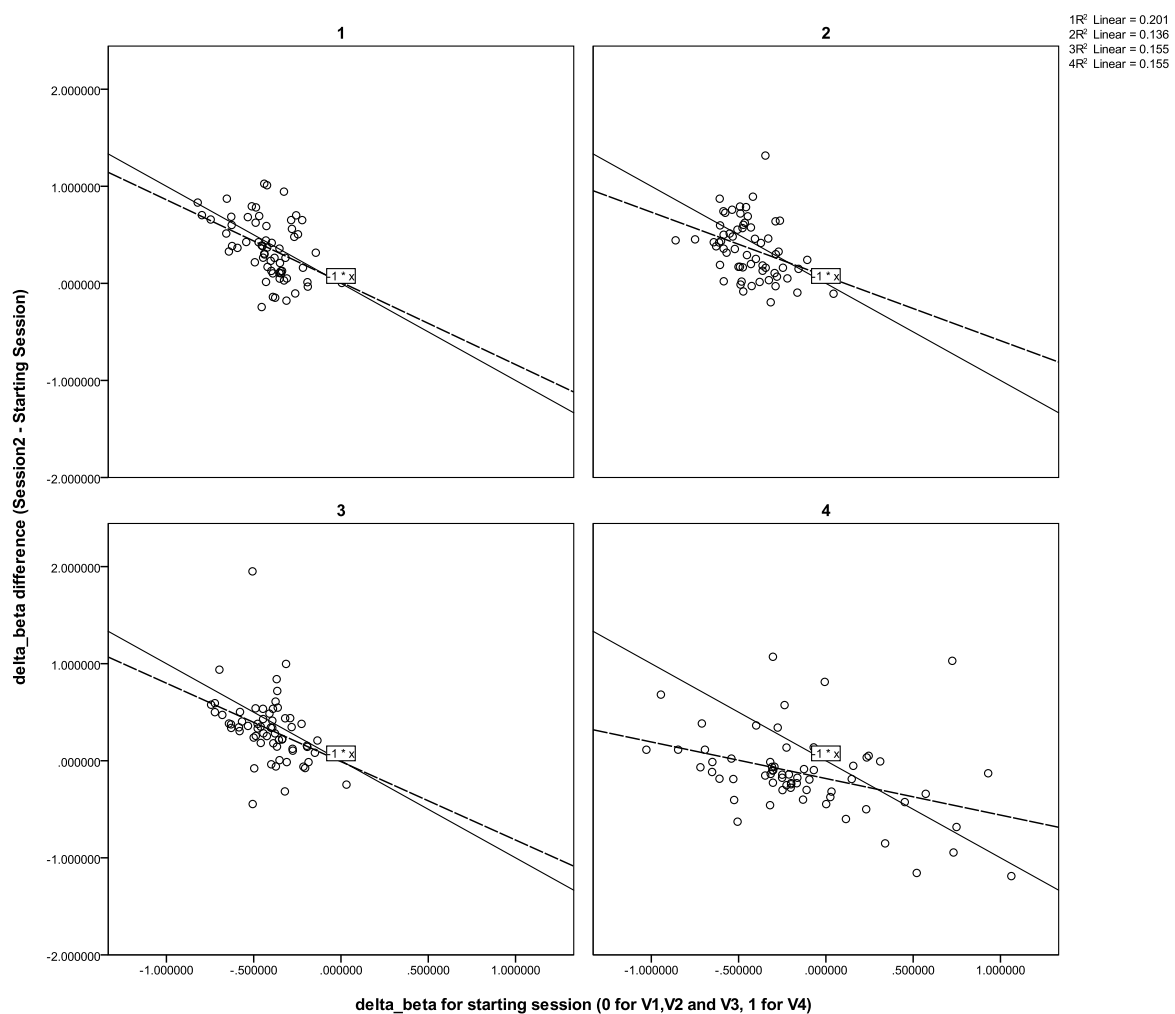


Figure 5.16 Difference between end and starting coefficients, over starting coefficient (free food)

Table 5.29 “free food & beverages” parameter learning by learning approach

	Overall (slope)	Overall (alpha)	Rank
Learning Approach 1	-0.849***	0.012	1
Learning Approach 2	-0.662***	0.072	3
Learning Approach 3	-0.808***	-0.009	2
Learning Approach 4	-0.377***	-0.183***	4

Table 5.30 summarises the rankings of all attributes, averaging the combination of all airlines and fare terms. Approach Two ranks first among all attributes but one, and it is the overall best performer. Approaches Three and One are in close proximity and each approach performs well for some attributes and poorly for others. In light of the overall averages for all nine attributes, Approach Three is slightly better than Approach One. Approach Four ranks last of all four approaches. However, if we carefully re-examine all attributes with regard to the slopes and intercepts shown in Tables 5.21 to 5.29, Approach Four always exhibited the right direction of improvement with relatively low systematic errors among all four approaches. In fact, the only reason that Approach Four has been ranked last is due to its slower learning pace as reflected in the slopes being more distant from -1 than other approaches. This is largely influenced by some learners who were not learning effectively, as shown by more scattered patterns of data points in graphs.

Table 5.30 Summary (mean) rankings of learning approaches

	Airline	Fare	Flying Time	Ticket Change	Food & Beverages	Overall
Learning Approach 1	4	2.7	2	4	1	2.8
Learning Approach 2	1.3	1	1	1	3	1.3
Learning Approach 3	1.7	3	3	3	2	2.3
Learning Approach 4	3	3.3	4	2	4	3.6

By combining all individual coefficients for the nine independent variables to create a representative coefficient for each individual for each session, the regression analysis was conducted in accord with Equation 5.20. This regression analysis provides coefficients for each

combination of learning approach and session. By comparing these coefficients, we can tell which approach caused learners to improve their prediction performance the most, as shown by the lowest coefficient (lower coefficients here represent which approach has less influence on differences between the target model and learners). These results validate the results of previous analysis using individual coefficients for each attribute. Approach Two is again the most effective approach in helping learners to learn the target model parameters. The coefficient for S2 under this approach is 1.514, the lowest for session S2 among all approaches. The Approach Three coefficients for S2 are again quite close to Approach One. However, given that the level of improvement for Approach Three is greater than for Approach One ( $2.035-1.726=0.309$  compared to  $1.950-1.723=0.227$ ), Approach Three is considered a more effective approach. Approach Four is again shown to be not effective compared to the other three approaches.

Table 5.31 Average learning of model parameters by approach & session

Source	SS	df	MS	Number of obs	=	693
				F( 11, 682)	=	668.77
<b>Model</b>	2277.661	11	207.06	Prob > F	=	0
<b>Residual</b>	211.158	682	0.310	R-squared	=	0.915
				Adj R-squared	=	0.914
<b>Total</b>	2488.818	693	3.591	Root MSE	=	0.556

Norm of coefficients	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]
session 0 (approach 1)	1.950	0.070	27.820	0.000	1.812 2.088
session 1 (approach 1)	1.623	0.070	23.150	0.000	1.486 1.761
<b>session 2 (approach 1)</b>	<b>1.723</b>	<b>0.070</b>	<b>24.580</b>	<b>0.000</b>	<b>1.586 1.861</b>
session 0 (approach 2)	1.979	0.070	28.230	0.000	1.841 2.117
session 1 (approach 2)	1.374	0.070	19.600	0.000	1.236 1.511
<b>session 2 (approach 2)</b>	<b>1.514</b>	<b>0.070</b>	<b>21.600</b>	<b>0.000</b>	<b>1.377 1.652</b>
session 0 (approach 3)	2.035	0.070	29.030	0.000	1.898 2.173
session 1 (approach 3)	1.673	0.070	23.860	0.000	1.535 1.810
<b>session 2 (approach 3)</b>	<b>1.726</b>	<b>0.070</b>	<b>24.630</b>	<b>0.000</b>	<b>1.589 1.864</b>
session 1 (approach 4)	2.088	0.070	29.780	0.000	1.950 2.226
<b>session 2 (approach 4)</b>	<b>2.090</b>	<b>0.070</b>	<b>29.820</b>	<b>0.000</b>	<b>1.953 2.228</b>

The analysis results yield an identical order for the four learning approaches in terms of helping model parameter learning. The consequent ranked approaches are in the order of Approach Two, Three, One and Four.

For Approach Four, a separate analysis was undertaken to ascertain the impact of target segments on model parameter learning. Table 5.32 shows that the results are consistent with the findings when testing prediction accuracy. That is, target segments have a significant impact on both prediction accuracy as well as target model parameter learning. Learners performed best in learning model parameters if the target segment to predict was consumer group B. The order of consumer groups A and C in the model parameter learning is different to the order found in testing prediction accuracy, namely that learners predicted consumer group C performed best.

Table 5.32 Average learning of model parameters by session & target segment (Approach Four)

Source	SS	df	MS	Number of obs	=	126
				F( 4, 122)	=	297.99
<b>Model</b>	575.192	4	143.798	Prob > F	=	0
<b>Residual</b>	58.872	122	0.483	R-squared	=	0.907
				Adj R-squared	=	0.904
<b>Total</b>	634.064	126	5.032	Root MSE	=	0.695
<b>Norm of coefficients</b>	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
session 1 (approach 4)	2.134	0.124	17.240	0.000	1.889	2.379
session 2 (approach 4)	2.136	0.124	17.260	0.000	1.891	2.381
Target segment A	0.478	0.152	3.150	0.002	0.178	0.778
Target segment B	-0.615	0.152	-4.060	0.000	-0.915	-0.315
Target segment C	0.000					

### 5.3.7 Summary of Hypotheses H1b, H2b and H3b

This section summarises the results and provides a basis to respond to research hypotheses H1b, H2b and H3b. Hypothesis H1b assumes that Approach Two is more effective than Approach One in helping learners to understand the target model. Hypothesis H2b assumes that Approach Three is more effective than Approach One for the same measure. Hypothesis H3b treats Approach Four separately and assumes it helps learners gain an understanding of a target class' model. It was not compared to other learning approaches due to incomparable tasks.

H1b states that:

Learners who receive outcome feedback (Approach Two) after each training task have a better understanding of the target model than those who perform self-regulated learning using a “plain vanilla” MDSS (Approach One).

Based on the results discussed in Section 5.3.6, the evidence is clear that Approach Two is the best learning approach to help learners improve quickly in understanding target model parameters. This result is consistent in the learning involving all attributes in the target model. On the other hand, Approach One, regardless of learners’ positive feedback and ease of use, performs poorly on some key attributes. It is not only the case that there is a large systematic error component in many attributes, but learners are leaning towards increasing and decreasing the gap between their own ideas and target model. Why outcome feedback has performed so well is an interesting finding in itself, because it does not match the findings of other researchers in the MCPL field (e.g. Cooksey 1996). This finding will be discussed further in Chapter 6.

H2b states that:

Learners who receive a diagnosis of their own model after training tasks (Approach Three) gain a better understanding of the target model than those who perform self-regulated learning using a “plain vanilla” MDSS (Approach One).

As discussed and shown in the results, Approach Three is more effective than Approach One but not to the same degree as Approach Two. For some attributes such as brand, Approach Two performed much better than Approach One in helping learners gain an improved understanding of target parameters quickly, but on some attributes Approach Two ranked behind Approach One. A further point of discussion is why providing more information has not achieved outcomes as great as expected. Approach Two does provide the most insights and in-depth analysis not only related to the task itself, but also to learners. Therefore, a more interesting discussion should also



cover a comparison of Approaches Three Two as well. This will also be further discussed in Chapter 6.

H3b states that:

Learners who receive class information and classification feedback after training tasks (Approach Four) gain a better understanding of a target class' model with more tasks and feedback given in training.

As earlier discussed, although in a comparison to Approaches One, Two and Three, the Approach Four seems to perform poorly, there is no doubt that learners were learning the model parameters. Learners have shown they were improving in the right direction to reduce the differences between their own judgements and the target consumer group's model. Moreover, due to other factors such as different levels of difficulties in learning target groups and whether learners had made incorrect assumptions about a consumer group if they happened to be members of that same group, the results showed a pattern of more individual differences. This pattern is evident in studying the graphs related to each attribute. This implies that some respondents who performed poorly may have influenced the results and overall performance of the approach. Whether this phenomenon is due to the constraints of the learning approach itself, or due to differences in consumer groups, will be further discussed in Chapter 6.

## **5.4 Summary**

This chapter covered methodologies and results testing prediction accuracy and target parameter learning. Among all four approaches, for both prediction accuracy and target model parameter learning, Approach Two performed the best among all approaches. Approach Three was found to be more effective than Approach One on both measures, but its advantages are much smaller than Approach Two compared to Approach One.

The exploratory nature of Approach Four and the assignment of different tasks, determined that it should be tested separately to other approaches. The results indicated that there are great individual differences in terms of performance for both measures. The best performance of learners was in improving the number of times they predicted the correct consumer group. On increasing prediction accuracy of full probabilities, less than half of the learners improved in S2 over S1. Others did not improve. With regard to model parameter learning, learners were improving in the right direction with less systematic errors than for other approaches, but the pace of improvement was much slower than for other approaches. It is possible that learning class models under this approach required much more time.

In Chapter 6, these findings will be further discussed by extending the implications theoretically and practically, and proposals will be made for future research directions.

# **Chapter 6 Conclusions and Implications**

## **6.1 Introduction**

Chapter 5 discussed analysis methods and presented the analysis results for the empirical study without extending discussions to the broader context of this research. Chapter 6 aims to fulfil this objective by further considering the contribution made by the present research to theory and practice, and suggesting some potential implications.

Section 6.2 links the research findings to the key research problem, improving subjective predictions of probabilities through learning from an external model to bridge the gap between two sources in making the predictions, namely, intuition and the model. In this section, each learning approach tested in the empirical study is discussed in relation to the key characteristics of learning reviewed in Chapter 2. Section 6.3 discusses the contributions that this research makes to marketing and decision-making theories and its practical implications. Section 6.4 concludes the thesis by considering the limitations of this research and suggesting possible direction for future research.

## **6.2 Conclusions of Research Problem and the Four Learning Approaches**

Key conclusions can be made regarding the research problem and the hypotheses of four learning approaches which are strongly supported by the analysis results. We will discuss them sequentially.

### **6.2.1 Conclusion regarding the Research Problem**

The most important finding of this research is that people can gain substantive understanding of key attributes of a target problem through learning from a formal model. As a result, they can improve the accuracy of subjective predictions. This is a positive response for the research direction proposed by this research. That is, the gap between intuitive predictions and model predictions can be narrowed and the quality of predictions and decision making can be improved.

The research problem as stated in Chapter 1 is about finding effective ways to help people improve their subjective predictions of probabilities through learning from formal analytical models. As explained in Chapter 5, in model learning, learners are converging to parameters of the model. During the process, subjective predictions are made by a person's updated mental models. We know from overwhelming evidence in judgement and decision making literature that models often outperform even the predictions of the acknowledged experts (e.g. Grove et al. 2000). However, it is also clear that decisions and predictions in practice, including marketing, are mostly made by people using their current state of knowledge (e.g. Burke & Miller 1999). Therefore, the solution that this research advocates is to improve people's own knowledge so updated intuition becomes more reliable in making predictions and furthermore in making decisions. Without this premise as the key consideration, findings of the research problem cannot be valued.

Thinking more about the nature of probability prediction, as mentioned in Chapter 1, probability predictions are levels of certainty with regards to propositions (e.g. events, choices and preferences). This is widely acknowledged by researchers (e.g. Winkler 1996). To establish levels of certainty, people either need to accumulate evidence, especially the frequency of events occurring in the past, to form their beliefs inductively, or to gain knowledge from certain sources to make deductive inferences based on new incidences. If we consider marketers gaining experience from practice as a way to establish their beliefs of marketing events through inductive learning, then becoming trained using formal models is to gain knowledge from external sources which allows the making of deductive inferences. This contrast of inductive and deductive inference is so relevant that it can be used to describe all approaches to the accretion of knowledge for predictions. Considering any learning approach tested in this research, it can be said that they are all combinations of the two types of inferences. For example, Approach One is mainly an inductive learning approach, and Approach Three is the opposite, mostly relying on deductive

inferences. These approaches are fundamental epistemological methods for gaining knowledge and establishing beliefs common to many generalised problems. The fact of subjective predictions is that the inferences people make, need to be supported by a knowledge base. This research proposes that knowledge base can be established from learning a reliable model. The further question is “how” to establish knowledge base effectively. Chapter 2 reviewed some key learning characteristics from learning theories in psychology and cognitive science. Four learning approaches were nominated for testing.

### **6.2.2 Conclusion on Learning Approach One**

Learning Approach One allows learners to use a “plain vanilla” style MDSS to conduct “what-if” analysis by themselves. The learning process is unsupervised and unguided. “Plain vanilla” style MDSS is the default form of decision aid in marketing practice. It covers many business intelligent tools such as dashboards and data visualisation tools. A common attribute of these tools is that they offer self-selected features to build example scenarios but offer little or no explanation as to outcomes. These tools are primarily built for the purpose of prediction or reporting, but not learning. If this tool is used as a learning tool, it requires people to perform self-regulated learning and select their own scenarios but offers no further assistance about how inferences were made. From testing Hypotheses 1a to 2b, this approach was not as effective as other approaches offering feedback to experimental design controlled training tasks (Approaches Two and Three) in both prediction and model learning.

The key characteristic of learning relating to this Approach One was covered in Section 2.3.2 regarding self-regulated learning versus controlled experiment. With the flexibility of selecting own tasks and the attractiveness of user friendly features, the preference of learners in favour of this tool and approach are high according to open-ended feedback. In the context of a preferred interface and features, there is no evidence available showing that learners may effectively converge

to a desired level of understanding of the target model and make more accurate predictions, compared to other approaches.

In light of unsatisfactory results that may emerge from this learning approach, the first possible interpretation is that neither learners' preference nor the system's technical features have direct relationships to prediction outcomes or model parameter learning. One argument this researcher put forward in Chapter 2 is that there should be more fundamental aspects in learning than interface design or visualisation features of a training tool. The characteristics of a learning approach should be task- and problem-specific, such as when we consider the hierarchy of feedback. Researchers studying self-regulated learning pay great attention to issues such as learners' motivation, initiative, perseverance and adaptive skills and they believe that a preponderance of these characteristics among learners generates good learning results (e.g. Zimmerman 2008). This was not the case for learners assigned to this approach.

The second possible interpretation is that learners cannot converge to the target model parameters to make more accurate predictions, if training scenarios are self-selected. When experimental design was discussed in Chapter 4, the most important reason offered for using experimental design was to identify a small and limited number of tasks and sets that can efficiently represent the problem space, because the number of combinations of attributes and levels is large and unmanageable. This is the case for the airline choice in this experiment. The number of combinations is a large number in the thousands (the total combination was 32,768 for the airline choice problem). In these circumstances it is difficult to believe that learners are able to select a set of well-represented combinations through self-selection.

In marketing practice, certain analyses are often conducted using SP or RP data with the models and analysis results built into a model-based MDSS, thus expecting marketers to improve their understanding of the target problem in order to make better predictions and eventually better decisions. The present research finds that this tool and learning Approach One are not effective in assisting marketers to better understand the target model and thus improve their predictions of probabilities.

### **6.2.3 Conclusion on Learning Approach Two**

As pointed out in Section 2.3.3, feedback is the focal point for studies in learning. The single key characteristic relating to Learning Approach Two is outcome feedback. It is the approach for providing immediate correct answers to learners after they have made predictions. As clearly demonstrated by the results, this approach performed the best in both improving prediction accuracy and model parameter learning. This finding conflicts with commonly held views by many MCPL researchers (e.g. Cooksey 1996; Hammond Summers & Deane 1973; Kluger & DeNisi 1996). According to these researchers, outcome feedback is supposed to perform worse than cognitive feedback, including in-depth statistical information regarding tasks and a summary of learners' performance (Learning Approach Three). In this experiment with a real choice problem related to airline offers, Approach Two improved learners' predictions faster than any other approach in the first session and maintained that position until the end of the second session. Why did this approach perform so well in this experiment? What could have made the experiment outcome different? These are important questions to further explore the research hypotheses, comparing this approach with the default approach (Approach One). From the results, there were no real differences between learners associated with different approaches when they started the training. Approach Two also performed well on all attributes. Therefore, individual differences and particular attributes are unlikely to have been the explanation because all other approaches asking learners the same questions included exactly the same training tasks (Approaches One and

Three). It is most likely that the use of interactive feedback in combination with experimental design contributed greatest to the performance of Approach Two.

According to MCPL researchers, the reason why outcome feedback may perform poorly is because this type of feedback provides neither direction nor detailed information on what to improve. In other words, outcome feedback lacks information to help learners identify and improve the particular areas that caused the errors. This belief is shared by other MCPL researchers and a general idea is that there is a negative relationship between outcome feedback and prediction performance (e.g. Steinmann 1976). It is implied that learners cannot learn from a trial and error style of inductive learning approach. Instead, they should be provided with in-depth feedback advising them the reasons behind the errors so that they may adjust their judgements. Based on the evidence of this research, it seems the abilities of the participants to achieve success through trial and error learning, were underestimated.

Empiricists always believe that inductive learning is a natural psychological process. As Godfrey-Smith (2003) puts it, there is more than one type of inductive learning. For example, projection is one type of inductive learning in which the purpose is not to reach the ideal conclusion but to use existing or past evidence to make a better prediction of the next case. Gradually, understanding can be accumulated to reach an ideal point. Perhaps this was the approach that learners were taking. It was certainly evident in learners' open-ended feedback that more than one learner considered this exercise as a good "game" in which they had to gauge where to improve. Also from feedback received, it is known that participants learned more when they achieved a close prediction or a very poor prediction. This is a clear indication of projecting future similar cases from past cases, either good or bad ones. This learning can be viewed as an inductive learning approach from "example comparison" (i.e. new case compared to past cases), a one key approach



proposed by Komatsu (1992) in discussing concept learning in Section 2.3.5. Another important factor is that in this approach, well-selected sets of training tasks based on experimental design were used so the problem space is evenly covered in a small but efficient number of sets. This was discussed in the previous section. Experimental design such as this, applied to a real-life problem where people make choices, was considered unimportant in MCPL studies by psychologists. Often hundreds if not thousands of training sets were used without an efficient or orthogonal design.

To summarise this approach, it is worthwhile mentioning the debate between Hogarth and Edwards on human capacities in improving probability assessment, as discussed in Section 2.3.2 (Edwards 1975; Hogarth 1975). It now seems reasonable, given the evidence of this research, that there is merit in Edwards' argument that humans do not have any inherent disadvantages when they have shown poor performance or lack of learning in probability assessment experiments. Bad experiment outcome can be caused by poorly designed experiment. At least in this study, when a proper experimental design was used to control training tasks, people showed they could effectively learn model parameters and improve their prediction accuracy.

#### **6.2.4 Conclusions on Learning Approach Three**

The type of learning that Learning Approach Three aims to trigger involves several capacities of learners. First, learners need to be able to synthesise and integrate key pieces of information from cognitive feedback. It is a continuous generalisation process. Alternatively, it can also be thought as a “feature abstraction” process as addressed by Komatsu (1992). Second, learners need to be able to apply deductive inferences to new cases on the basis of their gains from learning. From the results and learners' open-ended feedback, it seems clear that difficulties that learners had in integrating different key information to establish a new mental model is the key problem. As some learners put it, this process is too difficult with too many pieces of information to consider.

Recalling the main experimental conditions that were applied in Approach Three, as summarised in Table 4.2, this approach provided learners with detailed statistical information about each attribute and the choice probabilities of each level, as well as each attribute's weight in determining overall probabilities in feed-forward information before learners attempted the training tasks. Moreover, this information was provided by comparing each learner's own model with the target model. After each session, newly updated information was provided to each learner based on the analysis of their previous session. This process matched the concept of appropriate cognitive feedback as suggested by MCPL researchers, thus no problem arose as a consequence of providing inadequate information (e.g. Cooksey 1996; Hammond, Summers & Deane 1973; Kluger & DeNisi 1996).

Cognitive feedback and feed-forward information are supposed to perform best in helping learners to deal with complex learning problems. In contrast, simple outcome feedback is supposed to impede learning in complex tasks. This is not borne out by the findings of this research. Approach Three did not perform as well as Approach Two, although it performed better than Approach One. From this experiment with Approach Three, the evidence seems to support the following proposition: that trying to completely update mental models that learners previously applied through a single round of in-depth cognitive feedback with a great amount of information, is not the best approach to use. Such an approach may rely upon too many skill sets and too great an effort from learners, but it may be an even harder for learners to apply newly gained and undigested knowledge in new tasks, in order to gain immediate improvement. It is quite possible that if this approach is used in real-life training, more training sessions would be required, as applying deductive inferences from the newly updated mental model is essentially a harder task for most learners, than trial and error learning.

Another important point to discuss relating to Approach Three, is knowledge representation (KR). In Section 2.3.4 when KR was discussed, the focus was on the contrast between propositional and imagery representations. In Approach Three, the attention was clearly focussed on propositional KR. As Anderson (1978) explained, propositional KR has nothing to do with the actual form of information such as verbal or graphic formats, but instead, refers to the representation of concepts with true values and clear evaluation rules in unambiguous statements, descriptions or other formats. In Approach Three, clearly the “represented world” is a choice model with coefficients. The difficulty in representing this knowledge resides in the differences of coefficients and probabilities. Differences between coefficients and differences of probabilities are not the same. Building a computer-based model driven by MDSS is not difficult because it can be built as precise knowledge following exact rules, as in the MNL model. Coefficients are first converted to utilities and probabilities are then calculated from proportions of utilities. The same method cannot be expected to be performed by human brains because of the large volume of calculations. Even if learners can calculate probabilities with external help from tools, they most likely would fall into the same dilemma described in the “Chinese room” scenario as discussed in Section 2.2.3, which means that no real knowledge was gained even though the mechanism worked efficiently to provide answers (Searle 1980). Therefore, the challenging problem for KR when teaching a precise and complex model is to identify the most appropriate reasoning and semantics for KR. Representing MNL or similar models in KR by showing probabilities of attribute levels being chosen in each attribute may not be the only or the best approach for KR. This area should be further explored to ascertain what form of KR is the best at helping people understand the model and be able to synthesise knowledge to make better probability predictions.

#### **6.2.5 Conclusions on Learning Approach Four**

The results have shown that learners under this approach made slow improvement in both model learning and prediction accuracy. The most obvious improvement appears to be that learners’

predictions tend to reflect the consumer group they were assigned to predict. Their improvement in parameter learning was slow but headed in the right direction. One finding that stands out is that learners varied greatly in their prediction accuracy and parameter learning. One factor contributing to this variation was the target consumer groups that learners were assigned to predict. The results show that if the behaviour of a target group is more salient to explain, they tend to perform better.

In designing this learning approach, one key characteristic was category learning through prototypes (e.g. Markman & Ross 2003). As discussed in Section 2.3.5, a prototype does not refer to particular features or examples but the central tendency of a category (Rosch 1973). Such central tendencies are shared among all members in this category, and a category matching the prototypes is much easier to learn than a category that does not match. This can be used to explain the exact finding from this research. That is, a consumer cluster better matched with one prototype is easier to learn, while a consumer group that shares characteristics of more than one prototype is more difficult to learn. In this experiment, three consumer clusters were gained from archetypal analysis which provided extreme examples, such as people who strongly prefer the cheapest price or those who strongly prefer major airlines. However, not every cluster was defined by strong characteristics. In one of the three clusters (Group C), the definition was a mixture of several not so strong characteristics which were also similar to other clusters. The only difference in these characteristics among clusters is the degree of impact on preference behaviour. In the former case when a cluster has a single strong characteristic, learners were obviously performing better. When a cluster has a mixture of several features, learners did not perform as well.

Another point to discuss relates to feedback. In Approach Four the only feedback is similar to an outcome feedback informing learners whether they were predicting the right consumer group

based on results of a Bayesian classifier. While this feedback was useful in helping them approach the right consumer group, perhaps by adding outcome feedback showing probabilities of the target group, or all three groups, could help learners improve their probability predictions as well. Although the intention was to test whether learners can learn from similarities and dissimilarities of categories, from feed-forward information provided before and during the learning tasks, it is possible to include extra feedback on the true probabilities of the three groups, because they are also an important part of similarities and dissimilarities. This will be mentioned again in Section 6.5 as a topic for further research.

On knowledge representation (KR), Approach Four uses both graphic and verbal information to establish a mental image of consumer groups. This idea was formed by combining ideas of researchers in KR supporting imagery KR such as Kosslyn and the Conceptual Spaces theory of Gärdenfors (e.g. Gärdenfors 2000; Kosslyn 1981). In reality, this researcher is not convinced that all learners understand the purpose of this learning approach or whether they have constructed appropriate mental images to help them in their probability predictions. From the feedback of learners, some did not find this approach helpful. For other learners, it was the opposite effect as they found this approach quite useful. The proportion of learners represented by the two groups is roughly fifty-fifty, half in each group, based on results outlined in Chapter 5. From the experience of this research, in testing an approach built on rather abstract theory in cognitive science lacking suitable measurement and experimental methods, existing experimental methods are inadequate. Either the observation method or a different design needs to be developed. The most difficult aspect is how to determine whether a feature design based on an abstract concept can be realistically handled by an experimental condition and then determine what measurement process can be used to test it.

The positive message arising from studying Approach Four is that even though this approach is only an exploratory study and did not show satisfactory results, it highlights problems that are more obvious in testing abstract theories from cognitive science. Unlike testing theories from psychology in which observed behaviour has long served as primary evidence, learning theories from cognitive science such as theories relating to mental images, category learning and other mental process require new sources of observation data from experiments. The general trend of marketing researchers showing more interest in neuroscience and eye-tracking experiments is certainly a good start.

### **6.3 Contributions and Implications**

This research has made several unique theoretical contributions to marketing and decision making and decision support research.

First, in studying marketing decision support systems, the current focus is on either a model itself, or building computer systems to effectively support a model and its adoption in an organisation. This is a tradition emanating from areas such as operations research and marketing engineering in which research on MDSS was grounded. Learning theories from either traditional disciplines such as psychology, or new disciplines such as cognitive science, have had no impact in this area of research so far. This research integrates learning theories from these disciplines, decision support systems, a consumer choice model and intelligent tutoring mechanisms in order to bridge the gap between intuitive and the mechanistic judgements of models. There is a new direction in which decision support can be conducted, i.e. training decision makers through a tutoring system, supported by appropriate learning approaches improves decision makers' own knowledge.

The second contribution is that this research integrates both an experimental design and real-life consumer choice model with probability learning research. Probability learning research is

traditionally the research area of researchers in Social Judgement Theory under the MCPL framework, or psychologists working in the field of subjective probability assessment and forecast. Researchers working in choice modelling and experimental design have different research interests working in different fields. This research brings the two areas into a learning experiment in which learning approaches can be experimentally controlled and the learning subject is, for the first time, a real-life choice model. The results are satisfactory and more research such as this can certainly be conducted.

The third contribution is in proposing an extended framework for learning. In learning theories developed in psychology by behavioural psychologists, the standard framework is the traditional S-R-O association framework treating learning as a repeated loop from stimulus to response then to outcome. In this framework, the only observable component is response. Outcome is the evaluation of response. This research extended this standard framework to S-P-R-O with learners' mental processes included. This emphasises the importance of the association between stimulus and process, and process with response. Even though it still lacks observation and measurement methods in experiment in capturing process, this researcher believes that by applying a particular stimulus, certain learning process can be facilitated and triggered. The learning process plays an important role between stimulus and response and process determines the response and outcome. The design stimulus, including its contents KR and the learning approach, can be considered a complete "process design" which is to determine how to trigger the desired learning process. With further development in theories and techniques to capture process, it is likely that a focus on learning can be put on S-P and P-R associations.

The final theoretical contribution is in measuring the effectiveness of learning. In past research of probability, the only key measure is prediction accuracy. In this research, a new measure of

"substantive learning" was introduced. The accuracy measure is a normative standard corresponding to how closely the predicted and actual probabilities match. In past research of learning, a variety of analysis methods were used to measure accuracy. In this research, strictly proper scoring rules developed in the subjective probability assessment literature were emphasised and held as the most appropriate approach for analysing prediction accuracy between different probability distributions. "Substantive learning" specifically refers to gaining knowledge of key characteristics of the target problem, which in this research means understanding the parameters of a target training model (Hogarth 1975; Winkler 1996; Winkler & Murphy 1968). In this research, inferences were made based on an individual based regression model informing the impact of attributes to the differences of predicted and actual probabilities. In different contexts or different problems, inferences can be made using other types of models. However, learning is not directly observable in data; a two-step process is required with the first step providing a primary model and the second step providing inferences about learning.

One possible practical implication is to combine current model-based MDSSs with Intelligent Tutoring Systems (ITSs) to better support marketers. Model-based MDSSs provide answers to marketers' "what-if" queries. This tool ensures that predictions and decisions made by decision makers can be verified. ITS can be used to help marketers understand models and become better judges overall in making prediction and decisions. As proposed by Blattberg and Hoch (1990), intuition and a model can work jointly to produce better results in marketing decision making. If we consider MDSS as fundamentally an interactive and visualised model, then ITS is a tool to improve intuition. This is a new and unfamiliar field for those who work in the area of choice models and decision support tools. MDSS is more familiar to people who work in the computer science, training and education fields. It has been used widely from school education to training pilots. Learning theories and different learning approaches are fundamental to these systems



(Woolf 2009). Having an extra tool like ITS to train end-users (marketers and decision makers) to understand models may provide more benefits than improving the features of existing tools such as data visualisation or interface design perspectives.

## **6.4 Limitations and Future Research**

As mentioned in Chapter 1, this research is a proof-of-concept study integrating several disciplines to identify a new direction to improve probability predictions and people's intuition overall. The limitations of this study can be summarised in two points. First, due to the lack of previous research, some learning approaches especially Learning Approach Four are of an exploratory nature. At a minimum, some areas may be adjusted and improved to further test the same learning approach. For example, probability feedback can be incorporated and more sessions included. It is possible due to the differences in learning approaches, that one approach may naturally require more time and practice than other approaches. Although a process was proposed in the extended S-P-R-O framework and learning approaches and information were designed to trigger a certain desired mental process, there is no way to directly observe or capture this mental process to determine what learners relied upon during that process. This is not a limitation of this research alone but a limitation applicable to all studies of learning relying on observable response data.

There are several further research directions which are possible based on the findings of this research. First, the findings from this study can be re-tested using different product categories or choice models especially models with more parameters. One of the foci could be on whether Learning Approach Two also performs well when models become more complex. Another focus can be Learning Approach Three to ascertain whether the presentation of the same statistical information may be altered and become more effective. Research of the approaches could be further split into more conditions; for example, using an experimental design to control certain feedback elements, or a component of the learning approach. This refinement of the experiment

may further indicate which element is more crucial in driving effective learning. With the research findings from this study, it should be possible to test a learning approach in detail to understand the effect of each relative attribute. For researchers in the areas of judgement and decision-making, this experiment can be conducted among different groups; for example, among marketers who are working in a related industry to ascertain whether the combination of MDSS and ITS works more effectively than using MDSS alone. The results of this stream of research may have important managerial implications. The research can also be focused on developing new methodologies for probability learning studies. For example, we could aim to develop new ways to better capture and analyse underlying processes that learners adopt when making predictions. Some research in a laboratory environment may be required to closely examine learners' activities during learning.

These possible future research directions are certainly not an exclusive list. Researchers and practitioners may find other useful connections to topics and areas in which they are working. The positive message that flows from this is: theories and practices in marketing, psychology, cognitive science and computer science are in vastly better shape today than at the time when many of the studies cited in this thesis were conducted. Researchers and practitioners can certainly better equip themselves to further explore this traditional yet innovative field of subjective probability learning and prediction.

## Appendix 1 Experimental Design for Stage 1 Survey

Block	Subset	Alternative	Fare	Flying time	Ticket Change	Food/Beverages
1	1	Qantas	\$580	6 hours	Allowed	Not Free
1	1	Virgin	\$520	4 hours	Not allowed	Free
1	1	Jetstar	\$400	6 hours	Not allowed	Not Free
1	2	Qantas	\$460	4 hours	Not allowed	Not Free
1	2	Virgin	\$580	6 hours	Allowed	Free
1	2	Jetstar	\$400	6 hours	Not allowed	Not Free
1	3	Qantas	\$520	6 hours	Allowed	Not Free
1	3	Virgin	\$580	6 hours	Allowed	Not Free
1	3	Jetstar	\$460	4 hours	Allowed	Free
1	4	Qantas	\$460	6 hours	Allowed	Free
1	4	Virgin	\$520	6 hours	Not allowed	Free
1	4	Jetstar	\$580	4 hours	Not allowed	Free
1	5	Qantas	\$520	6 hours	Not allowed	Not Free
1	5	Virgin	\$400	6 hours	Allowed	Free
1	5	Jetstar	\$520	4 hours	Not allowed	Not Free
1	6	Qantas	\$400	4 hours	Not allowed	Not Free
1	6	Virgin	\$520	4 hours	Not allowed	Not Free
1	6	Jetstar	\$460	4 hours	Allowed	Free
1	7	Qantas	\$460	6 hours	Not allowed	Free
1	7	Virgin	\$460	6 hours	Not allowed	Not Free
1	7	Jetstar	\$400	4 hours	Allowed	Not Free
1	8	Qantas	\$400	6 hours	Not allowed	Free
1	8	Virgin	\$400	4 hours	Allowed	Free
1	8	Jetstar	\$460	6 hours	Not allowed	Free
1	9	Qantas	\$460	4 hours	Allowed	Not Free
1	9	Virgin	\$400	6 hours	Allowed	Not Free
1	9	Jetstar	\$580	6 hours	Allowed	Free
1	10	Qantas	\$580	4 hours	Not allowed	Free
1	10	Virgin	\$580	4 hours	Allowed	Free
1	10	Jetstar	\$580	4 hours	Not allowed	Free
1	11	Qantas	\$580	6 hours	Not allowed	Not Free
1	11	Virgin	\$460	4 hours	Not allowed	Not Free
1	11	Jetstar	\$580	6 hours	Allowed	Free
1	12	Qantas	\$400	4 hours	Allowed	Not Free
1	12	Virgin	\$460	4 hours	Not allowed	Free
1	12	Jetstar	\$520	4 hours	Not allowed	Not Free
1	13	Qantas	\$520	4 hours	Allowed	Free
1	13	Virgin	\$460	6 hours	Not allowed	Free
1	13	Jetstar	\$460	6 hours	Not allowed	Free
1	14	Qantas	\$580	4 hours	Allowed	Free
1	14	Virgin	\$400	4 hours	Allowed	Not Free
1	14	Jetstar	\$400	4 hours	Allowed	Not Free
1	15	Qantas	\$520	4 hours	Not allowed	Free
1	15	Virgin	\$520	6 hours	Not allowed	Not Free
1	15	Jetstar	\$520	6 hours	Allowed	Not Free
1	16	Qantas	\$400	6 hours	Allowed	Free
1	16	Virgin	\$580	4 hours	Allowed	Not Free
1	16	Jetstar	\$520	6 hours	Allowed	Not Free

Block	Subset	Alternative	Fare	Flying time	Ticket Change	Food/Beverages
2	1	Qantas	\$460	6 hours	Not allowed	Not Free
2	1	Virgin	\$520	4 hours	Allowed	Not Free
2	1	Jetstar	\$520	6 hours	Not allowed	Free
2	2	Qantas	\$580	4 hours	Not allowed	Not Free
2	2	Virgin	\$400	6 hours	Not allowed	Free
2	2	Jetstar	\$460	6 hours	Allowed	Not Free
2	3	Qantas	\$580	4 hours	Allowed	Not Free
2	3	Virgin	\$580	6 hours	Not allowed	Not Free
2	3	Jetstar	\$520	6 hours	Not allowed	Free
2	4	Qantas	\$520	4 hours	Allowed	Not Free
2	4	Virgin	\$520	4 hours	Allowed	Free
2	4	Jetstar	\$580	4 hours	Allowed	Not Free
2	5	Qantas	\$520	6 hours	Allowed	Free
2	5	Virgin	\$400	4 hours	Not allowed	Not Free
2	5	Jetstar	\$580	6 hours	Not allowed	Not Free
2	6	Qantas	\$400	4 hours	Not allowed	Free
2	6	Virgin	\$460	6 hours	Allowed	Not Free
2	6	Jetstar	\$580	6 hours	Not allowed	Not Free
2	7	Qantas	\$400	6 hours	Allowed	Not Free
2	7	Virgin	\$400	6 hours	Not allowed	Not Free
2	7	Jetstar	\$400	4 hours	Not allowed	Free
2	8	Qantas	\$460	4 hours	Allowed	Free
2	8	Virgin	\$580	4 hours	Not allowed	Not Free
2	8	Jetstar	\$460	4 hours	Not allowed	Not Free
2	9	Qantas	\$580	6 hours	Allowed	Free
2	9	Virgin	\$460	6 hours	Allowed	Free
2	9	Jetstar	\$520	4 hours	Allowed	Free
2	10	Qantas	\$400	6 hours	Not allowed	Not Free
2	10	Virgin	\$580	6 hours	Not allowed	Free
2	10	Jetstar	\$580	4 hours	Allowed	Not Free
2	11	Qantas	\$460	6 hours	Allowed	Not Free
2	11	Virgin	\$460	4 hours	Allowed	Free
2	11	Jetstar	\$460	6 hours	Allowed	Not Free
2	12	Qantas	\$580	6 hours	Not allowed	Free
2	12	Virgin	\$520	6 hours	Allowed	Not Free
2	12	Jetstar	\$460	4 hours	Not allowed	Not Free
2	13	Qantas	\$520	4 hours	Not allowed	Not Free
2	13	Virgin	\$460	4 hours	Allowed	Not Free
2	13	Jetstar	\$400	4 hours	Not allowed	Free
2	14	Qantas	\$520	6 hours	Not allowed	Free
2	14	Virgin	\$580	4 hours	Not allowed	Free
2	14	Jetstar	\$400	6 hours	Allowed	Free
2	15	Qantas	\$400	4 hours	Allowed	Free
2	15	Virgin	\$520	6 hours	Allowed	Free
2	15	Jetstar	\$400	6 hours	Allowed	Free
2	16	Qantas	\$460	4 hours	Not allowed	Free
2	16	Virgin	\$400	4 hours	Not allowed	Free
2	16	Jetstar	\$520	4 hours	Allowed	Free

## Appendix 2 Experimental Design for Stage 2 Tasks

Session	Subset	Alternative	Fare	Flying time	Ticket Change	Food/Beverages
1	1	Qantas	\$580	4 hours	Not allowed	Free
1	1	Virgin	\$580	4 hours	Not allowed	Free
1	1	Jetstar	\$400	4 hours	Allowed	Not Free
1	2	Qantas	\$460	6 hours	Allowed	Free
1	2	Virgin	\$460	6 hours	Not allowed	Not Free
1	2	Jetstar	\$400	6 hours	Allowed	Free
1	3	Qantas	\$580	4 hours	Allowed	Not Free
1	3	Virgin	\$460	6 hours	Allowed	Free
1	3	Jetstar	\$520	4 hours	Allowed	Not Free
1	4	Qantas	\$460	6 hours	Not allowed	Not Free
1	4	Virgin	\$580	4 hours	Allowed	Not Free
1	4	Jetstar	\$400	6 hours	Not allowed	Not Free
1	5	Qantas	\$400	4 hours	Not allowed	Not Free
1	5	Virgin	\$400	6 hours	Allowed	Not Free
1	5	Jetstar	\$580	6 hours	Not allowed	Not Free
1	6	Qantas	\$520	6 hours	Allowed	Not Free
1	6	Virgin	\$520	4 hours	Allowed	Free
1	6	Jetstar	\$580	4 hours	Allowed	Not Free
1	7	Qantas	\$400	4 hours	Allowed	Free
1	7	Virgin	\$520	4 hours	Not allowed	Not Free
1	7	Jetstar	\$520	4 hours	Not allowed	Free
1	8	Qantas	\$520	6 hours	Not allowed	Free
1	8	Virgin	\$400	6 hours	Not allowed	Free
1	8	Jetstar	\$580	4 hours	Not allowed	Free
1	9	Qantas	\$580	6 hours	Allowed	Free
1	9	Virgin	\$400	4 hours	Allowed	Not Free
1	9	Jetstar	\$520	6 hours	Allowed	Free
1	10	Qantas	\$460	4 hours	Not allowed	Free
1	10	Virgin	\$520	6 hours	Allowed	Free
1	10	Jetstar	\$400	4 hours	Not allowed	Free
1	11	Qantas	\$580	6 hours	Not allowed	Not Free
1	11	Virgin	\$520	6 hours	Not allowed	Not Free
1	11	Jetstar	\$460	4 hours	Not allowed	Free
1	12	Qantas	\$460	4 hours	Allowed	Not Free
1	12	Virgin	\$400	4 hours	Not allowed	Free
1	12	Jetstar	\$460	6 hours	Not allowed	Not Free
1	13	Qantas	\$400	6 hours	Allowed	Not Free
1	13	Virgin	\$580	6 hours	Not allowed	Free
1	13	Jetstar	\$580	6 hours	Allowed	Free
1	14	Qantas	\$520	4 hours	Not allowed	Not Free
1	14	Virgin	\$460	4 hours	Not allowed	Not Free
1	14	Jetstar	\$520	6 hours	Not allowed	Not Free
1	15	Qantas	\$400	6 hours	Not allowed	Free
1	15	Virgin	\$460	4 hours	Allowed	Free
1	15	Jetstar	\$460	4 hours	Allowed	Not Free
1	16	Qantas	\$520	4 hours	Allowed	Free
1	16	Virgin	\$580	6 hours	Allowed	Not Free
1	16	Jetstar	\$460	6 hours	Allowed	Free

Session	Subset	Alternative	Fare	Flying time	Ticket Change	Food/Beverages
2	1	Qantas	\$400	4 hours	Not allowed	Not Free
2	1	Virgin	\$400	4 hours	Not allowed	Not Free
2	1	Jetstar	\$580	4 hours	Allowed	Free
2	2	Qantas	\$520	6 hours	Allowed	Not Free
2	2	Virgin	\$520	6 hours	Not allowed	Free
2	2	Jetstar	\$580	6 hours	Allowed	Not Free
2	3	Qantas	\$400	4 hours	Allowed	Free
2	3	Virgin	\$520	6 hours	Allowed	Not Free
2	3	Jetstar	\$460	4 hours	Allowed	Free
2	4	Qantas	\$520	6 hours	Not allowed	Free
2	4	Virgin	\$400	4 hours	Allowed	Free
2	4	Jetstar	\$580	6 hours	Not allowed	Free
2	5	Qantas	\$580	4 hours	Not allowed	Free
2	5	Virgin	\$580	6 hours	Allowed	Free
2	5	Jetstar	\$400	6 hours	Not allowed	Free
2	6	Qantas	\$460	6 hours	Allowed	Free
2	6	Virgin	\$460	4 hours	Allowed	Not Free
2	6	Jetstar	\$400	4 hours	Allowed	Free
2	7	Qantas	\$580	4 hours	Allowed	Not Free
2	7	Virgin	\$460	4 hours	Not allowed	Free
2	7	Jetstar	\$460	4 hours	Not allowed	Not Free
2	8	Qantas	\$460	6 hours	Not allowed	Not Free
2	8	Virgin	\$580	6 hours	Not allowed	Not Free
2	8	Jetstar	\$400	4 hours	Not allowed	Not Free
2	9	Qantas	\$400	6 hours	Allowed	Not Free
2	9	Virgin	\$580	4 hours	Allowed	Free
2	9	Jetstar	\$460	6 hours	Allowed	Not Free
2	10	Qantas	\$520	4 hours	Not allowed	Not Free
2	10	Virgin	\$460	6 hours	Allowed	Not Free
2	10	Jetstar	\$580	4 hours	Not allowed	Not Free
2	11	Qantas	\$400	6 hours	Not allowed	Free
2	11	Virgin	\$460	6 hours	Not allowed	Free
2	11	Jetstar	\$520	4 hours	Not allowed	Not Free
2	12	Qantas	\$520	4 hours	Allowed	Free
2	12	Virgin	\$580	4 hours	Not allowed	Not Free
2	12	Jetstar	\$520	6 hours	Not allowed	Free
2	13	Qantas	\$580	6 hours	Allowed	Free
2	13	Virgin	\$400	6 hours	Not allowed	Not Free
2	13	Jetstar	\$400	6 hours	Allowed	Not Free
2	14	Qantas	\$460	4 hours	Not allowed	Free
2	14	Virgin	\$520	4 hours	Not allowed	Free
2	14	Jetstar	\$460	6 hours	Not allowed	Free
2	15	Qantas	\$580	6 hours	Not allowed	Not Free
2	15	Virgin	\$520	4 hours	Allowed	Not Free
2	15	Jetstar	\$520	4 hours	Allowed	Free
2	16	Qantas	\$460	4 hours	Allowed	Not Free
2	16	Virgin	\$400	6 hours	Allowed	Free
2	16	Jetstar	\$520	6 hours	Allowed	Not Free

## Appendix 3 Stage 1 Consumer Survey Screenshots



Thank you for doing this survey.

The survey will take about 15 to 20 minutes to complete and is being conducted by the Centre for the Study of Choice at the University of Technology Sydney (UTS).

Please take as much time as you need to answer the questions. Most questions only require you to tick a box, but a few questions will ask you to type in a response. All your answers to the questions are strictly anonymous. No one will contact you after the survey, and no sales solicitation is involved. Your answers will be used for research purposes only.

The survey begins with a few simple demographic questions about you. Please DO NOT USE the 'Back' and 'Forward' buttons in your browser. Please use the buttons at the bottom of each screen.

If you want to pause the survey to return to it later, simply close the window and click on the original link in the invitation. It will return you to the last point of entry in the survey.

*Please click on " >> " to start the survey.*



First, we will ask several simple questions about yourself.

Are you...?

- ☐ Male  
☐ Female

To which age group do you belong?

- ☐ Under 18  
☐ 18-24 years  
☐ 25-34 years  
☐ 35-44 years  
☐ 45-54 years  
☐ 55-64 years  
☐ 65-74 years  
☐ 75 years and over



Note: if respondents answer "Under 18", thank and terminate.



Where do you live?

- ☐ Sydney
- ☐ Other NSW
- ☐ Melbourne
- ☐ Other VIC
- ☐ Brisbane
- ☐ Other QLD
- ☐ Adelaide
- ☐ Other SA
- ☐ Perth
- ☐ Other WA
- ☐ Canberra
- ☐ Other ACT
- ☐ Hobart
- ☐ Other TAS
- ☐ Darwin
- ☐ Other NT

What is your postcode?

*Please type your four digit postcode number in the box below.*

Note: if respondents are not from Sydney, thank and terminate.

In the past 24 months, have you done any long-haul, cross-country travel by air where you purchased the tickets yourself?

*\* Below are some examples of long-haul, cross-country travel by air from Sydney:*

- from Sydney to Perth,
- from Sydney to Broom,
- from Sydney to Darwin, and
- from Sydney to Cairns

☐ Yes

☐ No

Are you planning to do any such travel by air in the next 24 months?

☐ Yes

☐ No



Note: if respondents choose "no" in both questions, thank and terminate.

Thank you for your responses so far.

This survey is the first of two surveys in the same study. Within the next two to three months, you will be invited again to participate in the second survey.

**The purpose of this first survey is to study how consumers make choices for cross country air travel.** You will be asked to make choices from 16 different choice scenarios. In each scenario, there are three flight offers, one from Qantas, one from Virgin Australia and one from Jetstar. You will be asked to select your most preferred offer and two other questions.

The second survey which will be conducted in the next two to three months is not a normal survey. Rather than considering it as an online survey, a more suitable description would be a *training program*.

In this training program, you will be presented with choice scenarios similar to the ones you see in the first survey and you will be asked to make predictions on how YOU THINK other consumers will make choices. You will then receive feedback to inform you how consumers actually make choices. You can learn from this feedback and improve your predictions. At the end of the training program, we will ask you to provide your thoughts about your learning experience.

We estimate the first survey will take about 15 to 20 minutes. The second survey (training program) may take a bit longer possibly 30 minutes.

Think carefully about the above information and the time that you may need to spend. Are you interested in participating in both surveys?

- ☐ Yes  
☐ No



Note: if respondents choose "no" in this question, thank and terminate.

In the following section, we will show you 16 choice scenarios.

Assume you have planned or are planning your next long-haul, cross country air travel. The holiday destination you choose may take any time between 4 to 6 hours from Sydney depending on the particular flight you choose.

What we want you to do is simple. In each choice scenario, we will show you offers from each of the three airlines *Qantas*, *Virgin Australia* and *Jetstar* for a return trip from Sydney to your holiday destination. Please select which options you would **MOST LIKELY** and **LEAST LIKELY** choose. In addition, we would like you to tell us your opinions on what other people would choose.

Please note, there are no right or wrong answers. We are only interested in your opinions.

Please click on ">>" to continue.



Scenario 1 of 16 :

	Qantas	Virgin Australia	Jetstar
Return Airfare	\$580	\$520	\$400
Flying Time	6 hours	4 hours	6 hours
Change Booking	Allowed	Not allowed	Not allowed
In-Flight Food and Beverages	Not Free	Free	Not Free

Which option would you **MOST LIKELY** choose?

- ☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Which option would you **LEAST LIKELY** choose?

- ☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Assume there are 100 other people also answering this question, how many of them do you think would choose each of the following options?

Please enter a number in every box and make sure the sum is 100. Assume all four boxes should have a number in it, excluding 0 and 100.

people would fly Qantas  
 people would fly Virgin Australia  
 people would fly Jetstar  
 people would not fly  
 =  0



Note: this is followed by scenarios 2 to 16.

The questions on the next few pages are exactly the same or very similar to those in the census of the population conducted by the Australian Bureau of Statistics. Your answers to these questions are confidential, and cannot be used to identify you personally.

Your answers will be used ONLY to compare answers of different types of people (e.g., compare younger people with older people; compare males with females) to understand any differences in preferences.

*Please click on " >> " to continue.*



What is your present marital status?

- ☐ Never married
- ☐ Widowed
- ☐ Divorced
- ☐ Separated but not divorced
- ☐ Married
- ☐ Living with long term partner

Which of the following best describes your work status?

- ☐ Employed Full time
- ☐ Employed Part Time
- ☐ Unemployed
- ☐ Not in the labour force - Retired from labour force
- ☐ Not in the labour force - Intends to look for work in the future
- ☐ Not in the labour force - Has never worked



What is the highest year of school you have completed?

- ☐ Year 12 or equivalent
- ☐ Year 11 or equivalent
- ☐ Year 10 or equivalent
- ☐ Year 9 or equivalent
- ☐ Year 8 or equivalent
- ☐ Year 7 or below
- ☐ Did not go to school

What is the highest non-school qualification you have?

- ☐ Postgraduate Degree or equivalent
- ☐ Graduate Diploma and Graduate Certificate from university or equivalent
- ☐ Bachelor Degree or equivalent
- ☐ Advanced Diploma and Diploma from university/TAFE or equivalent
- ☐ Certificate or equivalent (e.g. Certificate III & IV or Certificate I & II)
- ☐ None of the above



D7 Which one of the following categories best describes your annual gross personal income (before tax)?

- ☐ Nil Income
- ☐ \$1-\$7,799 (i.e. \$1-\$149 a week)
- ☐ \$7,800-\$12,999 (i.e. \$150-\$249 a week)
- ☐ \$13,000-\$20,799 (i.e. \$250-\$399 a week)
- ☐ \$20,800-\$31,199 (i.e. \$400-\$599 a week)
- ☐ \$31,200-\$41,599 (i.e. \$600-\$799 a week)
- ☐ \$41,600-\$51,999 (i.e. \$800-\$999 a week)
- ☐ \$52,000-\$67,599 (i.e. \$1,000-\$1,299 a week)
- ☐ \$67,600-\$83,199 (i.e. \$1,300-\$1,599 a week)
- ☐ \$83,200-\$103,999 (i.e. \$1,600-\$1,999 a week)
- ☐ \$104,000 or more (i.e. \$2,000 or more a week)
- ☐ Prefer not to answer



D8 Which one of the following categories best describes your annual total household gross income (before tax)?

- ☐ Nil Income
- ☐ \$1-\$7,799 (i.e. \$1-\$149 a week)
- ☐ \$7,800-\$12,999 (i.e. \$150-\$249 a week)
- ☐ \$13,000-\$18,199 (i.e. \$250-\$349 a week)
- ☐ \$18,200-\$25,999 (i.e. \$350-\$499 a week)
- ☐ \$26,000-\$33,799 (i.e. \$500-\$649 a week)
- ☐ \$33,800-\$41,599 (i.e. \$650-\$799 a week)
- ☐ \$41,600-\$51,999 (i.e. \$800-\$999 a week)
- ☐ \$52,000-\$62,399 (i.e. \$1,000-\$1,199 a week)
- ☐ \$62,400-\$72,799 (i.e. \$1,200-\$1,399 a week)
- ☐ \$72,800-\$88,399 (i.e. \$1,400-\$1,699 a week)
- ☐ \$88,400-\$103,999 (i.e. \$1,700-\$1,999 a week)
- ☐ \$104,000-\$129,999 (i.e. \$2,000-\$2,499 a week)
- ☐ \$130,000-\$155,999 (i.e. \$2,500-\$2,999 a week)
- ☐ \$156,000-\$181,999 (i.e. \$3,000-\$3,499 a week)
- ☐ \$182,000-\$207,999 (i.e. \$3,500-\$3,999 a week)
- ☐ \$208,000 or more (i.e. \$4,000 or more a week)
- ☐ Prefer not to answer

D9 Which of the following best describes your household?

**Family household**

- ☐ Couple family with no children
- ☐ Couple family with children
- ☐ One parent family
- ☐ Other family household

**Other household**

- ☐ Single person household
- ☐ Group household (i.e. shared)



This concludes the survey. Thank you very much for your valuable time and feedback.

*Please click on " >> " to claim your points.*



## Appendix 4 Stage 2 Learning Experiment Screenshots



Thank you for participating in this study.

The study is being conducted by the *Centre for the Study of Choice (CenSoC)* at the University of Technology Sydney (UTS). You are invited because you agreed to participate in a survey you completed about three months' ago on cross country travel.

In the *last* survey, you completed **16** different choice scenarios containing flight offers from Qantas, Virgin Australia and Jetstar. You selected the *most preferred* and *least preferred* options in each scenario. We also asked you to predict other consumers' choices for the same scenarios.

**This second survey is a training program designed to help you make more accurate predictions of consumer choices.** You will be asked to predict consumer choices in 32 choice scenarios and receive feedback on your performance. The findings of this study may help to improve customer service.

We think this training program may take you about 30 minutes. To achieve the BEST learning results, we recommend you complete all tasks in ONE session without interruption.

*If you are ready, please click on " >> " to start.*



## Common to All Learning Approaches



Hide in Live Survey - Learning version

- ☒ Learning from Practising Online DSS (control condition)
- ☐ Learning from receiving outcome feedback after each task
- ☐ Learning from comparing learner model with consumer model (cognitive feedback 1)
- ☐ Learning from segment classification and segment similarities (cognitive feedback 2)



Note: this is a hidden question not shown to respondents. Respondents are randomly assigned to one of the four learning approaches (experimental conditions) by the tutoring system.

## Learning Approach Four



Hide in Live Survey – randomly select one segment to learn

- ☐ Group A
- ☒ Group B
- ☐ Group C



Note: this is a hidden question not shown to respondents. Respondents are randomly assigned to learn one of the three consumer groups.



## Your Predictions in the Last Survey



Note: in the following section, all respondents are provided with general feedback to what they completed in the Stage 1 survey to warm up for training.



## Common to All Learning Approaches



In the last survey, you completed 16 choice scenarios like the example below. You provided your own choices as well as predicted other consumers' choices. Using information we collected from many consumers including you, we developed a model which can closely predict consumers' choices in ANY choice scenario.

	Qantas	Virgin Australia	Jetstar
Return Airfare	\$460	\$520	\$520
Flying Time	6 hours	4 hours	6 hours
Change Booking	Not allowed	Allowed	Not allowed
In-Flight Food and Beverages	Not Free	Not Free	Free

Which option would you **MOST LIKELY** choose?

- ☐ Fly Qantas      ☐ Fly Virgin Australia      ☐ Fly Jetstar      ☐ Not Fly

Which option would you **LEAST LIKELY** choose?

- ☐ Fly Qantas      ☐ Fly Virgin Australia      ☐ Fly Jetstar      ☐ Not Fly

Assume there are 100 other people also answering this question, how many of them do you think would choose each of the following options?

*Please enter a number in every box and make sure the sum is 100. Assume all four boxes should have a number in it, excluding 0 and 100.*

<input type="text"/>	people would fly Qantas
<input type="text"/>	people would fly Virgin Australia
<input type="text"/>	people would fly Jetstar
<input type="text"/>	people would not fly
<hr/>	
<input type="text"/>	0

Please click on ">>" to continue.

<<	>>
----	----

## Common to All Learning Approaches



We compared your predictions, average predictions and our model predictions to consumers' actual choices.

### **Your predictions on consumer choices are more accurate than average predictions.**

To explain this in detail, we asked you to allocate 100 consumers to 4 choice options (e.g., fly Qantas) in 16 choice scenarios. In any scenarios, when you allocate 100 consumers to 4 choice options, your prediction of the number of consumers who would choose a particular choice option differs from the actual number by 15 consumers. The smaller this difference is, the better your predictions are. On average, predictions made by all respondents differ from actual choices by 16 consumers. The most accurate predictions are made by the model that we developed using everyone's choices. Predictions made by this model only differ from the real figures by less than 4 consumers for any scenarios.

The following program will train you to learn this model in order to make more accurate predictions.

*Please click on " >> " to continue.*



Note: this feedback describes each respondent's prediction performance made in the Stage 1 survey. Without introducing complex statistical concepts, each respondent's own performance is compared with the average performance of all respondents. This comparison is done on the difference between the predicted number of consumers choosing each travel option with actual number of consumers who selected each option in the Stage 1 survey. Because no respondents in the Stage 1 survey were found to be more accurate than the aggregated MNL model in prediction choices, the last part of the feedback regarding the model is identical for all respondents.

## Steps in This Training Program



Note: learners are informed beforehand on the steps they will go through in this training program. These steps are different for each learning approach.

In the following screens, a tutoring system for each of the four learning approaches will be shown separately from the beginning to the end of the program.

## Learning Approach One



### There are four steps in this training:

1. **Decision Tool Practice One:** You will be shown an [online decision support tool](#). You can learn consumer choices by practising with this tool. You can use drop-down menus to build any choice scenarios, and immediately observe changes in consumers' choices. Please note, because this is your **ONLY** learning tool to help you make predictions, we have set a minimum learning time at 2 minutes. We suggest that you spend more time learning it until you feel that you are ready to do the tasks.
2. **Session One:** You will be shown 16 choice scenarios and asked to make predictions. During this session you will not have access to the decision support tool.
3. **Decision Tool Practice Two:** You will have a second chance to practise with the [online decision support tool](#).
4. **Session Two:** You will be shown another 16 choice scenarios and asked to make predictions. Again, during this session, you will not have access to the decision support tool.

Please click on ">>" to start training.



## Learning Approach Two



### There are two steps in this training:

1. **Session One:** You will be shown 16 choice scenarios and asked to make predictions. After you complete each task you will be immediately provided with the correct answers.
2. **Session Two:** After Session One you will be asked a simple question. This will be followed by another 16 choice scenarios asking you to make predictions. Again, after each task you will be immediately provided with the correct answers.

*Please click on ">>" to start training.*



## Learning Approach Three



### There are four steps in this training:

1. **Feedback One:** you will receive feedback to your predictions in the last survey. We will compare your views on how consumers choose features such as fare and flying time with actual consumer choices on these features. Please study this information carefully before starting Session One.
2. **Session One:** you will be shown 16 choice scenarios and asked to make predictions. Between each task, you will NOT receive any feedback.
3. **Feedback Two:** you will receive feedback similar to Feedback One regarding your predictions in Session One. Please study this information carefully before starting Session Two.
4. **Session Two:** you will be shown another 16 choice scenarios and asked to make predictions. Between each task, you will NOT receive any feedback. After 16 tasks, you will receive feedback similar to Feedback One and Feedback Two, but there will be no further tasks.

Please click on " >> " to start training.





## Learning Approach Four



### There are four steps in this training:

1. **Learning Consumer Groups:** you will receive detailed descriptions of three consumer groups (Group A, B and C). We will inform you of the consumer group you belong to and the consumer group that you will be asked to predict in this training. Please study this information carefully before starting Session One.
2. **Session One:** you will be shown 16 choice scenarios and asked to make predictions of choices by a particular consumer group. After each task, you will receive a quick answer informing you whether you are predicting the choices the right group.
3. **Session One Performance:** after completing Session One, you will receive a performance review informing you how you progressed in Session One.
4. **Session Two:** you will be shown another 16 choice scenarios and asked to make predictions of the same consumer group. Again, after each task, you will receive a quick answer informing you whether you are predicting the right group. After 16 tasks, you will receive a performance review similar to the one in Step Three, but there will be no further tasks.

Please click on ">>" to start training.





# Decision Tool Practice One

*A minimum of 2 minutes' practice*



Note: a minimum of two minutes' practice was required for using DSS to learn. This is controlled by the delayed appearance of the "Next" button. This is to deter learners from going through this exercise too quickly or by accident without obtaining any useful information before starting to do the training tasks. There is no restriction on the maximum time allowed.



## Learning Approach One



The " >> " button will appear after 2 minutes.

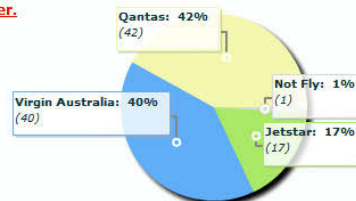
### Learning cross-country air travel consumer choices

1. Please use drop-down menus to build airline offers you are interested to know.

Qantas	Virgin Australia	Jetstar	Not Fly
<b>Return Airfare</b>	<b>Return Airfare</b>	<b>Return Airfare</b>	
\$400	\$400	\$400	
<b>Flying time</b>	<b>Flying time</b>	<b>Flying time</b>	
4 hours	4 hours	4 hours	
<b>Change Booking</b>	<b>Change Booking</b>	<b>Change Booking</b>	
Allowed	Allowed	Allowed	
<b>In-flight food and beverages</b>	<b>In-flight food and beverages</b>	<b>In-flight food and beverages</b>	
Free	Free	Free	

2. You can see immediately how many consumers will choose each offer.

Options	Choices
Fly Qantas	42%
Fly Virgin Australia	40%
Fly Jetstar	17%
Not Fly	1%



*Note: We strongly recommend that you spend enough time learning consumer choices using this system before you start training questions. To ensure this, there is a 2 minutes' delay before the ' >> ' button appearing below.*



# Feedback One

*Comparing your views with consumer choices*



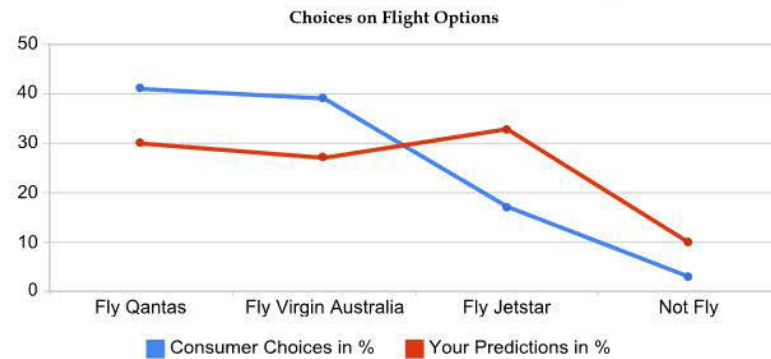
Note: this section provides learners with feed-forward information based on their prediction performance in the Stage One survey (labelled Session 0). Based on predictions in Session 0, a comparison is made between each learner's own model and the aggregated MNL model. The utilities of both models are transformed into probabilities for ease of explanation.

## Learning Approach Three

### Airline

Using your predictions in the last survey, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Flight Options	Consumer Choices	Your Predictions
Fly Qantas	41	30
Fly Virgin	39	27
Fly Jetstar	17	33
Not Fly	3	10



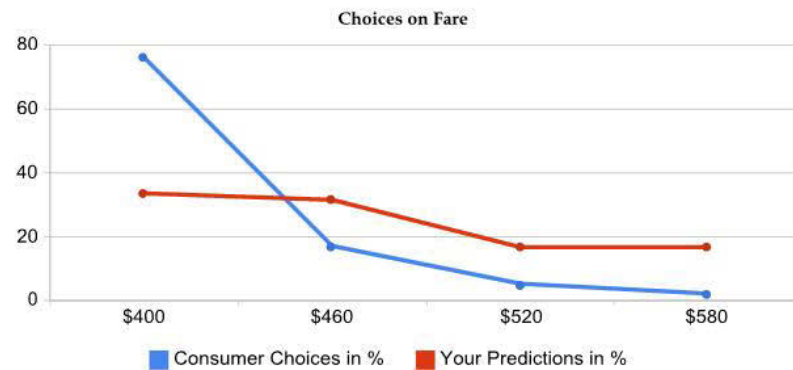
### Learning Approach Three (Continued)



#### Fare

Using your predictions in the last survey, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., flying time) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Fare	Consumer Choices	Your Predictions
\$400	76	34
\$460	17	32
\$520	5	17
\$580	2	17



## Learning Approach Three (Continued)

### Flying time

Using your predictions in the last survey, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Flying time	Consumer Choices	Your Predictions
4 hours	76	55
6 hours	24	45

<<

>>

### Learning Approach Three (Continued)



#### Ticket Change

Using your predictions in the last survey, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Ticket Change	Consumer Choices	Your Predictions
Allowed	61	49
Not allowed	39	51



## Learning Approach Three (Continued)



### In-flight food & beverages

Using your predictions in the last survey, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



In-flight food & beverages	Consumer Choices	Your Predictions
Free	70	58
Not free	30	42



### Learning Approach Three (Continued)



#### Importance of Features in Consumer Decision-Making

In making decisions, consumers think "airline" is the most important factor, accounting for 39% of total importance. This is followed by 35% for "fare", 12% for "flying time", 9% on "free in-flight food and beverages", and 5% for "allowing ticket change".

Importance of features in decision-making	Consumer Choices
Airline	39
Fare	35
Flying time	12
Ticket change	5
In-flight food & beverages	9

Please click on " >> " to continue.







# Learning Consumer Groups



## Learning Approach Four (Continued)

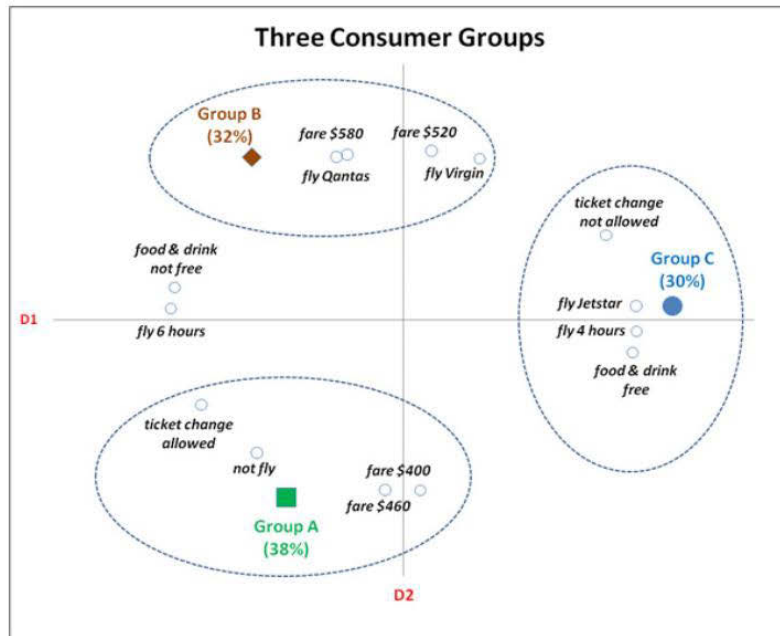


From responses we gathered, it is clear that there are **three** groups of consumers. We call them Group A, Group B and Group C. From your most/least preferred choices, we find you are MOST likely a **Group C** consumer, and you are LEAST likely a **Group B** consumer. **In this training program we will ask you to learn and predict the choices of Group C consumers.** Before you start the training tasks in Session One, **we strongly recommend that you spend enough time going through the information on this screen.**

The chart below shows

- 1) three consumer groups,
- 2) their preferences for four choices (fly Qantas, fly Virgin Australia, fly Jetstar and not fly), and
- 3) their connections to various flight features.

This should provide you with a direct visual description of the three consumer groups.



## Learning Approach Four (Continued)

### **Group A (38% of total consumers):**

Key words to describe Group A are "**cheaper fares**". The most obvious behaviour for this group is to choose cheaper fares, especially the cheapest fare. If an offer's return fare is at the lowest level (\$400), Group A consumers are 28 times more likely to choose rather than not choose the offer as their MOST preferred offer (the same ratio is only 2.7 times for Group B and 8.7 times for Group C). On the other hand, Group A's preference for higher fares like \$520 and \$580 is close to zero. In making decisions, Group A consumers treat "fare" as the most important factor, accounting for 48% of total importance. This is followed by 26% of total importance for "airline", 11% for "flying time", 7% for "allowing ticket change" and 8% for "free in-flight food and beverages".

### **Group B (32% of total consumers):**

Key words to describe Group B are "**airline and quality**". They strongly prefer better quality flight offers even if they have to pay higher fares. They almost always fly with Qantas and Virgin Australia (90%+ of total occasions). Their preference for low fare flight offers is much less than Group A and Group C (see previous example on \$400 fare in describing Group A). On the other hand, their preference for higher fare offers is much higher than the other two groups. For example, in contrast to Group A consumers who have almost 0% of occasions choosing higher fares (\$580 or \$520), 20% of Group B choices are made on these higher fares. In making decisions, Group B consumers think "airline" is the most important factor, accounting for 56% of total importance. This is followed by 23% for "fare", 10% for "flying time", 7% for "allowing ticket change" and 4% for "free in-flight food and beverages".

### **Group C (30% of total consumers):**

Key words to describe Group C are "**bargaining for better deals**". This group of consumers are not as extreme as Group A on fare and Group B on airline. Instead, they will look more into different features and look for better value for money. For example, they have a stronger preference than the other groups for free in-flight meals and drinks. When this service is offered, they are 2.5 times more likely to choose an offer (the same ratio is 1.5 times for Group A and 1.2 times for Group B). They also prefer flying for shorter time (4 hours). If flying time is 4 hours, they are 2.5 times more likely to choose the offer (the same ratio is 1.7 times for both Group A and Group B). On fare, they prefer the lowest fare but not at the level of Group A. In choosing airline, they choose Jetstar a lot more than the other two groups. In making decisions, Group C consumers think "airline" is the most important factor, accounting for 44% of importance. This is followed by 27% for "fare", 13% for "flying time", 4% for "allowing ticket change" and 13% for "free in-flight food and beverages".

## Learning Approach Four (Continued)

Tables below show summarise choices of three consumers groups' on features

### 1. Choices of "airline"

	Group A	Group B	Group C
Qantas	42%	50%	33%
Virgin Australia	38%	41%	41%
Jetstar	15%	9%	26%
Not fly	5%	1%	1%

### 2. Choices of "fare"

	Group A	Group B	Group C
\$400	91%	50%	78%
\$460	8%	31%	17%
\$520	0%	13%	4%
\$580	0%	6%	1%

### 3. Choices of "flying time"

	Group A	Group B	Group C
4 hours	75%	73%	86%
6 hours	25%	27%	14%

### 4. Choices of "ticket change"

	Group A	Group B	Group C
allowed	67%	65%	63%
not allowed	33%	35%	37%

### 5. Choices of "in-flight food and beverages"

	Group A	Group B	Group C
free	69%	60%	86%
not free	31%	40%	14%

If you are ready, please click on " >> " to continue.



Common to All Learning Approaches



# Session One

*16 choice scenarios*



Learning Approaches One, Two and Three (learners are asked to predict all consumers)



**Task 1 (Tasks 1 to 16 in Session One):**

	<b>Qantas</b> 	<b>Virgin Australia</b> 	<b>Jetstar</b> 
<b>Return Airfare</b>	\$580	\$580	\$400
<b>Flying Time</b>	4 hours	4 hours	4 hours
<b>Change Booking</b>	Not allowed	Not allowed	Allowed
<b>In-Flight Food and Beverages</b>	Free	Free	Not Free

Q1. If the above offerings are made in the market, which of the following choices do you think that consumers would **MOST LIKELY** make?

- ☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Q2. Out of **100** consumers, how many of them do you think would choose each option?

*Please enter a number in every box and make sure the sum is 100.*

people would fly Qantas  
 people would fly Virgin Australia  
 people would fly Jetstar  
 people would not fly  
 =  0



Learning Approaches Four (learners are asked to predict a particular group of consumers)



Task 1 (Tasks 1 to 16 in Session One)::

	<b>Qantas</b> 	<b>Virgin Australia</b> 	<b>Jetstar</b> 
<b>Return Airfare</b>	\$580	\$580	\$400
<b>Flying Time</b>	4 hours	4 hours	4 hours
<b>Change Booking</b>	Not allowed	Not allowed	Allowed
<b>In-Flight Food and Beverages</b>	Free	Free	Not Free

If you want to check details of the three consumer groups again, please click [here](#).

Q1. If the above offerings are made in the market, which of the following choices do you think that **Group C consumers** would **MOST LIKELY** make?

- ☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Q2. Out of **100 Group C consumers**, how many of them do you think would choose each option?

Please enter a number in every box and make sure the sum is 100.

people would fly Qantas  
 people would fly Virgin Australia  
 people would fly Jetstar  
 people would not fly  
 =  0



## Learning Approaches Two (feedback after each task)



### Task 1 (Tasks 1 to 16 in Session One):

	<b>Qantas</b> 	<b>Virgin Australia</b> 	<b>Jetstar</b> 
Return Airfare	\$580	\$580	\$400
Flying Time	4 hours	4 hours	4 hours
Change Booking	Not allowed	Not allowed	Allowed
In-Flight Food and Beverages	Free	Free	Not Free

### Correct Answers vs. Your Predictions

Options	Correct Choices	Your Predictions
Fly Qantas	8	15
Fly Virgin Australia	8	10
Fly Jetstar	83	60
Not Fly	1	15



## Learning Approaches Four (feedback after each task)



**Task 1 (Tasks 1 to 16 in Session One)::**

	Qantas 	Virgin Australia 	Jetstar 
<b>Return Airfare</b>	\$580	\$580	\$400
<b>Flying Time</b>	4 hours	4 hours	4 hours
<b>Change Booking</b>	Not allowed	Not allowed	Allowed
<b>In-Flight Food and Beverages</b>	Free	Free	Not Free

Based on your predictions, you are more likely making predictions on Group B consumers. Please pay attention to differences of Group B and Group C consumers.



Note: learners receive individualised feedback informing them whether they were predicting the correct consumer group or were predicting a different group. If they were predicting an incorrect group, they were asked to pay attention to the differences between the target group and the incorrect predicted group.

Learners continued with Tasks 2 to 16 until the Session One was completed. There was no feedback received between tasks. After the 16 Session One tasks, the program asked the following question before starting the Session Two training and tasks.

## Common to All Learning Approaches



After the first 16 tasks, what do you think about this learning approach?

- ☐ my predictions will improve using this learning approach
- ☐ my predictions will not improve using this learning approach





# Decision Tool Practice Two

*You can decide how long you want to practise*

*Please click on " >> " to continue.*



Note: for the second round of training on DSS, there was no restriction on the minimum time. Learners could totally control their own learning.

## Learning Approaches One



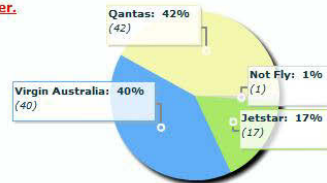
### Learning cross-country air travel consumer choices

1. Please use drop-down menus to build airline offers you are interested to know.

Qantas	Virgin Australia	Jetstar	Not Fly
			
<b>Return Airfare</b>	<b>Return Airfare</b>	<b>Return Airfare</b>	
\$400	\$400	\$400	
<b>Flying time</b>	<b>Flying time</b>	<b>Flying time</b>	
4 hours	4 hours	4 hours	
<b>Change Booking</b>	<b>Change Booking</b>	<b>Change Booking</b>	
Allowed	Allowed	Allowed	
<b>In-flight food and beverages</b>	<b>In-flight food and beverages</b>	<b>In-flight food and beverages</b>	
Free	Free	Free	

2. You can see immediately how many consumers will choose each offer.

Options	Choices
Fly Qantas	42%
Fly Virgin Australia	40%
Fly Jetstar	17%
Not Fly	1%



>>



## Feedback Two

*Comparing your predictions with consumer choices on each feature*



Note: after Session One, individual learner models were processed by the system and feedback was shown to learners in similar fashion to Feedback One.

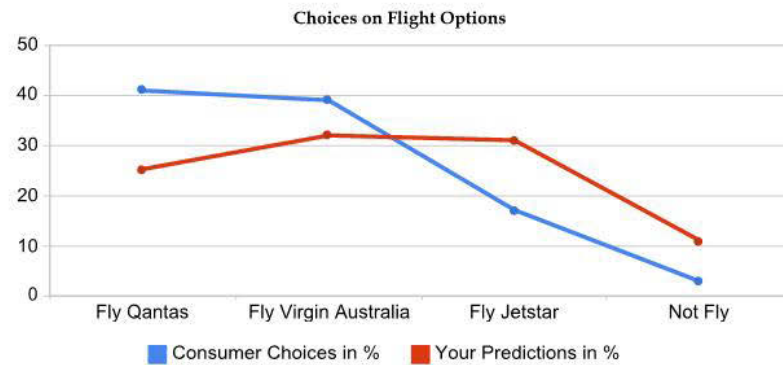
## Learning Approaches Three



### Airline

Using your predictions in Session One, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Flight Options	Consumer Choices	Your Predictions
Fly Qantas	41	25
Fly Virgin	39	32
Fly Jetstar	17	31
Not Fly	3	11



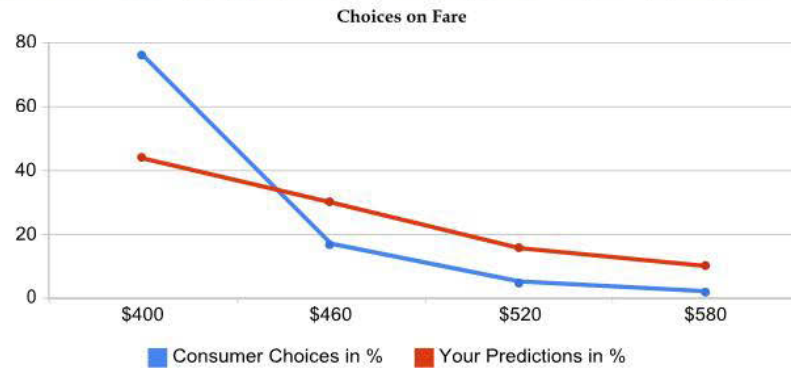
## Learning Approaches Three (Continued)



### Fare

Using your predictions in Session One, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., flying time) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Fare	Consumer Choices	Your Predictions
\$400	76	44
\$460	17	30
\$520	5	16
\$580	2	10



## Learning Approaches Three (Continued)



### Flying time

Using your predictions in Session One, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Flying time	Consumer Choices	Your Predictions
4 hours	76	53
6 hours	24	47





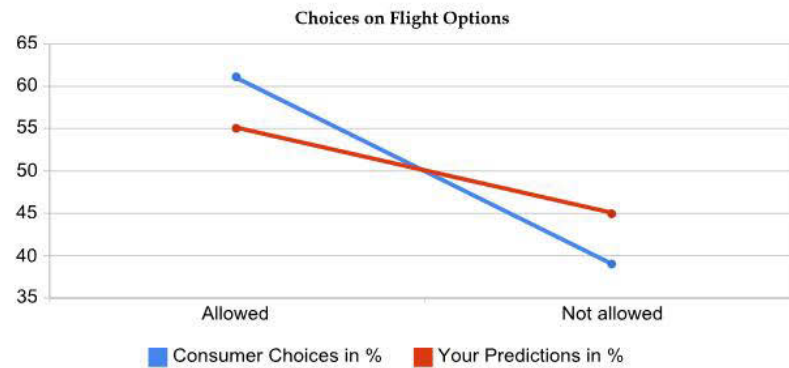
## Learning Approaches Three (Continued)



### Ticket Change

Using your predictions in Session One, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



Ticket Change	Consumer Choices	Your Predictions
Allowed	61	55
Not allowed	39	45



## Learning Approaches Three (Continued)



### In-flight food & beverages

Using your predictions in Session One, the chart below shows the differences between what you think consumers chose and what consumers actually chose.

If all other features (e.g., fare) are identical, choices implied by your predictions and consumers' actual choices are displayed below:



In-flight food & beverages	Consumer Choices	Your Predictions
Free	70	49
Not free	30	51



## Learning Approaches Three (Continued)



### Importance of Features in Decision-Making

In making decisions, consumers think "airline" is the most important factor, accounting for 39% of total importance. This is followed by 35% for "fare", 12% for "flying time", 9% for "free in-flight food and beverages", and 5% for "allowing ticket change".

Below is a table comparing importance of these features in your predictions with importance in consumer decision-making.

Importance of features in decision-making	Consumer Choices	Your Predictions
Airline	39	44
Fare	35	46
Flying time	12	4
Ticket change	5	5
In-flight food & beverages	9	1

Please click on " >> " to continue.





# Session One Performance



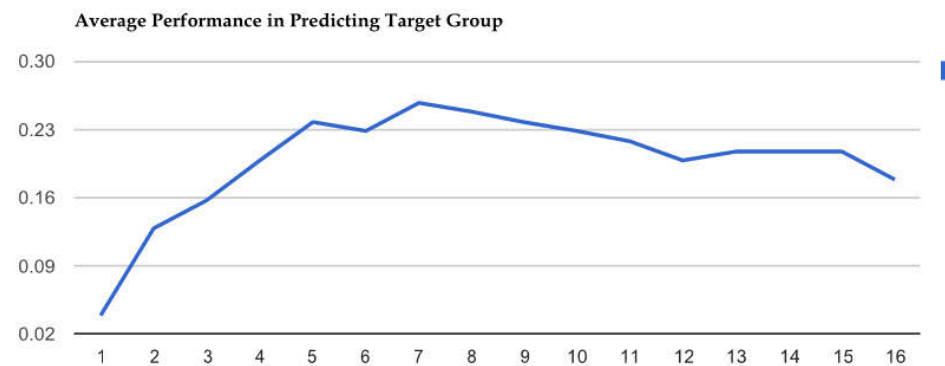
## Learning Approaches Four



### Summary of Your Performance in Session One

In total, you were predicting Group C consumers correctly 5 of 16 time(s) in Session One.

The chart below shows how your prediction accuracy increases or decreases over the 16 tasks. Movement of the line shows what your average performance is at a particular task with a score. For example, the position on Point 5 shows your average performance from choice scenario 1 to choice scenario 5.



Please click on ">>" to continue.



Common to All Learning Approaches



# Session Two


*16 choice scenarios*



Learning Approaches One, Two and Three (learners are asked to predict all consumers)



Task 17 (Tasks 17 to 32 in Session Two):

	<b>Qantas</b> 	<b>Virgin Australia</b> 	<b>Jetstar</b> 
<b>Return Airfare</b>	\$400	\$400	\$580
<b>Flying Time</b>	4 hours	4 hours	4 hours
<b>Change Booking</b>	Not allowed	Not allowed	Allowed
<b>In-Flight Food and Beverages</b>	Not Free	Not Free	Free

Q1. If the above offerings are made in the market, which of the following choices do you think that consumers would **MOST LIKELY** make?

- ☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Q2. Out of **100** consumers, how many of them do you think would choose each option?

*Please enter a number in every box and make sure the sum is 100.*

people would fly Qantas  
 people would fly Virgin Australia  
 people would fly Jetstar  
 people would not fly  
 =  0



Learning Approaches Four (learners are asked to predict a particular group of consumers)



Task 17 (Tasks 17 to 32 in Session Two):

	<div>Qantas</div> 	<div>Virgin Australia</div> 	<div>Jetstar</div> 
Return Airfare	\$400	\$400	\$580
Flying Time	4 hours	4 hours	4 hours
Change Booking	Not allowed	Not allowed	Allowed
In-Flight Food and Beverages	Not Free	Not Free	Free

If you want to check details of the three consumer groups again, please click [here](#).

Q1. If the above offerings are made in the market, which of the following choices do you think that **Group C consumers** would **MOST LIKELY** make?

- ☐ Fly Qantas
 ☐ Fly Virgin Australia
 ☐ Fly Jetstar
 ☐ Not Fly

Q2. Out of **100 Group C consumers**, how many of them do you think would choose each option?

Please enter a number in every box and make sure the sum is 100.

people would fly Qantas  
 people would fly Virgin Australia  
 people would fly Jetstar  
 people would not fly  
 =  0





Learning Approach Two (feedback after each task in the Session Two)



Task 17 (Tasks 17 to 32 in Session Two):

	Qantas 	Virgin Australia 	Jetstar 
Return Airfare	\$400	\$400	\$580
Flying Time	4 hours	4 hours	4 hours
Change Booking	Not allowed	Not allowed	Allowed
In-Flight Food and Beverages	Not Free	Not Free	Free

Correct Answers vs. Your Predictions

Options	Correct Choices	Your Predictions
Fly Qantas	50	60
Fly Virgin Australia	48	35
Fly Jetstar	2	5
Not Fly	0	0



## Learning Approach Four (feedback after each task in the Session One)



Task 17 (Tasks 17 to 32 in Session Two):

	<div><b>Qantas</b></div> 	<div><b>Virgin Australia</b></div> 	<div><b>Jetstar</b></div> 
<b>Return Airfare</b>	\$400	\$400	\$580
<b>Flying Time</b>	4 hours	4 hours	4 hours
<b>Change Booking</b>	Not allowed	Not allowed	Allowed
<b>In-Flight Food and Beverages</b>	Not Free	Not Free	Free

Based on your predictions, you are more likely making predictions on Group B consumers. Please pay attention to differences of Group B and Group C consumers.



## Learning Approach Four (feedback after each task in the Session One)



Task 25 (Tasks 17 to 32 in Session Two):

			
Return Airfare	\$400	\$580	\$460
Flying Time	6 hours	4 hours	6 hours
Change Booking	Allowed	Allowed	Allowed
In-Flight Food and Beverages	Not Free	Free	Not Free

Based on your predictions, you are making predictions on Group C consumers. It is the right consumer group.



Note: Learners continued with Tasks 18 to 32 until the Session Two was completed. There was no feedback received between tasks. After the 16 Session Two tasks, the program asked the following question before closing.

## Common to All Learning Approaches



After all 32 tasks, what do you think about this learning approach?

- ☐ my predictions will improve using this learning approach
- ☐ my predictions will not improve using this learning approach



This concludes the survey. Thank you very much for your valuable time and feedback.

*Please click on " >> " to claim your points.*



## Appendix 5 Socio-Demographic Background of Respondents

		Stage 1		Stage 2			
		n=485	LA1 (n=63)	LA2 (n=63)	LA3 (n=63)	LA4 (n=64)	Total (n=252)
<b>Gender (%)</b>							
Male		45%	38%	48%	54%	43%	46%
Female		55%	62%	52%	46%	57%	54%
<b>Age Group (%)</b>							
18-24 years		6%	3%	10%	5%	11%	7%
25-34 years		27%	29%	22%	35%	24%	27%
35-44 years		26%	25%	25%	21%	22%	23%
45-54 years		19%	21%	21%	22%	17%	20%
55-64 years		18%	17%	21%	16%	19%	18%
65-74 years		4%	3%	2%	2%	6%	3%
75 years and over		0%	2%				0%
<b>Where do you live? (%)</b>							
Sydney		100%	100%	100%	100%	100%	100%
<b>In the past 24 months, have you done any long-haul, cross-country travel by air where you purchased the tickets yourself? (%)</b>							
Yes		59%	63%	57%	65%	57%	61%
No		41%	37%	43%	35%	43%	39%
<b>Are you planning to do any such travel by air in the next 24 months? (%)</b>							
Yes		96%	97%	95%	92%	98%	96%
No		4%	3%	5%	8%	2%	4%

	Stage 1		Stage 2			
	n=485	LA1 (n=63)	LA2 (n=63)	LA3 (n=63)	LA4 (n=64)	Total (n=252)
<b>What is your present marital status?</b>						
Never married	21%	19%	17%	16%	29%	20%
Widowed	1%		2%			0%
Divorced	5%	8%	6%	2%	3%	5%
Separated but not divorced	3%	8%		2%	2%	3%
Married	59%	54%	62%	65%	62%	61%
Living with long term partner	11%	11%	13%	16%	5%	11%
<b>Which of the following best describes your work status?</b>						
Employed Full time	65%	67%	60%	65%	57%	62%
Employed Part Time	19%	17%	21%	22%	22%	21%
Unemployed	4%	5%	3%	5%	5%	4%
Not in the labour force - Retired from labour force	8%	5%	11%	8%	10%	8%
Not in the labour force - Intends to look for work in the future	4%	6%	5%		6%	4%
<b>What is the highest year of school you have completed?</b>						
Year 12 or equivalent	82%	81%	78%	89%	86%	83%
Year 11 or equivalent	4%	3%	3%	3%	5%	4%
Year 10 or equivalent	12%	13%	17%	6%	8%	11%
Year 9 or equivalent	1%				2%	0%
Year 8 or equivalent	1%		2%			0%
Year 7 or below	0%	2%		2%		1%
Did not go to school	0%	2%				0%

	Stage 1		Stage 2			
	n=485	LA1 (n=63)	LA2 (n=63)	LA3 (n=63)	LA4 (n=64)	Total (n=252)
<b>What is the highest non-school qualification you have?</b>						
Postgraduate Degree or equivalent	21%	27%	19%	14%	22%	21%
Graduate Diploma and Graduate Certificate from university or equivalent	9%	8%	8%	13%	6%	9%
Bachelor Degree or equivalent	31%	27%	33%	41%	33%	34%
Advanced Diploma and Diploma from university/TAFE or equivalent	14%	21%	10%	14%	13%	14%
Certificate or equivalent (e.g. Certificate III & IV or Certificate I & II)	13%	10%	16%	11%	14%	13%
None of the above	12%	8%	14%	6%	11%	10%
<b>Which one of the following categories best describes your annual gross personal income (before tax)?</b>						
Nil Income	2%	2%	3%	2%	3%	2%
\$1-\$7,799 (i.e. \$1-\$149 a week)	3%	5%	5%	3%	5%	4%
\$7,800-\$12,999 (i.e. \$150-\$249 a week)	2%	3%	3%	2%	3%	3%
\$13,000-\$20,799 (i.e. \$250-\$399 a week)	3%		3%	2%	3%	2%
\$20,800-\$31,199 (i.e. \$400-\$599 a week)	6%	10%	8%	5%	5%	7%
\$31,200-\$41,599 (i.e. \$600-\$799 a week)	5%	8%	6%	6%	8%	7%
\$41,600-\$51,999 (i.e. \$800-\$999 a week)	7%	2%	6%	10%	5%	6%
\$52,000-\$67,599 (i.e. \$1,000-\$1,299 a week)	11%	13%	5%	3%	11%	8%
\$67,600-\$83,199 (i.e. \$1,300-\$1,599 a week)	15%	11%	11%	14%	16%	13%
\$83,200-\$103,999 (i.e. \$1,600-\$1,999 a week)	14%	11%	14%	17%	11%	13%
\$104,000 or more (i.e. \$2,000 or more a week)	16%	16%	16%	19%	13%	16%
Prefer not to answer	15%	21%	19%	17%	17%	19%



	Stage 1		Stage 2			
	n=485	LA1 (n=63)	LA2 (n=63)	LA3 (n=63)	LA4 (n=64)	Total (n=252)
<b>Which one of the following categories best describes your annual total household gross income (before tax)?</b>						
Nil Income	0%	2%				0%
\$1-\$7,799 (i.e. \$1-\$149 a week)	0%			2%	2%	1%
\$7,800-\$12,999 (i.e. \$150-\$249 a week)	0%	2%				0%
\$13,000-\$18,199 (i.e. \$250-\$349 a week)	0%		2%			0%
\$18,200-\$25,999 (i.e. \$350-\$499 a week)	1%			2%		0%
\$26,000-\$33,799 (i.e. \$500-\$649 a week)	3%	5%			3%	2%
\$33,800-\$41,599 (i.e. \$650-\$799 a week)	2%	3%	2%	3%	3%	3%
\$41,600-\$51,999 (i.e. \$800-\$999 a week)	4%	8%	3%		5%	4%
\$52,000-\$62,399 (i.e. \$1,000-\$1,199 a week)	5%	2%	6%	5%	6%	5%
\$62,400-\$72,799 (i.e. \$1,200-\$1,399 a week)	5%	3%		5%	5%	3%
\$72,800-\$88,399 (i.e. \$1,400-\$1,699 a week)	6%	3%	8%	6%	8%	6%
\$88,400-\$103,999 (i.e. \$1,700-\$1,999 a week)	12%	8%	14%	5%	10%	9%
\$104,000-\$129,999 (i.e. \$2,000-\$2,499 a week)	13%	13%	13%	16%	16%	14%
\$130,000-\$155,999 (i.e. \$2,500-\$2,999 a week)	11%	6%	11%	14%	16%	12%
\$156,000-\$181,999 (i.e. \$3,000-\$3,499 a week)	8%	6%	5%	6%	5%	6%
\$182,000-\$207,999 (i.e. \$3,500-\$3,999 a week)	4%	5%	3%	5%		3%
\$208,000 or more (i.e. \$4,000 or more a week)	9%	14%	5%	16%	6%	10%
Prefer not to answer	17%	21%	29%	16%	16%	20%
<b>Which of the following best describes your household?</b>						
Couple family with no children	28%	30%	27%	32%	29%	29%
Couple family with children	43%	35%	46%	46%	43%	42%
One parent family	5%	6%	3%	5%	5%	5%
Other family household	4%	3%	6%		3%	3%
Single person household	10%	13%	6%	3%	8%	8%
Group household (i.e. shared)	10%	13%	11%	14%	13%	13%

## Bibliography

- Agresti, A. 2002, *Categorical data analysis*, 2nd edn, John Wiley & Sons, Hoboken, NJ.
- Anderson, J.R. 1978, 'Arguments concerning representations for mental imagery', *Psychological Review*, vol. 85, no. 4, pp. 249–76.
- Arunachalam, V. & Daly, B.A. 1996, 'An empirical investigation of judgment feedback and computerized decision support in a predicted task', *Accounting, Management Information Technology*, vol. 6, no. 3, pp. 139–56.
- Ashby, F.G. & Ell, S.W. 2001, 'The neurobiology of human category learning', *Trends in Cognitive Sciences*, vol. 5, no. 5, pp. 204–10.
- Atkinson, A.C. & Donev, A.N. 1992, *Optimum experimental designs*, Oxford University Press, Oxford.
- Balzer, W.K., Doherty, M.E. & O'Connor, R. 1989, 'Effects of cognitive feedback on performance', *Psychological Bulletin*, vol. 106, no. 3, pp. 410–33.
- Barber, L.K., Bagsby, P.G., Grawitch, M.J. & Buerck, J.P. 2011, 'Facilitating self-regulated learning with technology: Evidence for student motivation and exam improvement', *Teaching and Psychology*, vol. 38, no. 4, pp. 303–8.
- Baum, W.M. 1974, 'On two types of deviation from the matching law: Bias and undermatching', *Journal of the Experimental Analysis of Behavior*, vol. 22, no. 1, pp. 231–42.
- Bell, D.E., Raiffa, H. & Tversky, A. 1988, 'Descriptive, normative and prescriptive interactions in decision making', in D. Bell, H. Raiffa & A. Tversky (eds), *Decision making: Descriptive, normative and prescriptive interactions*, Cambridge University Press, Cambridge, pp. 9–30.
- Ben-Akiva, M. & Lerman, S.R. 1985, *Discrete choice analysis: Theory and application to travel demand*, MIT Press, Cambridge, MA.
- Ben-Akiva, M., McFadden, D., Gärling, T., Gopinath, D., Walker, J., Bolduc, D., Börsch-Supan, A., Delquie, P., Larichev, O., Morikawa, T., Polydoropoulou, A. & Rao, V. 1999, 'Extended framework for modeling choice behavior', *Marketing Letters*, vol. 10, no. 3, pp. 187–203.
- Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D.S., Daly, A., de Palma, A., Gopinath, D., Karlstrom, A. & Munizaga, M.A. 2002, 'Hybrid choice models: Progress and challenges', *Marketing Letters*, vol. 13, no. 3, pp. 163–75.
- Bickel, J.E. 2007, 'Some comparisons among quadratic, spherical, and logarithmic scoring rules', *Decision Analysis*, vol. 4, no. 2, pp. 49–65.
- Blattberg, R.C. & Hoch, S.J. 1990, 'Database models and managerial intuition: 50% model + 50% manager', *Management Science*, vol. 36, no. 8, pp. 887–99.
- Bliemer, M.C.J. & Rose, J.M. 2010, 'Serial choice conjoint analysis for estimating discrete choice models', in S. Hess & A. Daly (eds), *Choice modelling: The state-of-the-art and the state-of-practice* –

*Proceedings from the inaugural international choice modelling conference*, Emerald Group Publishing, Bingley, UK, pp. 139–62.

Bliemer, M.C.J. & Rose, J.M. 2011, 'Experimental design influences on stated choice outputs: An empirical study in air travel choice', *Transportation Research Part A*, vol. 45, pp. 63–79.

Bower, G.H. & Hilgard, E.R. 1981, *Theories of learning*, 5th edn, Prentice-Hall, NJ.

Bransford, J., Barron, B., Pea, R., Meltzoff, A., Kuhl, P., Bell, P., Stevens, R., Schwartz, D., Vye, N., Reeves, B., Roschelle, J. & Sabelli, N. 2006, 'Foundations and opportunities for an interdisciplinary science of learning', in R.K. Sawyer (ed.), *The Cambridge handbook of the learning sciences*, Cambridge University Press, Cambridge, pp. 19–24.

Brehmer, B. 1987, 'Note on subjects' hypotheses in multiple-cue probability learning', *Organizational Behavior and Human Decision Processes*, vol. 40, pp. 323–9.

Brier, G.W. 1950, 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3.

Burke, L.A. & Miller, M.K. 1999, 'Taking the mystery out of intuitive decision making', *The Academy of Management Executive*, vol. 13, no. 4, pp. 91–9.

Butler, D.L. & Winne, P.H. 1995, 'Feedback and self-regulated learning: A theoretical synthesis', *Review of Educational Research*, vol. 65, no. 3, pp. 245–81.

Camerer, C.F. & Johnson, E.J. 1997, 'The process-performance-paradox in expert judgment: How can experts know so much and predict so badly?', in W.M. Goldstein & R.M. Hogarth (eds), *Research on judgment and decision making – Currents, connections, and controversies*, Cambridge University Press, Cambridge, UK, pp. 342–64.

Carson, R., Bordes, N. & Pailthorpe, B.A. 1997, 'An archetypal representation of the history of the US economy', *Gather/ Scatter (San Diego Supercomputer Center)*, vol. 13, p. 14.

Castellan, N.J. 1974, 'The effect of different types of feedback in multiple-cue probability learning', *Organizational Behavior and Human Performance*, vol. 11, pp. 44–64.

Cherkassky, V. & Mulier, F. 2007, *Learning from data – Concepts, theory, and methods*, 2nd edn, Wiley-Interscience, Hoboken, NJ.

Cohen, G. 1983, *The psychology of cognition*, 2nd edn, Academic Press, London.

Cohen, H. & Lifebvre, C. 2005, 'Bridging the category divide', in H. Cohen & C. Lifebvre (eds), *Handbook of categorization in cognitive science*, Elsevier, Oxford, UK, pp. 2–15.

Cooksey, R.W. 1996, *Judgment analysis – Theory, methods, and applications*, Academic Press, San Diego, CA.

- Covin, J.G., Slevin, D.P. & Heeley, M.B. 2001, 'Strategic decision making in an intuitive vs. technocratic mode: Structural and environmental considerations', *Journal of Business Research*, vol. 52, no. 1, pp. 51–67.
- Cox, D.R. & Reid, N. 2000, *The theory of the design of experiments*, Chapman & Hall/CRC, Boca Raton, Florida.
- Cutler, A. & Breiman, L. 1994, 'Archetypal analysis', *Technometrics*, vol. 36, no. 4, pp. 338–47.
- Dawes, R. 1971, 'A case study of graduate admissions: Application of three principles of human decision making', *American Psychologist*, vol. 26, no. 2, pp. 180–8.
- Dawid, A.P., Lauritzen, S. & Parry, M. 2012, 'Proper local scoring rules on discrete sample spaces', *Annals of Statistics*, vol. 40, no. 1, pp. 593–608.
- de Palma, A., Ben-Akiva, M., Brownstone, D., Holt, C., Magnac, T., McFadden, D., Moffatt, P., Picard, N., Train, K., Wakker, P. & Walker, J. 2008, 'Risk, uncertainty and discrete choice models', *Marketing Letters*, vol. 19, pp. 269–85.
- Dhir, K.S. 2001, 'Enhancing management's understanding of operational research models', *Journal of the Operational Research Society*, vol. 52, no. 8, pp. 873–87.
- Dzyabura, D. & Hauser, J.R. 2011, 'Active machine learning for consideration heuristics', *Marketing Science*, vol. 30, no. 5, pp. 801–19.
- Edgell, S.E. 1978, 'Configural information processing in two-cue nonmetric multiple-cue probability learning', *Organizational Behavior and Human Performance*, vol. 22, pp. 404–16.
- Edgell, S.E. 1983, 'Delayed exposure to configural information in nonmetric multiple-cue probability learning', *Organizational Behavior and Human Performance*, vol. 32, pp. 55–65.
- Edwards, W. 1975, 'Cognitive processes and the assessment of subjective probability distributions: Comment', *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 291–3.
- Eisenstein, E.M. & Hutchinson, J.W. 2006, 'Action-based learning: Goals and attention in the acquisition of market knowledge', *Journal of Marketing Research*, vol. 43, no. 2, pp. 244–58.
- Eisenstein, E.M. & Lodish, L.M. 2002, 'Marketing decision support and intelligence systems: Precisely worthwhile or vaguely worthless?', in B. Weitz & R. Wensley (eds), *Handbook of marketing*, Sage Publications, London, pp. 436–56.
- Estes, W.K. 1950, 'Toward a statistical theory of learning', *Psychological Review*, vol. 57, pp. 94–107.
- Estes, W.K. 1972, 'Research and theory on the learning of probabilities', *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 81–102.
- Estes, W.K. 1976, 'The cognitive side of probability learning', *Psychological Review*, vol. 83, no. 1, pp. 37–64.

- Estes, W.K. 1994, 'Towards a statistical theory of learning', *Psychological Review*, vol. 101, no. 2, pp. 282–9.
- Estes, W.K. & Burke, C.J. 1953, 'A theory of stimulus variability in learning', *Psychological Review*, vol. 60, no. 4, pp. 276–86.
- Eugster, M. & Leisch, F. 2009, 'From Spider-Man to hero – Archetypal analysis in R', *Journal of Statistical Software*, vol. 30, no. 8, pp. 1–23.
- Evans, M., Hastings, N.A.J. & Peacock, J.B. 2000, *Statistical distributions*, 3rd edn, Wiley, New York.
- Fiebig, D., Keane, M., Louviere, J. & Wasi, N. 2010, 'The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity', *Marketing Science*, vol. 29, pp. 393–421.
- Field, A. 2005, *Discovering statistics using SPSS*, Sage Publications, London.
- Fishburn, P.C. 1986, 'The axioms of subjective probability', *Statistical Science*, vol. 1, no. 3, pp. 335–58.
- Friedman, D. 1983, 'Effective scoring rules for probabilistic forecasts', *Management Science*, vol. 29, no. 4, pp. 447–54.
- Gärdenfors, P. 2000, *Conceptual spaces – The geometry of thought*, MIT Press, Cambridge, MA.
- Gärdenfors, P. 2005, 'Concept learning and nonmonotonic reasoning', in H. Cohen & C. Lifebvre (eds), *Handbook of categorization in cognitive science*, Elsevier, Oxford, UK, pp. 824–42.
- Gärdenfors, P. 2008(a), 'Cognitive science: From computers to ant hills as models of human thoughts', in P. Gärdenfors & A. Wallin (eds), *A smorgasbord of cognitive science*, Bokförlager Nya Doxa, Nora.
- Gärdenfors, P. 2008(b), 'Concept learning', in P. Gärdenfors & A. Wallin (eds), *A smorgasbord of cognitive science*, Bokförlager Nya Doxa, Nora.
- Gärdenfors, P. & Williams, M.-A. 2001, 'Reasoning about categories in conceptual spaces', *Fourteenth international joint conference of artificial intelligence*, Morgan Kaufmann, San Francisco, California, pp. 385–92.
- Garrison, D.R. 1997, 'Self-directed learning: Toward a comprehensive model', *Adult Education Quarterly*, vol. 48, no. 1, pp. 18–33.
- Gneiting, T. & Raftery, A.E. 2007, 'Strictly proper scoring rules, prediction, and estimation', *Journal of American Statistical Association*, vol. 102, no. 477, pp. 359–78.
- Godfrey-Smith, P. 2003, *An introduction to the philosophy of science – Theory and reality*, University of Chicago Press, Chicago.
- Goldberg, L.R. 1970, 'Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences', *Psychological Bulletin*, vol. 73, pp. 422–32.

- Goldstein, W.M. & Hogarth, R.M. 1997, 'Judgment and decision research: Some historical context', in W.M. Goldstein & R.M. Hogarth (eds), *Research on judgment and decision making – Currents, connections, and controversies*, Cambridge University Press, Cambridge, UK, pp. 3–65.
- Greene, W.H. 2003, *Econometric analysis*, 5th edn, Pearson Education International, Upper Saddle River, New Jersey.
- Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E. & Nelson, C. 2000, 'Clinical versus mechanical prediction: A meta-analysis', *Psychological Assessment*, vol. 12, no. 1, pp. 19–30.
- Gulz, A. 2008, 'Researching virtual environments and virtual characters', in P. Gärdenfors & A. Wallin (eds), *A smorgasbord of cognitive science*, Bokförlager Nya Doxa, Nora, pp. 13–31.
- Hammond, K.R. 1955, 'Probabilistic functioning and the clinical method', *Psychological Review*, vol. 62, pp. 255–62.
- Hammond, K.R. & Stewart, T.R. 1975, 'Social judgment theory', in M.F. Kaplan & S. Schwartz (eds), *Human judgment and decision processes*, Academic Press, New York, pp. 271–312.
- Hammond, K.R., Summers, D.A. & Deane, D.H. 1973, 'Negative effects of outcome-feedback in multiple-cue probability learning', *Organizational Behavior and Human Performance*, vol. 9, pp. 30–4.
- Hampton, J.A. 1995, 'Testing the prototype theory of concepts', *Journal of Memory and Language*, vol. 34, pp. 686–708.
- Harnad, S. 2005, 'To cognize is to categorize: Cognition is categorization', in H. Cohen & C. Lifebvre (eds), *Handbook of categorization in cognitive science*, Elsevier, Oxford, UK, pp. 20–42.
- Hastie, R. & Dawes, R.M. 2001, *Rational choice in an uncertain world – The psychology of judgement and decision making*, Sage Publications, Thousand Oaks, CA.
- Hastie, T.J. & Tibshirani, R.J. 1994, 'Nonparametric regression and classification', in V. Cherkassky, J.H. Friedman & H. Wechsler (eds), *From statistics to neural networks – Theory and pattern recognition applications*, Springer-Verlag, Berlin; New York, pp. 62–82.
- Hastie, T., Tibshirani, R. & Friedman, J. 2009, *The elements of statistical learning – Data mining, inference, and prediction*, 2nd edn, Springer Science+Business Media, New York.
- Hattie, J. 1999, 'Influences on student learning', Inaugural Lecture, University of Auckland, Auckland.
- Hattie, J. & Timperley, H. 2007, 'The power of feedback', *Review of Educational Research*, vol. 77, no. 1, pp. 81–112.
- Hausman, J. 2001, 'Mismeasured variables in econometric analysis: Problems from the right and problems from the left', *The Journal of Economic Perspectives*, vol. 15, no. 4, pp. 57–67.
- Hausman, J. & McFadden, D. 1984, 'Specification tests for the multinomial logit model', *Econometrica*, vol. 52, no. 5, pp. 1219–40.

- Heckman, J.J. 1979, 'Sample selection bias as a specification error', *Econometrica*, vol. 47, no. 1, pp. 153–61.
- Heckman, J.J. & Hotz, V.J. 1989, 'Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training', *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 862–74.
- Hoch, S.J. & Kunreuther, H.C. 2001, 'A complex web of decisions', in S.J. Hoch, H.C. Kunreuther & R.E. Gunther (eds), *Wharton on making decisions*, John Wiley & Sons, New York, pp. 1–14.
- Hoch, S.J., Kunreuther, H.C. & Gunther, R.E. 2001, *Wharton on making decisions*, John Wiley & Sons, New York.
- Hoch, S.J. & Schkade, D.A. 1996, 'A psychological approach to decision support systems', *Management Science*, vol. 42, no. 1, pp. 51–64.
- Hogarth, R.M. 1975, 'Cognitive processes and the assessment of subjective probability distributions', *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 271–89.
- Hogarth, R.M. 1987, *Judgment and choice*, 2nd edn, John Wiley & Sons, New York.
- Huber, J. & Zwerina, K. 1996, 'The importance of utility balance in efficient choice designs', *Journal of Marketing Research*, vol. 33, no. 3, pp. 307–17.
- Hulse, S.H., Egeth, H. & Deese, J. 1980, *The psychology of learning*, McGraw-Hill, New York.
- Johnson-Laird, P.N. 1988, *The computer and the mind – An introduction to cognitive science*, Harvard University Press, Cambridge, MA.
- Jose, V.R.R., Nau, R.F. & Winkler, R.L. 2008, 'Scoring rules, generalized entropy and utility maximization', *Operations Research*, vol. 56, no. 5, pp. 1146–57.
- Kahneman, D. & Tversky, A. 1979, 'Prospect theory, an analysis of decision under risk', *Econometrica*, vol. 47, no. 2, pp. 263–91.
- Kamakura, W.A. & Russell, G.J. 1989, 'A probabilistic choice model for market segmentation and elasticity structure', *Journal of Marketing Research*, vol. 26, no. 4, pp. 379–90.
- Kayande, U., Bruyn, A.D., Lilien, G.L., Rangaswamy, A. & van Bruggen, G.H. 2009, 'How incorporating feedback mechanisms in a DSS affects DSS evaluations', *Information Systems Research*, vol. 20, no. 4, pp. 527–46.
- Kimble, G.A. 1985, 'The psychology of learning enters its second century', in B.L. Hammonds (ed.), *Psychology and learning – The master lecture series*, vol. 4, American Psychological Association, Washington, D.C.
- Klayman, J. 1984, 'Learning from feedback in probabilistic environments', *Acta Psychologica*, vol. 56, pp. 81–92.

- Kluger, A.N. & DeNisi, A. 1996, 'The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory', *Psychological Bulletin*, vol. 119, no. 2, pp. 254–84.
- Komatsu, L.K. 1992, 'Recent views of conceptual structure', *Psychological Bulletin*, vol. 112, no. 3, pp. 500–26.
- Kosslyn, S.M. 1981, 'The Medium and the Message in Mental Imagery: A Theory', *Psychological Review*, vol. 88, no. 1, pp. 46–66.
- Kosslyn, S.M. 1994, *Image and brain: The resolution of the imagery debate*, MIT Press, Cambridge, MA.
- Kosslyn, S.M., Ball, T.M. & Reiser, B.J. 1978, 'Visual images preserve metric spatial information: Evidence from studies of image scanning', *Journal of Experimental Psychology: Human Perception and Performance*, vol. 4, no. 1, pp. 47–60.
- Lancaster, K.J. 1966, 'A new approach to consumer theory', *Journal of Political Economy*, vol. 74, no. 2, pp. 132–57.
- Li, G., Wang, P., Carson, R. & Louviere, J. 2003, 'Archetypal analysis: A new way to segment markets based on extreme individuals', paper presented to the *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution*, ANZMAC, Adelaide, Australia.
- Lilien, G.L. & Rangaswamy, A. 2002, *Marketing engineering – Computer assisted marketing analysis and planning*, 2nd edn, Prentice Hall, Upper Saddle River, N.J.
- Lippmann, R.P. 1994, 'Neural networks, Bayesian a posterior probabilities, and pattern classification', in V. Cherkassky, J.H. Friedman & H. Wechsler (eds), *From statistics to neural networks – Theory and pattern recognition applications*, Springer-Verlag, Berlin; New York, pp. 83–104.
- Louviere, J.J., Hensher, D.A. & Swait, J.D. 2000, *Stated choice methods – Analysis and application*, Cambridge University Press, Cambridge, UK.
- Louviere, J.J. & Islam, T. 2008, 'A comparison of importance weights and willingness-to-pay measures derived from choice-based conjoint, constant sum scales and best-worst scaling', *Journal of Business Research*, vol. 61, no. 9, pp. 903–11.
- Louviere, J.J. & Meyer, R.J. 2008, 'Formal choice models of informal choices: What choice modeling research can (and can't) learn from behavioral theory', *Review of Marketing Research*, vol. 4, pp. 3–32.
- Louviere, J.J. & Woodworth, G. 1983, 'Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregated data', *Journal of Marketing Research*, vol. XX, no. November 1983, pp. 350–67.
- Luce, R.D. 1959, *Individual choice behavior*, John Wiley, Oxford, UK.
- Luce, R.D. & Raiffa, H. 1957, *Games and decisions – Introduction and critical survey*, John Wiley and Sons, New York.



Luger, G.F. 2009, *Artificial intelligence – Structures and strategies for complex problem solving*, 6th edn, Pearson Addison Wesley, Boston.

Luger, G.F., Johnson, P., Stern, C., Newman, J.E. & Yeo, R. 1994, *Cognitive science – The science of intelligent systems*, Academic Press, San Diego.

Manski, C. 1977, 'The structure of random utility models', *Theory and Decision*, vol. 8, pp. 229–54.

March, J.G. 1988, 'Bounded rationality, ambiguity, and the engineering of choice', in D. Bell, H. Raiffa & A. Tversky (eds), *Decision making: Descriptive, normative and prescriptive interactions*, Cambridge University Press, Cambridge, UK, pp. 33–57.

Markman, A.B. 1999, *Knowledge representation*, Lawrence Erlbaum Associates, Mahwah, NJ.

Markman, A.B. & Ross, B.H. 2003, 'Category use and category learning', *Psychological Bulletin*, vol. 129, no. 4, pp. 592–613.

McFadden, D. 1974, 'Conditional logit analysis of qualitative choice behavior', in P. Zarembka (ed.), *Frontiers of economics*, Academic Press, New York.

McFadden, D. 1979, 'Quantitative methods for analysing travel behaviour of individuals: Some recent developments', in D.A. Hensher & P.R. Stopher (eds), *Behavioural travel modelling*, Groom Helm, London, pp. 279–318.

McFadden, D., Train, K. & Tye, W. 1977, 'An application of diagnostic tests for the irrelevant alternatives property of the multinomial logit model', *Transportation Research Record*, vol. 637, pp. 39–46.

Medin, D.L. & Schaffer, M.M. 1978, 'Context theory of classification learning', *Psychological Review*, vol. 85, no. 3, pp. 207–38.

Meehl, P.E. 1954, *Clinical versus statistical prediction – A theoretical analysis and a review of the evidence*, University of Minnesota Press, Minneapolis.

Merrill, M.D. 1994, *Instructional design theory*, Educational Technology Publications, Englewood Cliffs, NJ.

Meyer, R.J. 1987, 'The learning of multiattribute judgment policies', *Journal of Consumer Research*, vol. 14, no. 2, pp. 155–73.

Mitchell, T.M. 1997, *Machine learning*, WCB/McGraw-Hill, Boston, MA.

Mowrer, R.R. & Klein, S.B. 2001, 'The transitive nature of contemporary learning theory', *Handbook of contemporary learning theories*, Lawrence Erlbaum Associates, London, pp. 1–22.

Nau, R.F. 1985, 'Should scoring rules be 'effective'?', *Management Science*, vol. 31, no. 5, pp. 527–35.

Norman, D.A. 1985, 'Twelve issues for cognitive science', in A.M. Aitkenhead & J.M. Slack (eds), *Issues in cognitive modeling: a reader*, Lawrence Erlbaum Associates, London, pp. 309–36.

- Nosofsky, R.M. 1986, 'Attention, similarity, and the identification-categorization relationship', *Journal of Experimental Psychology: General*, vol. 115, no. 1, pp. 39–57.
- O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. & Rakow, T. 2006, *Uncertain judgments: Eliciting experts' probabilities*, John Wiley & Sons, Chichester, UK.
- Pandya, A.S. & Macy, R.B. 1995, *Pattern recognition with neural networks in C++*, CRC Press, Boca Raton, FL.
- Payne, J.W., Bettman, J.R. & Johnson, E.J. 1993, *The adaptive decision maker*, Cambridge University Press, Cambridge, UK.
- Pekalska, E. & Duin, R.P.W. 2005, *The dissimilarity representation for pattern recognition – Foundations and applications*, World Scientific Publishing, Singapore.
- Pollard, D. 2002, *A user's guide to measure theoretic probability*, Cambridge University Press, Cambridge, UK.
- Powell, W.B. & Ryzhov, I.O. 2012, *Optimal learning*, John Wiley & Sons, Hoboken, NJ.
- Puckett, S.M. & Rose, J.M. 2010, 'Observed efficiency of a D-optimal design in an interactive agency choice experiment', in S. Hess & A. Daly (eds), *Choice modelling: The state-of-the-art and the state-of-practice – Proceedings from the inaugural international choice modelling conference*, Emerald Group Publishing, Bingley, UK, pp. 163–94.
- Pylyshyn, Z.W. 1973, 'What the mind's eye tells the mind's brain', *Psychological Bulletin*, vol. 80, no. 1, pp. 1–12.
- Pylyshyn, Z.W. 1981, 'The imagery debate: Analogue media versus tacit knowledge', *Psychological Review*, vol. 87, pp. 16–45.
- Revelt, D. & Train, K. 1998, 'Mixed logit with repeated choices: Households' choices of appliance efficiency level', *Review of Economics and Statistics*, vol. 80, pp. 647–57.
- Rosch, E.H. 1973, 'Natural categories', *Cognitive Psychology*, vol. 4, pp. 328–50.
- Rosch, E.H. 1975, 'Cognitive reference points', *Cognitive Psychology*, vol. 7, pp. 532–47.
- Rosch, E.H., Mervis, C., Gray, W., Johnson, D. & Boyes-Braem, P. 1976, 'Basic objects in natural categories', *Cognitive Psychology*, vol. 8, pp. 382–439.
- Rose, J.M., Bliemer, M.C.J., Hensher, D.A. & Collins, A.T. 2008, 'Designing efficient stated choice experiments in the presence of reference alternatives', *Transport Research Part B*, vol. 42, pp. 395–406.

- Rumelhart, D.E. & Norman, D.A. 1985, 'Representation of knowledge', in A.M. Aitkenhead & J.M. Slack (eds), *Issues in cognitive modeling: a reader*, Lawrence Erlbaum Associates, London, pp. 15-62.
- Sándor, Z. & Wedel, M. 2001, 'Designing conjoint choice experiments using managers' prior beliefs', *Journal of Marketing Research*, vol. 38, no. 4, pp. 430-44.
- Savage, L.J. 1954, *The foundations of statistics*, John Wiley & Sons, New York.
- Savage, L.J. 1971, 'Elicitation of personal probabilities and expectations', *Journal of American Statistical Association*, vol. 66, no. 336, pp. 783-801.
- Savage, L.J. 1972, *The foundations of statistics*, Dover Publications, New York.
- Scarpa, R., Campbell, D. & Hutchinson, G. 2007, 'Benefit estimates for landscape improvements: Sequential Bayesian design and respondents' rationality in a choice experiment', *Land Economics*, vol. 83, no. 4, pp. 617-34.
- Schmitt, N., Coyle, B.W. & King, L. 1976, 'Feedback and task predictability as determinants', *Organizational Behavior and Human Performance*, vol. 16, pp. 388-402.
- Searle, J.R. 1980, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417-57.
- Sengupta, K. 1995, 'Cognitive feedback in environments characterized by irrelevant information', *Omega, International Journal of Management Science*, vol. 23, no. 2, pp. 125-43.
- Simon, H.A. 1979, 'Information processing models of cognition', *Annual Review of Psychology*, vol. 30, pp. 363-96.
- Simon, H.A. 1983, 'Why should machine learn?', in R.S. Michalski, J.G. Carbonell & T.M. Mitchell (eds), *Machine learning – An artificial intelligence approach*, Morgan Kaufmann, San Francisco, California, pp. 25-38.
- Simon, H.A. 1997, *Models of bounded rationality: Empirically grounded economic reason*, MIT Press, Cambridge, Massachusetts.
- Skouras, K. & Dawid, P. 1999, 'On efficient probability forecasting system', *Biometrika*, vol. 86, no. 4, pp. 765-84.
- Slovic, P., Fishhoff, B. & Lichtenstein, S. 1977, 'Behavioral decision theory', *Annual Review of Psychology*, vol. 28, pp. 1-39.
- Slovic, P. & Lichtenstein, S. 1971, 'Comparison of Bayesian and regression approaches to the study of information processing in judgment', *Organizational Behavior and Human Performance*, vol. 6, pp. 649-744.
- Steinmann, D.O. 1976, 'The effects of cognitive feedback and task complexity in multiple-cue probability learning', *Organizational Behavior and Human Performance*, vol. 15, pp. 168-79.

Street, D. & Burgess, L. 2007, *The construction of optimal stated choice experiments – Theory and methods*, John Wiley & Sons, Hoboken, New Jersey.

Suppes, P. 1994, 'Qualitative theory of subjective probability', in G. Wright & P. Ayton (eds), *Subjective probability*, John Wiley & Sons, Chichester, UK.

Swait, J. & Andrews, R.L. 2003, 'Enriching scanner panel models with choice experiments', *Marketing Science*, vol. 22, no. 4, pp. 442–60.

Thaler, R.H. 1980, 'Toward a positive theory of consumer choice', *Journal of Economic Behavior and Organization*, vol. 1, no. 1, pp. 39–60.

Thurstone, L.L. 1927, 'A law of comparative judgment', *Psychological Review*, vol. 34, no. 4, pp. 273–86.

Tolman, E.C. 1948, 'Cognitive maps in rats and men', *Psychological Review*, vol. 55, no. 4, pp. 189–208.

Train, K.E. 2009, *Discrete choice methods with simulation*, 2nd edn, Cambridge University Press, Cambridge.

Tufte, E.R. 1983, *The visual display of quantitative information*, Graphics Press, Cheshire, CT.

Tufte, E.R. 1997, *Visual explanations – Images and quantities, evidence and narrative*, Graphics Press, Cheshire, CT.

Tversky, A. & Kahneman, D. 1974, 'Judgment under uncertainty: Heuristics and biases', *Science*, vol. 185, no. 4157, pp. 1124–31.

van Bruggen, G.H. & Wierenga, B. 2010, *Marketing decision making and decision support: challenges and perspectives for successful marketing management support systems*, Now Publishers, Hanover, MA 02339 USA.

Vanharanta, M. & Easton, G. 2010, 'Intuitive managerial thinking; the use of mental simulations in the industrial marketing context', *Industrial Marketing Management*, vol. 39, no. 3, pp. 425–36.

Vapnik, V.N. 1999, *The nature of statistical learning theory*, Springer, New York.

Vella, F. 1998, 'Estimating models with sample selection bias: A survey', *Journal of Human Resources*, vol. 33, no. 1, pp. 127–69.

von Neumann, J. & Morgenstern, O. 1944, *Theory of games and economic behavior*, Princeton University Press, Princeton, NJ.

Vovk, V. 2001, 'Competitive on-line statistics', *International Statistical Review*, vol. 69, no. 2, pp. 213–48.

Walker, J.L. 2001, 'Extended discrete choice models: Integrated framework, flexible error structures, and latent variables', PhD thesis, Massachusetts Institute of Technology.

- Wedel, M. & Kamakura, W. 2000, *Market segmentation – Conceptual and methodological foundations*, 2nd edn, Kluwer Academic Publishers, Norwell, Massachusetts.
- Wilhelm, A. 2005, 'Interactive statistical graphics: The paradigm of linked views', in C.R. Rao, E.J. Wegman & J.L. Solka (eds), *Data mining and data visualization*, Elsevier, Amsterdam, pp. 437–534.
- Winkler, R.L. 1996, 'Scoring rules and the evaluation of probabilities', *Test*, vol. 5, no. 1, pp. 1–60.
- Winkler, R.L. & Murphy, A. 1968, "'Good" probability assessors', *Journal of Applied Meteorology*, vol. 7, pp. 751–8.
- Woiceshyn, J. 2009, 'Lessons from "good minds": How CEOs use intuition, analysis and guiding principles to make strategic decisions', *Long Range Planning*, vol. 42, no. 3, pp. 298–319.
- Woolf, B.P. 2009, *Building intelligent interactive tutors – Student-centered strategies for revolutionizing e-learning*, Morgan Kaufmann Publishers, Burlington, MA.
- Yellott, J.I. 1977, 'The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution', *Journal of Mathematical Psychology*, vol. 15, no. 2, pp. 109–44.
- Zimmerman, B.J. 1990, 'Self-regulated learning and academic achievement: An overview', *Educational Psychologist*, vol. 25, no. 1, pp. 3–17.
- Zimmerman, B.J. 2008, 'Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects', *American Educational Research Journal*, vol. 45, no. 1, pp. 168–83.
- Zimmerman, B.J. & Martinez-Pons, M. 1990, 'Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use', *Journal of Educational Psychology*, vol. 82, no. 1, pp. 51–9.