A METHODOLOGICAL INVESTIGATION OF THE JUSTER SCALE: CONTEXTUAL REQUIREMENTS AND MUTUALLY EXCLUSIVE BEHAVIOURS

By

Mathew Parackal

A thesis submitted for the degree of
Doctor of Philosophy in Marketing at
Massey University
Palmerston North
New Zealand

April 2004

ABSTRACT

The relatively poor performance of intention scales to forecast future purchase behaviour turned researchers' attention to testing probability scales and the 11 point Juster Scale has become a preferred instrument for this task. The scale has undergone considerable testing and has been implemented successfully in a variety of research environments including self-completion mail surveys, telephone surveys and Internet-based surveys. Nevertheless, several methodological challenges remain, each of which produce some variation in the scale's accuracy. In particular, the review of Juster Scale literature revealed that accuracy of the scale was not consistent across product categories. This raised concerns about the reliability of the scale with both the context of the scale as it is presented to respondents and the nature of samples used to test the scale cited as possible causes for inconsistency. The review also identified two areas of development for the Juster Scale. These were to examine whether the structure of the scale could improve its performance and a problem that researchers encountered when using the Juster Scale to forecast mutually exclusive behaviours.

The research carried out for this thesis aimed to address two of the four issues raised above. They were, one, to address the contextual requirements of the Juster Scale and, two, to resolve the problem that researchers encountered when using the Juster Scale to forecast mutually exclusive behaviours. Data required to address the two issues were secured by implementing two Internet-based surveys. One was carried out on the clientele of Vodafone New Zealand (Vodafone survey) and the other on a sample of the national population (New Zealand survey). The test products were WAP-capable mobile phones and the payment plans offered by mobile telephony companies. Purchase probability data for these products were obtained in separate treatments to produce the required comparisons.

The review of literature identified three factors that exhibited tendencies to alter context, namely, question order, the practice of testing the Juster Scale concurrently on product categories and respondent's interpretation of the question accompanying the Juster Scale. Prior to addressing these issues, it was necessary to standardise the contextual requirements of the Juster Scale. Investigation was undertaken by implementing the Juster Scale in separate treatments with and without providing

additional contextual inputs. Results showed that the Juster Scale implemented on its own without additional contextual information produced mean probability scores similar to when the scale was implemented after contextual information was provided.

The Juster Scale has also been successfully employed in the forecasting of mutually exclusive behaviour. The review of literature revealed two methods namely weighting and the Constant Sum Scale for the above purpose. However, no comparisons were previously made to test whether the forecasts made by these two methods were similar or not, and this became the second major objective addressed in the current research. Investigation was undertaken by implementing the two methods in separate treatments (Weighted-scores and Constant Sum Scale) in the New Zealand survey. Results produced were mixed hence it was not possible to conclusively establish one method's superiority. The topic remains open for further investigation to test a method that is best suited for collecting probability data of mutually exclusive behaviour.

The investigation on the contextual requirements of the Juster Scale concluded, at least for the test products (WAP-capable mobile phones) used in this research, that the Juster Scale is a robust forecasting instrument in a typical purchasing environment. However, contextual requirements of the Juster Scale were examined here for just one product. Future studies might investigate whether the results obtained here can be reproduced for other product categories.

Results from the investigation to resolve the problem researchers encountered when using the Juster Scale in forecasting mutually exclusive behaviours were not conclusive. This research, however, showed that the Constant Sum Scale was a better instrument to produce clear data, ready for analysis. Nonetheless, the topic remains open for more investigation. In any future research, selection of test products will be crucial. Frequently purchased products may not generate the necessary comparisons.

The major contribution of this thesis to the academic community is that the Juster Scale is successful in collecting probability data in a purchasing context. While the objective regarding mutually exclusive behaviour yielded less conclusive results, the research showed that either of the two methods compared (Constant Sum Scale and weighting process) could be used when respondents are confident about their behaviour.

ACKNOWLEDGMENTS

My foremost acknowledgement goes to God Almighty through the Lord Jesus Christ. I have had the greatest experience of trusting in God all the years I have worked on this thesis. I would like to offer my praises to God with the opening verses of Psalm 103

> "Praise the Lord, O my soul; all my inmost being, praise his holy name.
> Praise the Lord, O my soul, and forget not all his benefits…"

I take this opportunity to express my appreciation to my wife, Sherly for her tireless love and support. She has with tremendous patience helped me in each stage of this thesis. Without her, I would not have completed this piece of work. My three daughters, Sharon, Sarah and Raquel, have been a source of motivation for me to complete this thesis. I've enjoyed the love and care of my family during times of difficulties and times of success. I would like to also mention my appreciation to fellow-believers at the Victoria Gospel Hall, Palmerston North and Mailer Street Gospel Hall, Dunedin. Thank you for the many prayers and help given; may God continue to bless and prosper your services that you do in the Lord Jesus.

I have had a great team of supervisors behind this work. To begin, I'd like to thank Dr Ron Garland for stepping in as my chief supervisor. Your guidance was precise and always right. Thanks for the timely feedback and comments; they were very useful. Next on my list is Associate Professor Tony Lewis; your arriving to my help was timely which turned this thesis around. Thank you for your advice and support, they had great value to this thesis. I was introduced to this research by Dr Mike Brennan; thanks Mike for setting me up with this research. Thanks for those early papers and conference experiences. One person who has followed me through my work is Mr Barry Jackson. I have used him as my punching bag; thanks Barry for keeping up with me. I would like to express my thanks to Professor Phil Gendall for offering me the opportunity to do my PhD in the department. I have enjoyed working under you and the experience gained in survey research has been very useful in doing the research for this thesis. I would like to convey my thanks to Dr Zane Kearns and Mr Don Esslemont. I thoroughly enjoyed learning under you both; thanks for your help, support and friendship. My thanks to Mrs Sue Johns for reading the thesis before I submitted; thank you very much Sue. There were some who silently did their part to make my PhD days go easy. Mrs Maureen

MacDonald, Ms Pru Robbie, Mrs Rosemary McDonald, Mrs Pat Collins and Mr Craig Goodwin, thanks for your best wishes and help from time to time.

In closing this brief acknowledgement, I would like to mention my thanks to Mum and Dad. They have sacrificially given a lot of themselves to me. They gave me a great home to grow up, good education and wonderful days. They have faithfully followed me with their prayers. My thanks also go to Sherly's Mum and Dad for their love and prayers. To all my brothers (Achachan, Geogie, Binod, and Babuchayan) and sisters (Sheena, Sheela, Elsa and Beena), thank you for praying for me. I am thankful to God Almighty for all you wonderful people in my life.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER TWO: PROBABILITY SCALES USED FOR FORECASTING PURCHASE BEHAVIOUR

CHAPTER THREE: PROBLEMS OF THE JUSTER SCALE

CHAPTER FOUR: INTERNET-BASED SURVEYS

CHAPTER FIVE: CONCEPTUAL FRAMEWORK AND HYPOTHESES

CHAPTER SIX: METHODOLOGY

CHAPTER SEVEN: CONTEXT OF THE JUSTER SCALE: RESULTS AND DISCUSSION

LIST OF TABLES

CHAPTER THREE: PROBLEMS OF THE JUSTER SCALE

CHAPTER FOUR: INTERNET-BASED SURVEY

CHAPTER SIX: METHODOLOGY

CHAPTER SEVEN: CONTEXT OF THE JUSTER SCALE: RESULTS AND DISCUSSION

CHAPTER EIGHT: MUTUALLY EXCLUSIVE BEHAVIOURS: RESULTS AND DISCUSSION

1. INTRODUCTION

1.1 Background

In survey research, the poor performance of intention scales to forecast purchase behaviour led researchers to test probability scales. The earlier studies that compared probability scales with intention scales found the former to be more accurate (Theil & Kosobud 1968; Juster 1966; Ferber & Piskie 1965). Since then, researchers have tested an eleven-point scale pioneered by Juster (1964) and found it to produce satisfactory results (Day *et al.* 1991; Gan, Esslemont & Gendall 1986; Clawson 1971; Gabor & Granger 1972; Gruber 1970; Clancy & Garsen 1970; Stapel 1968). This scale was named after its author and became the Juster Scale in the academic literature.

Researchers have customised the Juster Scale for use in self-completion questionnaires (Gendall, Esslemont & Day 1991), telephone surveys (Brennan, Esslemont & Hini 1995a) and Internet-based surveys (Parackal & Brennan 1999). Different versions of the scale have been successfully tested to forecast purchase rates (Gendall, Esslemont & Day 1991; Gan *et al.* 1986; Gabor & Granger 1972; Clawson 1971; Juster 1966), purchase levels (Brennan, Esslemont & U 1995b; Brennan & Esslemont 1994a; Seymour, Brennan & Esslemont 1994; Hamilton-Gibbs, Esslemont & McGuinness 1992), mutually exclusive behaviours (Parackal & Brennan, 1999; Hoek & Gendall, 1997a), demand schedules (Brennan 1995) and customer loyalty (Garland 2002; Danenberg & Sharp 1999; Riebe *et al.* 1998; Danenberg & Sharp 1996). In recent years the scale has been used in choice modelling studies (Rungie & Danenberg 1998) and in modelling repeat purchases using the Dirichlet model (Wright, Sharp & Sharp 2002).

The Juster Scale studies cited were successful in achieving their respective objectives. However, in some of these studies, the accuracy of the scale was not very impressive (Brennan, Esslemont & U 1995; Day *et al.* 1991; Gan *et al.* 1986; Clawson 1972). Also, the accuracy of the scale was seen to vary considerably across product categories. For example, forecasts of automobiles were reasonably accurate (Juster 1966; Stapel 1968; Pickering & Isherwood 1974; Gan *et al.* 1986), but forecasts of other durables were not so accurate (Brennan, Esslemont & Hini 1995c; Pickering & Isherwood 1974; Clawson

1971; Heald 1970; Juster 1966). This observation raised serious questions about the reliability of the Juster Scale.

In the review of literature carried out for this thesis, two issues were identified as causing some of the variations in the accuracy of the Juster Scale. They were the context of the Juster Scale and the nature of samples used in the Juster Scale studies. The review also identified two areas of development required for the scale. One was to improve the structure of the Juster Scale and the other was to address a problem faced when using the scale to collect probability data for mutually exclusive behaviours.

In the research carried out for this thesis, two of the above issues, namely, the context of the Juster Scale and the problem faced when using the scale to collect probability data for mutually exclusive behaviour were addressed. In the following sections, the two issues are introduced and the research objectives stated. In the final section, outlines of each chapter are provided.

1.2 Context of the Juster Scale

The academic literature on the context of questions (hereafter referred to as the contextual literature) suggests that context influences the responses obtained by survey questions (Schuman & Presser 1981). The contextual literature demonstrates that a question asked in different contexts produces different response distributions. Consequently, results based on such response distributions are incomparable (Schuman, Kalton, & Ludwig 1983; Sudman & Bradburn 1982; Schuman, Presser & Ludwig 1981; Schuman & Presser 1981; Duncan & Schuman 1980). Observations from the contextual studies, cited above, led to this thesis reviewing the Juster Scale studies from a contextual point of view. The review identified a number of factors that tended to modify the context of the Juster Scale, even causing the Juster Scale to be implemented in a context different to that which was originally intended. Results became specific to the context of the study and may not be comparable. If this were true for Juster Scale studies reviewed, then the scale would require fresh testing with a standardised context to produce comparable results.

Juster (1966) was of the view that asking questions about the individual's income, income prospects, asset shares, economy, and previous purchases, before presenting the Juster Scale could improve its accuracy. Pickering & Isherwood (1974) made similar observations about their probability scale. These were the earliest references to the context of probability scales. It was after most of the developmental studies had been completed that Brennan (1995) made a direct reference to the context of the Juster Scale. In his article, the irrational forecasts of an innovation made over a period of two years were attributed to the context of the scale. Following this observation, a polling study compared two methods of providing context to the Juster Scale (Hoek & Gendall 1996). This study tested a premise that recommends asking attitudinal and opinion questions before behavioural questions to set the latter's context (Labaw 1990). Hoek & Gendall's study, however, did not enquire whether the Juster Scale required such additional input to collect probability data in the desired context.

A stream of studies, separate to the ones cited so far, implemented the Juster Scale after respondents had viewed contextual information about the product (Urban, Hauser, Quall & Weinberg 1997; Urban, Weinberg & Hauser 1996; Hauser, Urban & Weinberg 1993). Contextual information was provided via a simulation of a purchasing environment that respondents were asked to experience. The simulation comprised of visiting a virtual showroom, reading articles, viewing advertisements and listening to word of mouth communication. Respondents used computer terminals to access the information sources. Absolute forecasting errors ranged from 5% to 10% in these studies. Urban *et al.* (1997) also compared the simulated environments with real life environments and found no difference between the two in terms of accuracy of performance. This thesis adapted that approach for its implementation over the Internet to investigate whether the Juster Scale required additional input to collect purchase probability data in the desired purchasing context.

1.3 Mutually Exclusive Behaviours

One application of the Juster Scale that has considerable practical use is in the forecasting of mutually exclusive behaviours. Researchers have employed this application successfully to forecast election results (Hoek & Gendall 1993; 1996), switching behaviour between competing products (Parackal & Brennan 1999) and

market shares (Brennan & Esselmont 1994). The challenging part of this application has been to get respondents to indicate probability scores that were relative to the available alternatives. Respondents, in general, failed to understand how probability operated when making a choice from a set of mutually exclusive alternatives (Flannelly, Flannelly & McLeod 2000; 1999; Parackal & Brennan 1998; Hoek & Gendall 1997a; 1996; 1993). That is, they failed to see that they would adopt only one of the alternatives, hence the chances assigned to the alternatives must total up to ten. As probability scores were assigned to alternatives without considering their relative influence, probability scores did not add up to ten. The mean probability scores calculated failed to explain the behaviour of the sample toward the alternatives.

To use the data to explain the behaviour in question, probability scores had to be weighted to ten or one across the alternatives (Hoek & Gendall 1993). This was done to the probability scores of each respondent before calculating the means. The weighting procedure was successful in fixing the problem of illogical assigning of probabilities in the sample. It is, however, not clear whether the forecasts obtained by this approach were the same as those obtained when respondents by themselves gave probability scores that added up to one.

The Constant Sum Method has also been used to forecast and explain future behaviours. It asks respondents to assign a constant number of points, percentage, or tokens across a set of alternatives (Reibstein 1978; Alexrod 1964; Metefessel 1947). In the studies cited, the Constant Sum Method produced more accurate forecasts than the methods against which it was compared. Based on this principle, a Constant Sum Scale was developed by Hamilton-Gibbs *et al.* (1992) to collect purchase probability data for fast moving products. The scale was found to produce satisfactory results in all its tests (Brennan *et al.* 1995b; Brennan & Esslemont 1994; Seymour *et al.* 1994; Hamilton-Gibbs *et al.* 1992). Subsequently, Hoek & Gendall (1997b) employed this scale to collect voting probability. The scale was successful in getting respondents to distribute voting probability scores across all the competing candidates and parties and they always added up to ten.

The two methods introduced above have shown satisfactory results in separate studies. There was, however, no comparison between the two methods in the literature. In the

current research, the two methods were compared to establish which was most suitable for collecting probability data of mutually exclusive behaviours. The comparisons were also seen as being appropriate to verify whether forecasts based on the weighting process was similar or different to those based on probability data that did not require the weighting process.

1.4 Research Objectives

To address the two issues introduced above, the following objectives were set for the research carried out in this thesis:

- To investigate whether the Juster Scale requires additional contextual information to collect purchase probability data in a purchasing context.
- To investigate whether forecasts of mutually exclusive behaviours based on probability scores that did not add up to ten (weighted probability scores) are more accurate than those based on scores that added up to ten (Constant Sum Scale).

1.5 Research Approach

The method of using probability data to make forecasts requires the collection of quantitative data. The Juster Scale was developed for this very purpose. Hence quantitative survey research methods had to be employed to address various issues of the scale. This was the research approach used in the Juster Scale literature and in this research also. Based on this research approach, the current research implemented two separate quantitative surveys; one on a random sample drawn from the client base of Vodafone New Zealand and the other on a random sample drawn from the New Zealand electoral roll. For simplicity and for all discussions in this thesis, the two surveys will be called "Vodafone survey" and "New Zealand survey" respectively.

To achieve the first objective, purchase probability data for WAP-capable mobile phones were collected using the Juster Scale. Data were collected using questionnaire versions implemented in separate treatments (Standard, Point & Click, and Search

Engine). Mean probability scores obtained in the treatments were compared for statistical differences.

To achieve the second objective, probability data for subscribing to payment plans offered by mobile telephony service providers (Telecom and Vodafone) were collected in separate treatments. In one, an electronic version of the Constant Sum Scale was implemented and in the other, an approach that did not require respondents to give scores that added up to ten was implemented (Weighted-scores). The treatments were fielded in the New Zealand survey. As the sample was selected from the electoral roll, it included respondents who did not use mobile phones. These respondents were asked to give their probability to sign up with mobile telephony service providers, providing an additional group on which the comparisons were performed. Mean probability scores of each item obtained in the treatments were compared.

1.6 Outline of Chapters

In Chapter Two, a review of literature pertaining to the Juster Scale is covered. The literature that established the ascendancy of probability surveys over intention surveys to forecast future purchase behaviour is first examined. In the latter half of this chapter, the developmental works that established the Juster Scale as the preferred scale is covered.

Chapter Three follows on to provide a critical review of the Juster Scale studies.

In Chapter Four, a review of literature pertaining to the application of the Internet in survey research is covered. The review includes the adoption of the Internet, Internet technology and different ways of using the Internet in survey research. The purpose of this review is to select a suitable survey approach to collect the required data for this thesis.

In Chapter Five, the methodology employed to achieve the research objectives is provided. In this chapter, the issues raised in the review of literature are prioritised and the objectives set out in this opening chapter are reiterated. Then the methodology adopted is explained.

Chapter Six reports the results of the survey approach adopted in this research. This chapter is included to demonstrate the success of the survey approach employed in collecting data required to achieve the research objectives.

In Chapter Seven and Chapter Eight, results of the investigation carried out pertaining to the two research objectives are reported.

In Chapter Nine, the conclusions of the investigation and the recommendations for future investigation in these areas are provided.

2. PROBABILITY SCALES USED FOR FORECASTING PURCHASE BEHAVIOUR

2.1 Introduction

The market environment under which businesses operate is subject to frequent and violent changes (Golicic, Davis, McCarthy & Mentzer 2001). Under such conditions, marketers rely on forecasts to make decisions (Waddell & Sohal 1994). Companies such as Coca Cola (Carroll 1989), Janssen Pharmaceutica (Bowditch, Fitall & Wilde 1995) and General Motors (Urban *et al.* 1996), in spite of the substantial amount of information about past performance available to them, continue to employ forecasting methods to facilitate marketing decisions.

Forecasts are used as input variables in making production, personnel, finance, and marketing decisions (Armstrong 1986). Hence, it is important to employ reliable and valid forecasting methods (Golden *et al.* 1994). A plethora of literature on forecasting methods exists from the early 1960s (Armstrong 1986). Very few studies have compared these methods to establish their relative performance. Hardie, Fader & Wisniewski (1998) compared eight mathematical models used for forecasting trial sales of consumer packaged goods. This study showed that simple models based on time series were more accurate than complex ones. In a more recent study, forecasts based on purchase probability data were found to be more accurate than those based on extrapolation of past sales (Armstrong, Morwitz, & Kumar 2000). Prior to this, Alexrod (1964) compared sixteen forecasting methods for sensitivity, stability and predictivity and found that the Constant Sum Scale, a method used to collect purchase probability data, as being superior to the rest on each of the above three dimensions. Subsequently a research stream that compared forecasts based on purchase probability data with those based on intention data, found the former to be more accurate (Day *et al.* 1991; Gan *et al.* 1986; Gruber 1970; Juster 1966).

This review of literature concentrates on the method that uses purchase probability data to forecast future behaviour. In this chapter the history of this method will be traced; it

stemmed from the dissatisfaction of intention scales to produce accurate forecasts. Then the different scales used to collect purchase probability data will be covered. Following this, the developmental work done on one probability scale called the Juster Scale (Juster 1966) is reviewed.

2.2 Intention Scales

Researchers have used different intention scales to collect information about purchase behaviour. All these have been Likert-type scales with scale-points ranging from two to nine. In most two and three-point scales, options were described using verbal descriptions such as "yes", "no", "don't know" (Heald 1970; Tobin 1959; Klein & Lansing 1955). Other scales have used semantic differentials such as "probably", "maybe", "unlikely", or "most likely" to describe the options (Pickering & Greatorex 1980).

This type of scale, however, had a number of problems that led to it being abandoned. One was that it offered respondents only a limited number of choices (depending on the scale-point) to convey their purchasing plan (Ferber & Piskie 1965). In the cases of dichotomous and trichotomous scales (yes/no and yes/no/don't know), choices were restricted to two and three options. Such limited options may not be sufficient to discriminate one's behaviour satisfactorily. Respondents were forced to choose one of the options on this scale, even if that did not best describe their behaviour.

In the case of semantic differential scales, the problem lies with the meanings of words used to describe the differentials. Meanings of verbal expressions are known to vary across individuals (Juster 1966, Ferber & Piskie 1965). When this happened to differentials on scales, it affected the responses collected using them (Worcester and Burns 1975).

Another problem of intention scales was the difficulty to quantify the behaviour in question (Day *et al.* 1991). At best, frequencies of the descriptors could be produced that could be expressed in terms of proportions. For example, in the case of a semantic differential scale, the frequency produced could be interpreted as 20% stated that they would most probably

buy, 30% stated that they would probably buy, 35% stated that they would most probably not buy, and 15% stated that they would probably not buy. Such descriptive statistics cannot provide definitive indications of how many would actually purchase the product in question.

Finally, and most importantly, accuracy of forecasts obtained on intention scales was not satisfactory (Day *et al* 1991; Gan *et al* 1986; Guber 1970; Theil & Kosobud 1968; Juster 1966; Byrnes 1964). The studies cited compared intention scales with probability scales. The results provided some compelling reasons to abandon the former in favour of the latter for forecasting purposes. Inadequacy of the intention scale to explain actual purchase behaviour was shown by experimental surveys built into the US Quarterly Survey of Intention (QSI). The US Bureau of Census conducted these surveys (QSI) in the 1960s to measure purchases of household goods by consumers. The Detroit Experiment was the first of the experimental surveys built into the 1963 (November) QSI (Bymes 1964). Intention and purchase probability data for household goods and automobiles were collected from 192 households. Cross tabulating the intention and probability scores of respondents showed that a large number who indicated no intention to purchase went on to indicate a non-zero probability score on the probability scale. This observation suggests that the intention scale had limited discrimination power when compared with the probability scale.

In the study by Juster (1966) probability data were collected from respondents who had recently participated in the 1964 (July) QSI. Juster employed an 11-point probability scale, whereas, the QSI employed a 5-point intention scale. The two surveys collected intention and probability data for the same product, providing comparisons between the two scales. These comparisons brought to light a number of limitations of the intention scale. Response distributions produced by the two scales were markedly different. A good number of respondents who indicated "no intention to buy" on the intention scale, indicated non-zero probability scores on the probability scale. The reverse pattern was also observed with respondents who indicated “definite or probable intentions to buy” on the intention scale indicating a zero probability score on the probability scale. Respondents who indicated “no intention to buy” using the intention scale, eventually ended up making most of the actual

purchases. Respondents who were unsure of their behaviour and indicated, "don't know" on the intention scale expressed their purchase behaviour with non-zero probability scores on the probability scale. Actual purchases made by "intenders" and "non-intenders" within probability groups varied systematically. In contrast, actual purchases made by those who indicated "zero-probability" and "non-zero probability" varied randomly within intention groups.

Forecasts of automobile purchases made on the probability scale were more accurate than those made on the intention scale. Regressing actual purchases of automobiles with intentions scores showed that intention data explained most of the variations in the actual purchase behaviour. However, when probability scores were included in the regression analysis, intention scores failed to show any association with actual purchase behaviour. On the contrary, probability scores exhibited significant association with actual purchase by itself and along with intention scores.

The poor performance of intention scales was observed in a study that validated the intention data for automobiles collected in another QSI (Theil & Kosobud 1968). The results showed that those who indicated "no intention to buy" made 70% of the actual purchases. Those who indicated "some intention to buy" or "intention to buy" made less than 40% of the purchases. Thus the accuracy of forecasts was not satisfactory for both those who purchased and those who did not purchase the product.

2.3 Probability Scales

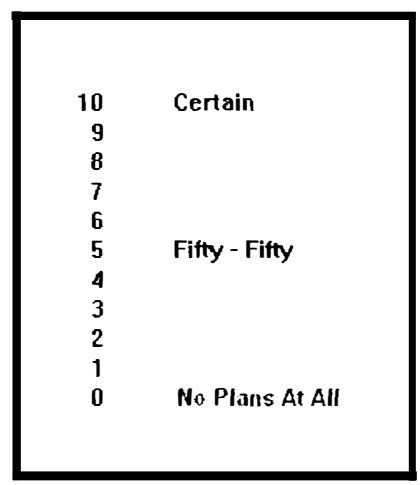
Observations made in the above studies provided convincing evidence for investigators to give up intention scales in favour of probability scales. Subsequent research in this area concentrated on developing probability scales (Day *et al.* 1991; Gan *et al.* 1986; Pickering & Isherwood 1974; Gabor & Granger 1972; Gruber 1970; Clancy & Garsen 1970; Stapel 1968; Ferber & Piskie 1965). In this section, the developmental work done on probability scales is examined. To start, the different probability scales used in the literature are explained. The review then centres on the Juster Scale that attracted much of the

investigation in this area.

2.3.1 Plan-O-Meter

The first reported attempt to use probability scales in forecasting dates back to 1958 (Ferber & Piskie 1965). An eleven-point scale was used to collect buying intention data for 14 expenditure categories (nine durable goods; home repairs, improvements, redecorating, vacation and education). This scale was called Plan-O-Meter (see Figure 2.1) by its authors and had probability statements against the two terminals (0 - No Plans At All; 10 - Certain) and against the mid point (5 - Fifty-Fifty).

Figure 2.1 Plan-O-Meter (cited in Juster 1966 p 12)



The following question was used along with the Plan-O-Meter:

> *"Do you plan to purchase any of these goods (whether owned presently or not) between now and _____. Let's take the first one_____, how likely are you to purchase it during this period?"(p 322).*

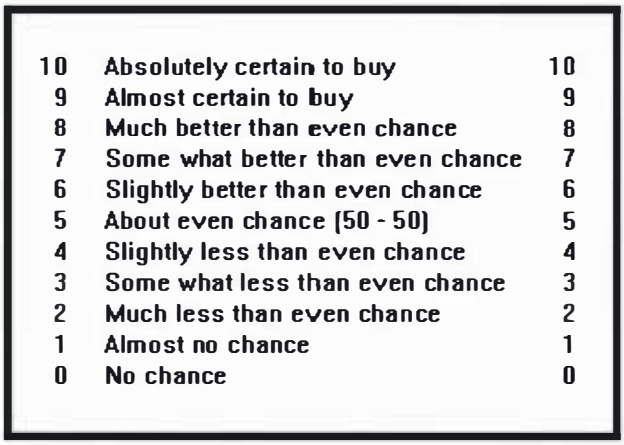
Ferber & Piskie (1965) observed that their scale was successful in getting respondents to convey their buying plan. Responses, however, were concentrated at the three labelled scale points (at 0, 5 and 10), forming a tri-modal distribution. Obviously, the tri-modal distribution was due to the labels against the points, making them stand out more than the other points.

Regression analysis was used to assess the amount of variation in the actual purchase behaviour explained by responses collected on the scale. Independent variables (selected based on chi-square tests) used in the regression were family size, age, education, income, income change, home ownership and buying plans measured on the Plan-O-Meter. Four variations (three transformations that reduced the raw scores into dichotomous scores; the fourth was retained as the raw scores) of scores collected on the Plan-O-Meter were used in separate analyses. In all the analyses, scores collected on the Plan-O-Meter (reduced and raw scores) accounted for most of the variation in the actual purchase behaviour. The R square obtained remained very much the same for all four variations. This led the authors to suggest that the Plan-O-Meter in its full form (eleven-scale point) did not produce any additional information over its dichotomous forms. This, however, may not be a correct reasoning as the four variations originated from the same data. To make such a conclusion, a systematic comparison of the eleven-point scale with its dichotomous variations must be made.

2.3.2 Byrnes' Scale

In the Detroit Experiment (1963, November QSI), comparisons were made between a five-point intention scale and an eleven-point probability scale (Byrnes 1964). Probability and intention data for eight durable products (new and used automobiles, kitchen range, refrigerator, washing machine, clothes dryer, room air conditioner, television set and dishwasher) were collected using these scales. Cross tabulation of probability and intention scores showed that the probability scale was successful in getting respondents to express their purchase behaviour, which the intention scale had failed to do. The probability scale used was different from the one used by Ferber & Piskie (1965) in the sense that it had probability statements against all eleven scale points (Figure 2.2). The mid point (5) had an additional label, "50-50" included.

Figure 2.2 Byrnes' Scale (cited in Juster 1966, p 13)



10	Absolutely certain to buy	10
9	Almost certain to buy	9
8	Much better than even chance	8
7	Some what better than even chance	7
6	Slightly better than even chance	6
5	About even chance (50 - 50)	5
4	Slightly less than even chance	4
3	Some what less than even chance	3
2	Much less than even chance	2
1	Almost no chance	1
0	No chance	0

The following question accompanied the above scale:

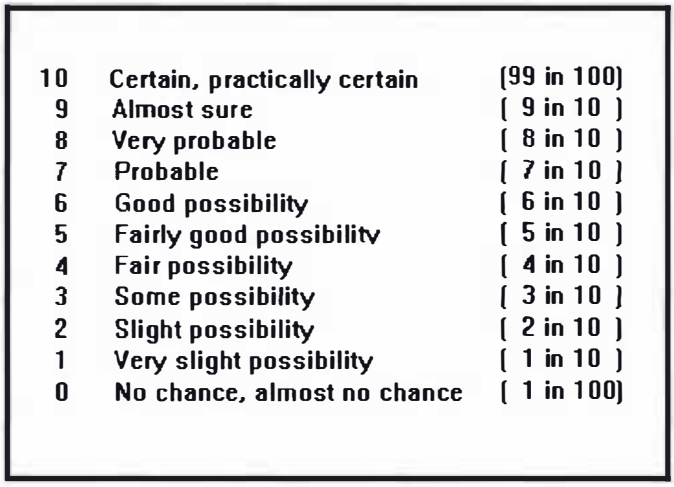
> *"During the next (six, twelve and twenty four) months, that is between now and the next_____, what do you think the chances are that you or someone in the household will buy a _____?"*

Response distributions obtained on the above scale emulated the "inverse J" shape curve and had a distinct peak at the mid point (5). This perhaps could be because of the over emphasis of the scale's mid point (Juster 1966). The scores were more evenly distributed across the scale, which was an improvement over the scale used by Ferber & Piskie (1964).

2.3.3 The Juster Scale

Juster (1966) conducted a full-scale experiment to investigate differences between probability and intention scales. He surveyed 800 households who participated in the July 1964 QSI, a few days after the US Census Bureau conducted its survey. Respondents were asked to indicate their purchase probabilities on a modified version of the Byrnes' Scale (see Figure 2.3) for the same products that the QSI collected intention data.

Figure 2.3 The Juster Scale (Juster 1966, p 15)



10	Certain, practically certain	(99 in 100)
9	Almost sure	(9 in 10)
8	Very probable	(8 in 10)
7	Probable	(7 in 10)
6	Good possibility	(6 in 10)
5	Fairly good possibility	(5 in 10)
4	Fair possibility	(4 in 10)
3	Some possibility	(3 in 10)
2	Slight possibility	(2 in 10)
1	Very slight possibility	(1 in 10)
0	No chance, almost no chance	(1 in 100)

The above scale became known as the Juster Scale in the academic literature, after its author. The Juster Scale was accompanied by the following question:

> *"Taking everything into account, what are the prospects that some member of your family will buy a _____ sometime during the next_____ months: between now and next_____?"*

The response distributions obtained for all the products on the Juster Scale exhibited a rather smooth "inverse J" shaped curve. Respondents were able to give probability scores using all the scale points. This was suggestive of the Juster Scale's better discrimination power over the previous two scales (Byrnes 1964; Ferber & Piskie 1965). The difference between the Juster Scale and the previous scales (Byrnes 1964; Ferber & Piskie 1965) was that it did not overemphasise any of the scale points.

The six months recall data of automobiles (recall data was obtained for automobiles only) showed that 11% of non-intenders (measured on the Intention Scale) actually made a purchase. The corresponding figures for those who gave zero probability scores were comparatively low (8%, 6% and 5%) for the three forecasting periods (six, twelve and twenty-four months). Of all the actual purchases made by non-intenders, 25% (8) were made by non-intenders who indicated zero probability scores for all three forecasting periods. The remaining 75% (24) of purchases were made by non-intenders who indicated

non-zero probability scores. Overall, the accuracy of forecast obtained on the Juster Scale was better than that obtained on the intention scale. Other researchers who tested the Juster Scale also reported similar results regarding its accuracy (Gan *et al.* 1986; Gruber 1970; Gabor & Granger 1972). In the years that followed, the Juster Scale received more research attention and became the preferred probability scale (Armstrong 1986).

2.3.4 Likelihood Scale

One study in the Netherlands, during the same period when Juster was testing his scale, used a scale to collect percentage chances for buying a new or old car (Stapel 1968). The scale used was an eleven-point one that had percentages (0 to 100%) against statements that expressed the likelihood of buying (see Figure 2.4).

Figure 2.4 Likelihood Scale (Stapel 1968, p 100)

100%	**(One hundred percent sure)**
90%	**(Almost certain)**
80%	**(Very big chance)**
70%	**(Big chance)**
60%	**(Not so big a chance)**
50%	**(About even)**
40%	**(Smaller chance)**
30%	**(Small chance)**
20%	**(Very small chance)**
10%	**(Almost certainly not)**
0	**(Certainly not)**

The scale was accompanied by the following question:

> *"Will you indicate how big you yourself estimate the chance of someone in this family buying a car, either new or used, this year? What percent chance?"*

This scale also produced the "inverse J" shape distribution that the previous scales produced (Byrnes 1964; Ferber and Piskie 1965; Juster 1966). Actual purchase rate

increased as probability scores increased. This observation was in line with those made for the other scales (Ferber and Piskie 1965; Juster 1966). The response distribution had a distinct peak at the mid point (5) that was also seen on the distributions produced by Byrnes' and Ferber and Piskie's scales.

2.3.5 Pickering & Isherwood's Scale

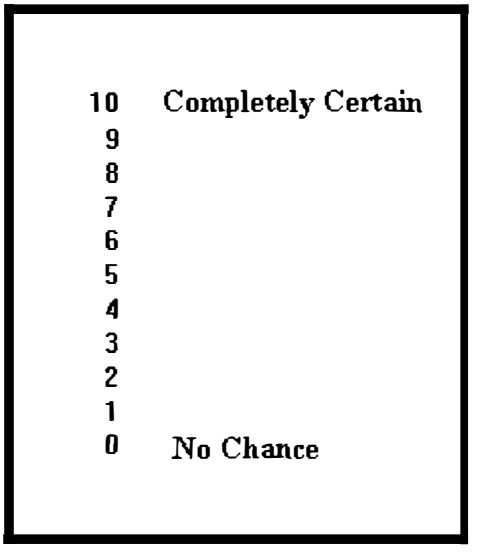
Pickering & Isherwood (1974) used an eleven-point scale (Figure 2.5) that had probability statements attached at the terminals (0 and 10). Their show card included the following explanation:

> *"A score of 8 would mean that you were 80% certain, a score of 1 that you were 10% certain and so on."*

Respondents were asked the following question:

> *"What is the probability of buying a [item specified] in 3 month's time, that is [month specified]?"*

Figure 2.5 Pickering & Isherwood's Scale (Pickering & Isherwood 1974, p 208)



Probability data for 18 durable goods over three time periods (three, six, and twelve months) were obtained using the above scale. Response distributions produced the "inverse J" shaped curve. Responses aggregated at the mid point to exhibit a distinct blip. This was

the case for all the test products and across the three time periods (three, six and twelve months). Probability scores and actual purchase exhibited a positive correlation, confirming the observation made by the earlier investigators (Stapel 1968, Juster 1966, Ferber & Piskie 1965). Of the respondents who gave zero probability scores, a small proportion (5%) accounted for 55% of the purchases. Sixty percent of the respondents, who indicated "Completely certain", made purchases. The aggregate predictive error (across all test products) based on a 14 months validation was only four percent.

2.4 Development of the Juster Scale

The favourable results obtained in the initial study (Juster 1966) led others to test the Juster Scale (Gan *et al.* 1986; Gabor & Granger 1972; Gruber 1970). In this section, the developmental works done on the Juster Scale are reviewed. To begin, studies that replicated Juster's experiment are covered. Following this, studies that compared the Juster Scale with other probability scales are reviewed. In the final section, studies done to customise the Juster Scale for various survey types are examined.

2.4.1 Replications of Juster's Experiment

Gruber (1970) replicated Juster's experiment on 16 new food product concepts. Probability and intention data were collected from 200 female shoppers. They were shown pictures of the product concepts, followed by a questionnaire with the Juster Scale, a five-point Intention Scale and other questions. This study did not validate the scales but provided cross-tabulation of responses obtained using the two scales.

The two scales exhibited high positive correlation with one another. The scores collected on the two scales were concentrated towards the upper end of the scale. The response distributions emulated the "J-shape" curves, and were different to the previous shapes observed so far. Being innovative products, most respondents appeared to express interest in purchasing them; both scales were able to show this.

A large majority of respondents who stated "might or might not buy" (mid point) and "probably would not buy" on the intention scale, gave non-zero probability scores on the Juster Scale. This observation was in alignment with that made by Juster (1966). Respondents seem to be able to discriminate their purchase behaviour better using the Juster Scale than the Intention Scale. According to Gruber, the better spread of data obtained on the Juster Scale should translate into better forecasts in terms of accuracy. Guber (1970), however, did not verify this claim.

The Juster Scale was tested in a study in Britain (Nottingham) to forecast the purchase behaviour of eight different durable items (Gabor & Granger 1972). Respondents were asked to indicate their probability of purchasing the products within the next twelve months. This study carried out a 12-month validation for the forecasts made. Results of the validation survey showed that respondents who gave zero-probability scores made most of the purchases (65%). A large majority (80%) of these were replacement purchases; apparently respondents did not anticipate them when they indicated their purchase probability. In the non-zero probability category, actual purchases were high for those who gave high probability scores (0.7 and above). This observation was consistent with that made in the earlier studies (Pickering & Isherwood 1974; Stapel 1968; Juster 1966; Ferber & Piskie 1965). Because of the strong correlation between probability scores and actual purchase behaviour, the overall picture constructed by averaging the probability scores provided a reasonably accurate estimate of the actual behaviour.

Clawson (1971) examined the accuracy of the Juster Scale to forecast purchase behaviour of nine products and services (attend movies; travel outside South California; ride local bus; trip in a camper, motor home, or travel trailer; buy common stock, preferred stock or mutual fund; move to a different house; open a saving account; buy or lease automobile; buy TV set) over short-term periods (3 months). He carried out an initial survey of 299 households in Los Angeles and Orange Counties in the USA in June 1969. A second survey in September 1969 was carried out to find out the actual purchases made by respondents during the test period. The second survey collected actual purchases from a total of 327 respondents (176 re-interviews of respondents of the first survey and 151 new interviews).

This design enabled Clawson to compare forecasts with actual purchase behaviours in two contrasting samples (related and independent samples). A total of 17 separate comparisons to validate forecasts were made. In all the comparisons, forecasts were overestimated. In three comparisons, differences between forecasts and actual purchase behaviours were significant. Two of them were to "buy common stock", in the related and independent samples ($p = 0.01$) and one to "travel outside California", in the related samples ($p = 0.05$). Differences between forecast and actual purchase behaviour was not significance ($p > 0.05$) in the remaining 14 comparisons. Regression analysis was executed to find out how much of the variations in the actual purchase could be explained by the purchase probability data. Actual purchases of all items were regressed with the mean purchase probability scores of all the test items. R squares produced for the independent and related samples were 0.98 and 0.96 respectively. The F-values were significant at the 0.001 levels in both analyses. The high R squares were indicative of the probability data's ability to explain actual behaviour satisfactorily.

Clawson (1971) pointed out that the accuracy of forecasts was a functional relation between confidence respondents placed in their purchase plans (measured on the Juster Scale) and the extent to which they were able to fulfil them. This relationship was best seen in the related samples that allowed comparisons of purchase probability scores and actual purchases of the same respondents. To show this Clawson aggregated probability scores into five categories (0; 1 to 3; 4 to 6; 7 to 9; and 10). Regression analysis was executed with actual purchases of all products as the dependent variable and probability categories of all products as the independent variable. The R square produced was 0.99 and was significant at the 0.01 levels. This result suggests that the probability categories accounted for almost all of the variations in the actual purchase behaviour.

Regression analyses were executed on each test product separately. R-square was highest for "local bus rides" and "attend movies" (0.98 and 0.96, significant at the 0.001 and 0.01 levels) and lowest for "buy a TV set" and "open a savings account" (0.0018 and 0.22; not significant at the accepted level). Clawson was not able to give a satisfactory explanation for the vast variation in R squares between the individual items. He observed that 104

respondents gave high probability scores (0.7, 0.8, 0.9, and 1.0) to "attend movies", in contrast, only eight gave high probability scores to "buy a TV set". The author attributed the number of high probability scores to the better R-square result.

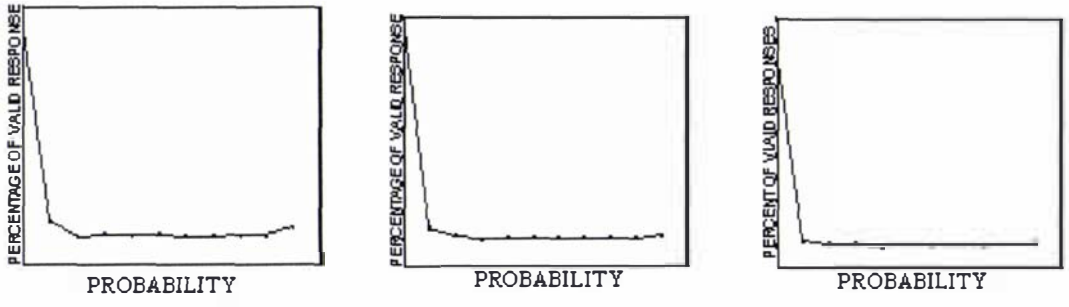
Perhaps respondents were confident of their plans to "attend movies" and hence were able to express their behaviour accurately on the Juster Scale. In the case of "buy a TV set" respondents may not have been so confident of their plans and hence were not accurate in expressing their behaviour on the scale. Examining the response distribution obtained for the two items raised the suspicion of the sample being biased towards certain items. Discussion on this matter is further extended in Chapter Three.

In New Zealand, Gan *et al.* (1986) confirmed the previous results in a study that compared the Juster Scale with a five-point intention scale. Probability and intention data for three product categories (durables, services, and fast moving consumer goods) were collected. Respondents were first asked to express their purchase intention using the intention scale, following this, they were asked to express their purchase probability using the Juster Scale. Cross tabulating the responses collected using the two scales showed that 25% of those who indicated 'definitely will not buy' on the intention scale indicated non-zero probability score on the probability scale. About 90% of those who indicated 'don't know' on the intention scale indicated non-zero probability scores on the probability scale. Correlations of probability scores and actual purchase were much higher than correlations of intention scores and actual purchases. Respondents who indicated no intention to buy on the intention scale but indicated a non-zero probability on the Juster Scale made most of the purchases. Respondents in general were able to express their purchase behaviour better using the Juster Scale. Validation of forecasts showed that the Juster Scale was more accurate than the intention scale.

Gan *et al.*'s study revealed further shapes to the response distributions obtained using the Juster Scale. In the case of durable goods there was a high proportion of zero probability. The response distributions (see Figure 2.6) of all products in this category (car, video, compact disc player) produced the "inverse J" shape curve. This was in agreement with the

earlier studies (Byrnes 1964; Stapel 1965; Juster 1966).

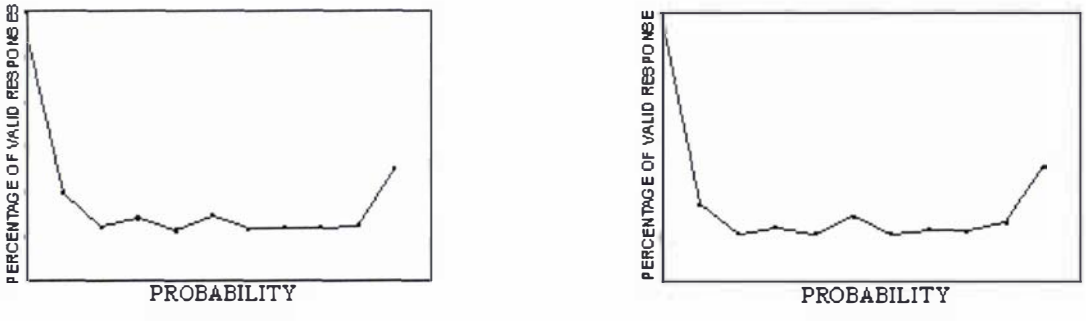
Figure 2.6 Probability Distribution of Durable Goods (Gan *et al.* 1986)



Car **Video** **Compact Disc**

In the case of services (shares, hotel accommodation) that were less expensive than durables, responses were unevenly distributed across the scale (see Figure 2.7). The proportion of zero probability was much less when compared with the durables. The response distributions exhibited a trough between 0 and 10 and a distinct peak at the mid point (5) for all products in this category.

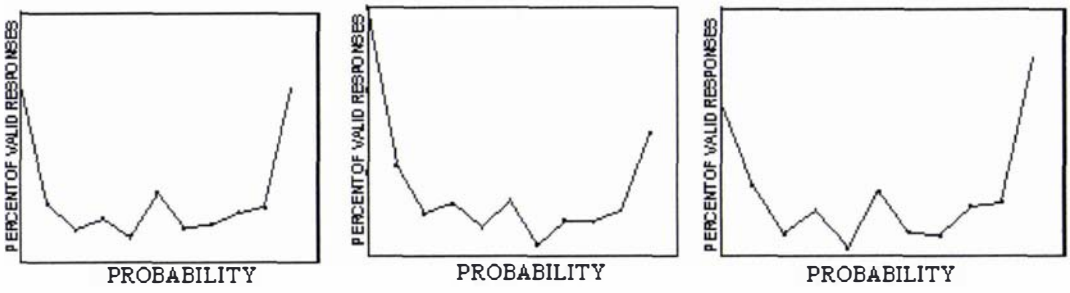
Figure 2.7 Probability Distribution of Service Goods (Gan *et al.* 1986)



Shares **Hotel**

Response distributions of fast moving consumer products (LP record, movie, book) were even more unevenly distributed across the scale. Unlike durables and services, the distribution of this product category exhibited a "U" shape curve. The distribution peaked at two distinct points (3 and 5) within the trough (Figure 2.8).

Figure 2.8. Probability Distribution of Fast Moving Products (Gan *et al.* 1986)



LP Record **Movie** **Book**

Shapes of response distributions (Figures 2.6, 2.7 & 2.8) shown above were typical of the product categories. In the case of expensive and seldom-purchased products, proportions of zero probability scores were comparatively high. The response distributions took the shape of the "inverse J" curve. This was only to be expected, as few people can afford to buy such products at any given time. In the case of products not so expensive (services and fast moving products), proportions of zero probability scores were comparatively less than those of the durables. Probability scores collected for the latter two categories were distributed across the scale.

As far as the smoothness of the distribution was concerned, durables clearly exhibited a smooth "inverse J" shape curve. Distributions of services and fast moving products were uneven, with distinct blips within the trough (a single blip at scale point 5 for services; two blips at 3 and 5 scale points for fast moving products). From the above observations, the blips on the response distributions appear to be specific to the products tested and may not be because of the scale attachments as argued by Juster (1966). Juster's (1966) study was based on the argument that the earlier scales (Ferber & Piskie 1965; Byrnes 1964) were causing blips and regarded as being inferior. While there may be some truth in Juster's argument, it was not based on any empirical evidence. Hence, research evidence ought to be sought to find out in what sense the Juster Scale was an improvement over other scales in the literature (Pickering & Isherwood 1975; Ferber & Piskie 1965; Byrnes 1964). This raises the need for a systematic comparison of probability scales to establish the one that could be recommended to collect purchase probability data. This is further discussed in Chapter Three.

According to the literature on scale reliability, respondents tend to choose the mid point of a scale to give a safe answer (Peter 1979; Garner 1960). This could result from ambiguity of the response options provided on the scale. In which case, the scale fails to discriminate respondents satisfactorily. It could also arise when there is ambiguity about the behaviour in question. The possibility of the latter being true in the case of the Juster Scale appears to be evident from the observations made in Gan *et al.* (1986).

In the case of fast moving products, two distinct blips were noticed within the trough. This again could be specific to the product category. For this product category, the decision is not whether to buy or not to buy, but rather when to purchase (whether to purchase this month or next month). Although there was no clear evidence found in the literature, it could be the uncertainty of when to purchase that causes the uneven distribution in this category. As products in this category are purchased frequently, what would be useful to know is how much would be purchased or the purchase level. Hamilton-Gibbs *et al.* (1992) developed a method to forecast the number of units purchased or the purchase level. The method will be discussed in this chapter in section 2.4.2 (page 31) while dealing with different versions of the Juster Scale.

The shapes of response distributions observed so far were all typical Beta distributions ("U" shaped, "inverse J" shaped, "J" shaped and skewed distributions). This suggests that the Juster Scale and the other probability scales were successful in collecting probability data that essentially was a choice between two alternatives (to purchase or not) (Ehrenberg & Uncles 1998; Ehrenberg 1988; Riebe, Danenberg, Sharp & Rungie 1999). Hence, it may not be right to discount a scale because of the shape of its response distribution or blips on its response distribution. The observations regarding the shapes of response distribution and blips in Gan *et al.* (1986) raised the requirement for a systematic comparison of probability scales to establish the one that produces the best results.

2.4.2 Comparisons of the Juster Scale with Other Probability Scales

Literature corroborates probability scales as being a better alternative to intention scales to forecast purchase behaviour (Day *et al.* 1991; Gan *et al.* 1986; Pickering & Isherwood 1974; Gabor & Granger 1972; Stapel 1968; Juster 1966). In all the studies reviewed, respondents were able to discriminate their behaviour better on probability scales, a state that intention scales have clearly failed to achieve (Gan *et al.* 1986; Guber 1970; Juster 1966; Byrnes 1964). The Juster Scale has become the preferred probability scale and has attracted much of the investigation in this area. Empirical evidence to support this scale over other probability scales however was limited. In the following section studies that compare the Juster Scale with other probability scale are discussed.

The Juster Scale versus the Numerical Probability Scale

The academic literature recognises the ambiguity of verbal expressions in survey questions (Belson 1986; Gendall & Hoek 1990) and measurement scales (Worcester & Burns 1975). Worcester & Burns (1975), in particular, showed that meanings assigned to verbal descriptors varied considerably across respondents, influencing their responses given on the scale. For this reason, Juster (1966) suggested that a scale with reduced verbal descriptions could be more accurate than his own scale. All the same, he never endeavoured to investigate the matter. Others have used probability scales with reduced verbal descriptors with reasonable satisfaction (Pickering and Isherwood 1974; Stapel 1968; Ferber & Piskie 1965). The latter investigators mentioned also did not endeavour to find out whether their scales produced better forecasts. To resolve this dilemma, Gendall *et al.* (1991) compared the Juster Scale with a version that did not include the probability statements (Numerical Probability Scale) for forecasting accuracy and overall performance. The two versions were implemented in a split sample study using self-completion questionnaires (method discussed under Survey types on page 32).

For seven of the ten products tested, forecasts made on the Juster Scale were more accurate than those made on the Numerical Probability Scale. The proportion of non-zero

probability scores was slightly greater for the Juster Scale. Whether this has anything to do with the better accuracy of the Juster Scale needs investigation. Both scales produced the "inverse J" shape response distribution curve. The only difference observed was the blip at the mid point. The Juster Scale produced a smooth curve with no obvious blip, whereas the Numerical Probability Scale produced a distinct blip at the mid point. Apart from the numerical statements, no additional emphasis was given to any of the scale points on the Numerical Probability Scale, hence, the argument that over emphases of attachments cause blips may be ruled out. The products tested were kept constant for both scales, hence, the argument that products cause blips may also be ruled out. The only difference between the scales was the probability statements. Therefore, the blip observed on the Numerical Probability Scale could be because of the scale not having those statements. If respondents were choosing the mid point to give a safe answer (Peter 1979; Garner 1960) on this scale then it has obviously failed to discriminate all behaviours satisfactorily. Some evidence on this effect was seen with a small number of respondents (n = 7) who made an additional point at the lower end of this scale to convey absolutely no chance of buying (zero probability). There was no such problem (respondents adding points) with the Juster Scale. The smooth curve obtained on the Juster Scale could be taken as the scale's better discriminatory power over the Numerical Probability Scale.

The Juster Scale versus the Verbal Probability Scale

Brennan *et al.* (1995a) tested the Juster Scale in a telephone survey. They compared the Juster Scale with a scale that required respondents to give a number between zero and ten verbally over the telephone (Verbal Probability Scale) (the method adopted is discussed under Telephone Survey in section 2.4.3, page 32). The Juster Scale was mailed out to respondents with instructions to use it when an interviewer contacted them over the telephone. Interviewers subsequently called respondents over the telephone and asked them to indicate their probability scores using the Juster Scale. In the case of the Verbal Probability Scale, interviewers called respondents directly by telephone and asked them to indicate a number between zero and ten that best represented their chances out of ten to purchase the product in question. This study found that both scales produced similar levels

of forecasting accuracy. The authors recommended the Verbal Probability Scale to collect probability data over the telephone, as it worked out to be more cost efficient.

The Verbal Probability Scale was subsequently tested for question wordings in a split sample test (Dawes 2000). The following two probability questions with the words "change" and "renewed" being interchanged were implemented on separate groups:

> *"What are the chances that on your next renewal for your building insurance, you will **change** from your existing provider?"* (p 8)
>
> *"What are the chances that on your next renewal for your building insurance, you will **renew** with your existing provider?"* (p 8)

All instructions accompanying the questions were kept constant. This study reported that there was no difference in the probability scores obtained with the two questions.

The Juster Scale versus the Constant Sum Scale

For fast moving consumer products such as toothpaste and margarine, consumers are bound to purchase more than one unit over a period. In such cases, forecasting the proportion of the sample that would purchase (purchase rate) has very little value. What is worth knowing is the number of units that respondents would buy during a given period (purchase level) (Day *et al.* 1991). Hamilton – Gibbs *et al.* (1992) tested two techniques to collect probability data to forecast purchase levels of fast moving consumer products. The first technique called the "Multiple Question Method" involved interviewers initially asking respondents to indicate a probability score on the Juster Scale to purchase one unit of the product. Interviewers continued to ask respondents for probability scores to purchase two units, three units, and so on until a unit number was reached for which respondents a gave zero probability. Purchase level was calculated by summing the product of unit numbers and the corresponding probability scores given (Equation 1).

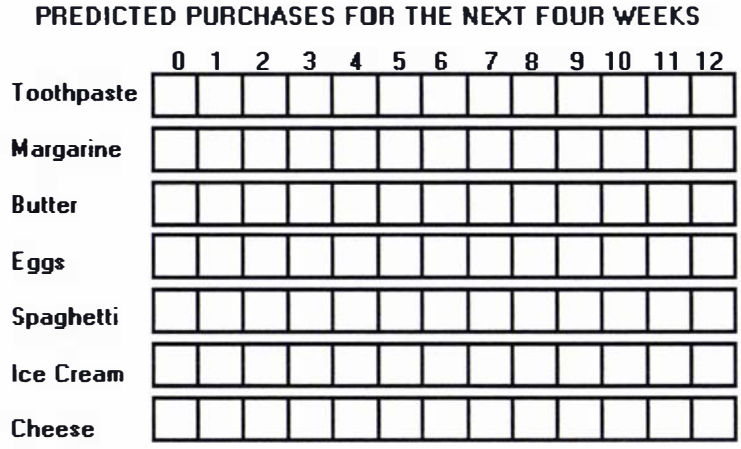
$$\sum(n_i * p_i)$$

----------Equation 1

Where
n_i = number of units
p_i = probability score

The second technique used the Constant Sum Scale (see Figure 2.9) (Reibstein 1978; Alexrod 1964; Metefessel 1947) that required respondents to distribute 10 tokens across the different units of the product. Test products were listed in rows on a grid. Each row had 13 boxes, numbered from zero to twelve and they represented the number of units purchased. Each token was equivalent to 0.1 probabilities or one in ten chances to purchase.

Figure 2.9 Constant Sum Scale (Hamilton-Gibbs, Esslemont, & McGuinness 1992, p 20)



Respondents were asked to distribute the ten token across the 13 boxes to convey their chances or probabilities of purchasing different units of the product. The following example was included to convey the task required of respondents:

> *"... if respondents felt there was a 50-50 chance the household would purchase either two or three tubes of toothpaste over the four week period, they would assign five tokens to each of the squares representing two and three tubes of toothpaste. All 10 counters, and only 10, had to be used"* (p 19).

The number of tokens assigned to a particular unit number was treated as the probability to purchase that many units. Unit numbers that did not have any tokens assigned were given a zero probability. Purchase level was calculated by summing the product of the unit numbers and the number of tokens places against the units (see Equation 1). This method forced respondents to give scores that added to one, and hence the scale got its name Constant Sum Scale.

For all seven products tested, the Multiple Question Method underestimated the purchase levels, whereas the Constant Sum Scale overestimated the purchase levels. In the case of five of the seven products, forecasts made on the Constant Sum Scale were more accurate. The authors also noted that the Constant Sum Scale was easier to implement.

Brennan *et al.* (1995b) compared the Multiple Question Method with the Constant Sum Scale to forecast purchase level of branded products (Coca Cola and Campbell's Red & White Label Soup). In this study, the Multiple Question Method underestimated, whereas the Constant Sum Scale overestimated the actual purchase level of Coca Cola (both within 5% of the actual purchase level). This result was consistent with the previous study (Hamilton-Gibbs *et al.* 1992). However, in the case of Campbell's Soup both scales abnormally overestimated the actual purchases (+ 158% for the Multiple Question Method and +102% for the Constant Sum Scale). The overestimation made for this brand was attributed to respondents reporting purchases of all type of soups, and ignoring that the question was about a specific brand (Campbell's Soups). Similar behaviour was observed in a subsequent omnibus survey on the same population (cited in Brennan *et al.* 1995b). It appears that respondents had incorrectly reported their purchases of Campbell's Soup. The authors argued that Campbell's Soup was a small brand in New Zealand, hence, was not popular enough in the market to be purchased frequently. It was also possible that respondents had averaged their envisaged purchases over time, resulting in a recall error called "averaging" (Cook 1987). This occurs when respondents are asked to recall purchase of products not frequently purchased.

Results of the Campbell's Soup research was not clear enough to make any strong conclusions. All the same it is worth mentioning that the forecasting error was less on the Constant Sum Scale. The result obtained for Cocoa Cola, however, was encouraging enough to recommend the Constant Sum Scale to forecast purchase levels of similar branded items. Based on these observations the Constant Sum Scale was recommended over the Multiple Question Method to forecast purchase levels of fast moving branded products.

Hamilton-Gibbs *et al.* (1992) found the Constant Sum Scale to be easy to implement and the task required of respondents simpler compared to the Multiple Question Method. Brennan *et al.* (1995b), however, did not share Hamilton-Gibbs *et al.*'s view, hence recommended further testing for the Constant Sum Scale. Seymour *et al.* (1994) replicated Hamilton-Gibbs *et al.*'s (1992) and Brennan *et al.*'s (1995) experiments. In that study, comparisons were made between three methods (Multiple Question Method, Constant Sum Scale that required respondents to stack the tokens and Constant Sum Scale that required respondents to lay the tokens flat) for accuracy, ease of implementation and respondents' ability to understand. The Constant Sum Scale that required respondents to stack the tokens was employed by Hamilton-Gibbs *et al.* (1992) and is called Constant Sum Scale (stack) in the following discussion for simplicity. The Constant Sum Scale that required respondents to lay the token flat was employed by Brennan *et al.* (1995b) and is called Constant Sum Scale (flat) in the following discussion for simplicity. To assess the accuracy of the methods, forecasts were compared with actual purchases collected by a recall survey implemented 28 days after the main survey. Respondents' ability to understand the methods was assessed using a seven-point scale (where "1" implied an excellent understanding level and "7" a poor one). The interviewers made the assessment at the end of each interview using the above scale. The means obtained on this seven-point scale were compared between three age categories (under 30, 31-60, over 60), two education levels (high and low) and across the three methods using Analysis of Variance (ANOVA). Purchase probability data were collected for six frequently purchased grocery items (toothpaste, butter, margarine, eggs, spaghetti, and ice-cream).

Seymour *et al.* (1994) observed that the interviews lasted fifteen minutes for the Multiple Question Method whereas it lasted about twelve minutes for the Constant Sum Scale (stacked and flat). The level of understanding differed significantly ($p < 0.01$) between the age categories across the three methods. Respondents with high levels of education were seen to understand all three methods better than those with low levels of education; this was particularly true for the under 30-age group.

Respondents had a number of difficulties in understanding the methods. Seymour *et al.*

(1994) suggest simple procedures to reduce them in future studies, these are explained along with the difficulties encountered here. In the Multiple Question Method, there was confusion as to whether respondents were asked to give the chances to buy "exactly" or "at least" 'n' items. Stressing the word "exactly" could reduce this difficulty. Many respondents who used the Multiple Question Method had difficulty understanding the idea of using probabilities to express their purchasing behaviour. Consequently, many gave probabilities that were irrational and had to be weighted to one before applying the analyses.

Elderly respondents had difficulty understanding how the counters were to be used in the Constant Sum Methods (stacked and flat). For such respondents, additional examples could be provided to help them understand the scale and the task required. Some respondents were confused as to what the counters represented, that is, whether they represented the unit numbers or probabilities. Providing additional explanation about the grid could solve this problem.

Forecasting errors varied considerably across the three methods. For the Multiple Question Method, forecasting error ranged from –6% to + 56%. Accuracy of forecast for this method was best for one product (butter underestimated by 6%). The Constant Sum Scale (flat) was observed as being the most accurate with the forecasting errors ranging from +1% (butter) to -21% (spaghetti). Accuracy of forecast for this scale was best for three products. The Constant Sum Scale (stack) was seen to overestimate actual purchases for all but one item. Forecasting error for this scale ranged between 2% and 19%. The one item (ice cream) that was an exception had a forecasting error of -24%. Accuracy of forecast for this latter scale was best for two products. The overall forecasting accuracy was best for the Constant Sum Scale (flat) advocated by Brennan *et al.* (1995b) (absolute errors ranged from 1% to 21%).

2.4.3 Implementations of the Juster Scale in Other Survey Media

The Juster Scale was originally designed for implementation in face-to-face settings (Day *et al.* 1991; Gan *et al.* 1986; Gabor & Granger 1972; Clawson 1971; Juster 1966). The scale was printed on a show card that interviewers presented to respondents with the relevant question. Gendall *et al.* (1991) implemented the Juster Scale and the Numerical Probability Scale in a self-completion questionnaire. Following this, the scale was tested for implementation in telephone surveys (Brennan *et al.* 1995a) and Internet-based surveys (Parackal & Brennan 1999). In the following section, these studies are examined in detail.

Self-Completion Questionnaire

In face-to-face surveys, interviewers can provide additional explanation whenever respondents fail to understand a scale or a question. This, however, is not possible in surveys that use self-completion questionnaires. In such surveys, appropriate explanations of the task required must be provided. Gendall *et al.* (1991) tested two probability scales (Juster Scale and Numerical Probability Scale; see Section 2.4.2, page 25) by implementing them in self-completion questionnaires. To explain how the scales were to be used the following explanation was included:

> *"If you are certain, or practically certain that you would buy < > you should choose the answer '10'. If you think there is no chance or almost no chance of buying < > you should choose '0'. If you are uncertain about the chances, choose a number as close to '0' or '10' as you think it should be."*

The objective of this study was to compare two probability scales (The Juster Scale and the Numerical Probability Scale) for accuracy. The study did not attempt to ascertain whether the explanation was sufficient. On face value, the explanation seems to be straightforward. Subsequent studies that implemented the scale in self-completion questionnaires have adopted the above explanation. While there has been no criticism of the above explanation in the literature, an empirical examination could contribute to the academic literature.

Telephone Survey

The fast spreading telecommunication network has made telephone surveys very popular among market researchers (Aaker & Day 1990). With a majority of households in the western world being accessible by telephone, it was advantageous to customise the Juster Scale to be used in telephone surveys. Brennan *et al.* (1995) tested two approaches ("Mail Group" and "No Mail Group") to collect probability data over the telephone.

Mail Group: Respondents in this group received the Juster Scale by post with instructions to use it when an interviewer called over the telephone. The Juster Scale was printed behind the cover letter. Respondents were asked to keep the letter near their telephone so that it would be readily available when an interviewer called later in the week. Two to three days later, interviewers contacted these respondents. After introducing and reading a statement of confidentiality, interviewers instructed respondents to have the letter with the scale in front of them. To those who had the letter, interviewers read out the following instructions:

> "*We would like to know what the chances are of you buying certain products during the next four weeks. The answers you may give are provided on the Juster Scale that is printed on the back of the letter we sent you. The answers are arranged on a scale a bit like a thermometer. If you are certain, or practically certain that you will purchase a product then you could choose the answer 10. If you think there is no chance or almost no chance of purchasing the best answer would be zero. If you are uncertain about the chance of purchasing choose an answer either as close to 0 or 10 as you think it should be.*"

The interviewer carried on to read out the following probability questions:

> *"Taking everything into account what is the chance that you or anyone else in your household will buy the following product in the next four weeks, that is between now and Christmas. Choose your answer from the Juster Scale on the rear of the letter sent to you.*
>
> - *one or more containers of margarine <RECORD RESPONSE>*
> - *one or more tins or packets of spaghetti <RECORD RESPONSE>*

- *one or more whole pre - cooked chicken. <RECORD RESPONSE>*

Now thinking just about yourself, and taking everything into account, what are the chances that you personally will do any of the following within the next four weeks. Again please give me a number between 0 and 10 on the Juster Scale.

- *buy a CD <RECORD RESPONSE>*
- *buy a paperback book <RECORD RESPONSE>*
- *buy a pair of shoes <RECORD RESPONSE>*
- *eat a meal at a restaurant <RECORD RESPONSE>*
- *go to the movies <RECORD RESPONSE>*
- *travel in a taxi" <RECORD RESPONSE>*

If a respondent was unable to trace the letter then, the interviewer called off the interview. Arrangement was made to send him or her, a copy of the covering letter.

No Mail Group: In this group, interviewers directly contacted respondents and requested them to indicate a purchase probability on a scale of zero to ten. After introducing and obtaining agreement to be interviewed, the following statement was read:

> *"We would like to know what the chances are of you purchasing certain products during the next four weeks. I would like you to answer on a scale of 0 to 10. If you are certain or practically certain that you will purchase a product, you would choose the answer '10'. If you think there is no chance, or almost no chance of purchasing the best answer would be '0'. If you are uncertain about the chances, choose another answer as close to '0' or '10' as you think it should be. You can think of the numbers as chances out of ten. For example, 3 would mean 3 chances in 10 that you would buy the product, while a 7 would mean that there are 7 chances in 10 that you would buy the product and so on."*

Interviewers then read out the specific question as used in the Mail Group.

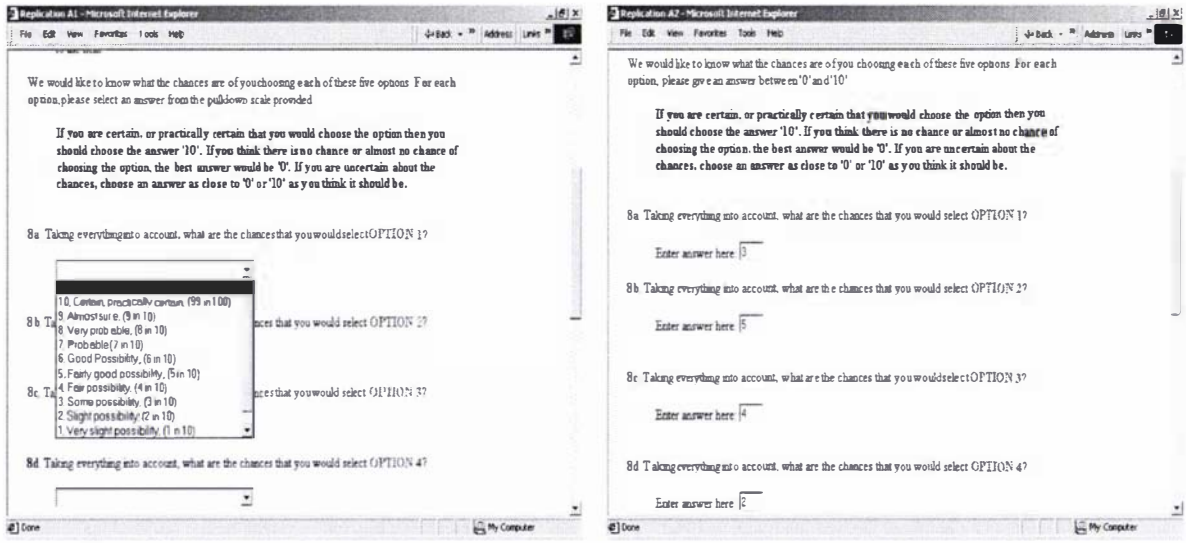
Both approaches produced a response rate of around 75%, suggesting that they were equally effective to collect probability data. The forecasts made in the No Mail Group were slightly more accurate than those made in the Mail Group. The overall satisfactory

performance, and more importantly the cost effectiveness of the Verbal Probability Scale, led the authors to recommend this scale to collect probability data over the telephone survey.

Internet-Based Survey

The prospect of reaching a wide audience in many countries has led researchers to develop the Internet for survey purposes (Brennan, Rae & Parackal 1999; Kottler 1997a; 1997b; Burke 1997; Knoth 1997). Parackal & Brennan (1999) tested the Juster Scale against a Verbal Probability Scale for implementation in an Internet-based survey. The Juster Scale was made into a pull-down menu format and the Verbal Probability Scale was a printed version of Brennan *et al.* (1992) (see Figure 2.10). These scales were implemented in separate treatments. Respondents were randomly assigned to the treatments and were asked to use the corresponding scale to indicate their probabilities.

Figure 2.10 The Juster Scale and the Verbal Probability Scale (Parackal & Brennan 1999)



The study was done on the clientele of an Internet Service Provider and respondents were asked to indicate their probabilities to subscribe to five new payment plans that the provider was planning to introduce. These plans were mutually exclusive and clients had to decide

on one of them if they were to continue as customers to the provider. Hence, probability scores assigned had to add up to ten to convey the purchase behaviour of the sample (this issue will be discussed further in Chapter Three). The authors observed that the scores of most respondents failed to add up to ten. The scores had to be weighted to one to logically convey the purchase behaviour of the sample towards each plan. Student's t-test was executed to see if the weighted mean probability scores obtained in the treatments were similar or different. For all five options, differences observed were not significant ($p > 0.05$), suggesting that both scales produced similar forecasts. The results concluded that either of the versions could be used in Internet-based surveys.

2.5 Chapter Summary

The poor performance of intention scales to forecast future purchase behaviour turned researchers' attention to probability scales. The pioneering studies compared an eleven-point probability scale with a five-point intention scale. Results of these studies have conclusively established probability scales as being better suited for forecasting purchase behaviour. Subsequent research efforts were directed at developing this type of scale. Much of these were concentrated on the Juster Scale. This scale was successfully tested to forecast purchase rates and purchase levels. Versions of the scale have been successfully tested for implementation in self-completion questionnaire, telephone survey and Internet-based surveys.

The overall performance of the Juster Scale has been satisfactory. In general, literature recommends this scale for collecting probability data (Armstrong 1986). All the same, there were problems that require research attention. Some of them were mentioned in this chapter. These problems are further scrutinised and discussed in Chapter Three.

3. PROBLEMS OF THE JUSTER SCALE

3.1 Introduction

The Juster Scale is one of the preferred probability scales used to collect probability data (Armstrong 1986). In recent years the scale has been successfully employed in assessing customer loyalty (Garland 2002; Riebe *et al.* 1998; Danenberg & Sharp 1996a; 1996b; 1999), in choice modelling studies (Rungie & Danenberg 1998), modelling repeat purchase using the Dirichlet model (Wright, Sharp & Sharp 2002) and in segmentation studies (Reid & Wood 2002; Riquier, Luxton, & Sharp 1997). The scale is successfully used in expectation and intentions surveys in many countries (Nysveen & Pedersen 2004; Wright, Lees & Garland 2002; Ryan & Huyton 2000; Corkindale & List 1999; Ryan & Huyton 1998).

While the Juster Scale has been subjected to repeat testing, there has been no satisfactory assessment of its reliability. In all the studies reviewed, the scale exhibited high positive correlations with actual behaviour data. This was the only indicator of its overall reliability. When the accuracy of the scale was compared across studies vast variation existed (absolute forecasting error ranged from 4% to 150%), raising concerns about its reliability. Accuracy of the scale was also inconsistent across product categories. The lack of consistency across product categories further raised concerns about its reliability, prompting the current review of literature.

So far the review of literature done for this thesis has identified two issues that appear to be responsible for some of the variations in the accuracy of the Juster Scale. They are the use of sub-optimal samples in Juster Scale studies and the context of the Juster Scale. The review carried out also led to the identification of two areas of improvements for the Juster Scale. One was to verify whether the attachments of the Juster Scale (verbal and numerical descriptors) were aiding in the collection of accurate probability data and the other was to address a problem that probability data of mutually exclusive behaviours collected on the Juster Scale had. In the following sections the issues and areas of development introduced here are discussed under separate headings.

3.2 Sub-Optimal Samples

Forecasts based on probability data were obtained by averaging the probability scores collected in the sample (Day *et al.* 1991). To a large extent, averaging cancels out random errors, making the forecasts resemble the actual behaviour. To facilitate the natural removal of random errors it is essential to collect data from a probability sample. Many of the Juster Scale studies reviewed (Brennan *et al.* 1995 b; Seymour *et al.* 1994; Day *et al.* 1991; Gan *et al.* 1986; Clawson 1971) failed to use probability samples. Further, the approach of testing the scale concurrently on different products (Brennan *et al.* 1995a, 1995b; Brennan & Esslemont 1994a, 1994b; Seymour *et al.* 1994; Hamilton-Gibbs *et al.* 1992; Day *et al.* 1991; Gan *et al.* 1986; Gabor & Granger 1972; Clawson 1971; Gruber 1970; Heald 1970; Juster 1966) made randomisation of samples difficult for all the test products.

Three studies (Day *et al.* 1991; Gan *et al.* 1986; Clawson 1971), in particular, tested the Juster Scale concurrently on three product categories, providing comparisons of their results. The accuracy of forecasts obtained varied considerably across the categories and the studies (ranging from –5% to +245%) (see Tables 3.1, 3.2, and 3.3). Two of these studies used a cluster sampling method (Gan *et al.* 1986; Clawson 1971), while the third (Day *et al.* 1991) used a panel list obtained from a research company. These studies also failed to set criteria to ensure that a reasonable cross-section of individuals was surveyed with respect to the test products.

Clawson (1971) tested the Juster Scale on nine different products (attend movies; travel outside South California; ride local bus; trip in a camper, motor home, or travel trailer; buy common stock, preferred stock or mutual fund; move to a different house; open a savings account; buy or lease automobile; buy TV set) concurrently and found quite different levels of accuracy across the categories. Forecasts were satisfactory for "ride local bus" and "attend movies", whereas, they were rather poor for "buy TV set" and "open a savings account".

The study used a cluster sampling method, selecting groups of six respondents from a 50-block area. The general rule to obtain sufficient variation in the sample, when using this method, is by maximising the within group variance and minimising the between

group variance on the criteria variables (Hair *et al.* 2000). It appears that Clawson (1971) did not employ such a procedure. Even if it was employed, it is doubtful whether sufficient variance could be achieved for all the test products.

Clawson (1971) categorised the test products into high (e.g. attend movies) and low (e.g. buy TV set) frequency activities. He observed that more respondents gave high probability scores (0.7, 0.8, 0.9, & 1.0) to high frequency activities, and comparatively very few gave high probability scores to low frequency activities (e.g. 104 gave high probability scores to "attend movies" compared to only 8 to "buy TV set"). Regression analysis executed on individual items showed that R-squares produced for high frequency activities were large (0.96 for Movie attendance and 0.98 for Bus ride), whereas for low frequency activities, it was considerable small (0.0018 for buying a TV set). The author attributed the large R-squares obtained for the high frequency activities to the absolute number of high probability scores.

If the reason attributed by Clawson (1971) was true then merely increasing the sample size could improve the R-square. A closer examination of Clawson's two examples revealed a sampling issue rather than the absolute number of high and low probability scores as the cause of the variations in the R-squares. From what was reported for "attend movies", it could be worked out that 59%[1] of the sample gave scores across the upper end of the scale (0.7 or more). The remaining 41% gave scores across the lower end of the scale; this includes those who failed to give a response. The proportions of high and low probability scores, though not equal, were reasonably spread across the scale. This suggests that the sample contained a satisfactory cross-section of respondents with respect to this product category. When probability scores in the sample were averaged, individual errors got cancelled out, minimising the forecasting error. In contrast, only 4%[2] of the sample gave high probability scores to "buy TV set". The sample apparently was biased towards respondents who had very little or no interest in this purchase. This was also reflected by the small R-square (0.0018) obtained for this product ("buy TV set"). The poor performance of the Juster Scale as

[1] 104/176*100 = 59; where 104 is the number of high probability scores, 176 is the sample size

[2] 8/176*100; 8 is the number of high probability scores, 176 is the sample size

far as this product was concerned could be because of the sub-optimal sample used rather than the scale.

Gan *et al.* (1986) used a cluster sampling method, selecting 12 household members from each of 10 clusters. Analyses done on the raw probability scores showed that this study also failed to ensure sufficient variance in the sample. To examine the spread of the data, respondents were grouped into high (0.7, 0.8, 0.9 and 1.0) and low (0, 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6) probability scores. For fast moving products (high frequency purchases) the ratio of high and low probability was 59:41 (see Table 3.1). In contrast, for durables (low frequency purchases) the ratio was 97:3 (see Table 3.1).

Table 3.1 Proportion of High and Low Probability Scores Collected for Durables, Services and Fast Moving Products Observed in Gan *et al.* (1986)

	High Probability (%)	**Low Probability (%)**
Durables	3	97
Services	17	83
Fast moving products	59	41

Note: High probability = 7 to 10; Low probability = 0 to 6

In the case of fast moving products, the ratio of the high and low probability scores was close to being balanced. The sample seemed to have included a reasonably wide cross-section of respondents as far as this product category was concerned. Averaging the probability scores in the sample cancelled out much of the individual errors. This was reflected in the better forecasting accuracy obtained for this category in comparison with the others (see Table 3.2). The ratio of high and low probability shown in Table 3.1 followed the same pattern seen for the two examples in Clawson's paper.

Table 3.2 Forecasts and Actual Purchase Rates Reported in Gan *et al.* (1986)

	Forecast (%)	Actual (%)	Error* (%)	Absolute Mean Error
Durables				
New or used cars	18.5	17.4	+6.3	
Video recorder	3.8	1.1	+245.0	
Washing machine	5.9	2.2	+168.0	121.46
Electric jug	8.0	3.3	+143.0	
Food processor	6.4	4.4	+45.5	
Service	46.4	41.3	+12.4	
Bed night	16.4	12.0	+36.7	29.6
Shares or debentures	13.7	9.8	+39.8	
Skiing trip				
Fast-moving consumer products	48.4	37.0	+30.8	
	52.8	55.4	-4.7	14.9
LP recorder or tape	45.1	38.0	+18.7	
Pair of shoes				
Hard cover book				

* Error % = [(predicted purchase rate - actual purchase rate)/actual purchase rate] x 100

Source: Day *et al.* (1991) p 25

Table 3.3 Forecasts and Actual Purchase Rates Reported in Day *et al.* (1991)

	Forecast %	Actual %	Error* %	Absolute Mean Error
Durables				
New or used cars				
Video recorder	8.6	7.4	+16.2	
Washing machine	4.7	6.4	+26.6	89.9
Compact discs	4.9	1.5	+227.0	
	2.5	1.5	+67.0	
Service				
Bed night				
Shares or debentures	30.8	22.5	+36.9	
Meal out	30.2	36.5	-17.3	15.7
Movie	69.9	71.7	-2.5	
	38.7	36.7	+5.5	
Fast-moving consumer products				
Paperback book	52.7	38.0	+4.6	4.6

*Error % = [(predicted purchase rate - actual purchase rate)/actual purchase rate] x 100

Source: Day *et al.* (1991) p 26

Day *et al.* (1991) also concurrently tested the Juster Scale on the same three categories, using a sample of panel members. The sample make-up of this study could not be ascertained, as the data was not available. Rank ordering the absolute mean errors, however, showed that fast moving products (high frequency purchases) had the least error followed by services and durables (low frequency purchases) (see Table 3.3). This pattern was consistent with the previous two studies discussed above (Clawson 1971; Gan *et al.* 1986).

From the proportions of high and low probability scores, one could see that the variances in purchase behaviour could be different for the three categories (Gan *et al.* 1986; Clawson 1971). Examining the distribution of probability scores in Gan *et al.* (1986) and Clawson (1971) revealed that for high frequency purchases the samples included a reasonably satisfactory cross-section of individuals who gave high and low probabilities, whereas for low frequency purchases the samples had fewer individuals who gave high probabilities. The pattern in the absolute mean errors across the product categories observed in Day *et al.* (see Table 3.3) also suggests a similar outcome. In all three studies forecasting accuracy was better when probability data were obtained from a wide cross-section of respondents. A reasonable cross-section of the target population could be sampled by using an appropriate probability sampling technique. The three studies examined however failed to use such a sampling technique.

Defining the target population is an essential practice in survey research (Hair *et al.* 2000). It ensures the collection of data from individuals for whom the topic has relevance. Results of such samples could be confidently extrapolated to the target population. Not defining the target population, however, could result in the generation of forecasts that are not reflective of the behaviour in the target population. Such forecasts when validated would result in large forecasting error (Kingsley & Anderson 1998). This appears to have been the case with many Juster Scale studies reviewed. Studies that did adopt such practice, however, have reported comparatively low forecasting errors. For example, Urban *et al.* (1996) employed screening techniques to ensure that the sample was appropriate to collect purchase probability data for a new electric vehicle and an existing car brand (Toyota Celica). The forecast for the electric vehicle could not be validated because actual behaviour was not available, however, the forecast of the existing car brand was validated and the absolute forecasting error

reported was 10%. Another study that employed similar procedures to collect purchase probability data of a new camera reported an even lower absolute forecasting error of 5% (Urban *et al.* 1997).

Brennan *et al.* (1994a) collected purchase probability data for fast moving items from a random sample. The forecasting errors ranged between –8% and +30% in this study and were comparable to those made for the same category in the three studies (Day *et al.* 1991; Gan *et al.* 1986; Clawson 1971) discussed earlier (-5% to +31%) (see Tables 3.2 & 3.3). Perhaps, the samples used in Day *et al.* (1991), Gan *et al.* (1986) and Clawson (1971) were suitable only to collect purchase probability data for fast moving products.

Observations from studies reviewed so far suggest that some of the variations in the accuracy of forecasts made on the Juster Scale could be caused by the sampling method used. From a theoretical perspective this is a valid argument, however, no empirical studies were available to support the argument. Comparing the accuracy of forecasts made on the Juster Scale in an optimal and a sub-optimal sample would allow one to establish the extent sample nature influenced the performance of the scale.

To carry out such a study, a sample drawn from a population defined by certain criteria to match the study and test products could be thought of as being optimal for collecting probability data. In contrast, a sub-optimal sample would be one drawn from a population not defined by any specific criteria. To further explain, consider Internet-based services and products. These are services and products targeting Internet-users (Internet population), who might use them at some time. Hence they could be thought of as being prospective customers of such products. A probability sample of such Internet-users would be optimal to collect purchase probability data for such services and products. A probability sample selected from the general population would comprise of both Internet-users and non-Internet-users. As non-Internet users would have no use of such services and products their presence in the sample would make it less optimal than the former. By comparing the forecasts made on the Juster Scale in these samples (optimal and sub-optimal samples), the extent to which sample nature influences the performance of the Juster Scale could be established.

To explain the sampling procedure for this comparison, consider the New Zealand Internet population that comprises of 75% of those over the age of 10 (Ministry of Economic Development 2003). According to Statistics New Zealand, the spread of households with Internet access across the country ranges from 25% for Gisborne and West Coast to 45% for Auckland and Wellington (Statistics New Zealand 2001). With such a wide spread across the country, Internet-users could be thought to be normally distributed or nearing normal distribution in the New Zealand population. If this was true, then a random sample would include Internet users in the proportion they are present in the general population. As the sample is drawn from the national population it would also be representative of the Internet population of New Zealand. Respondents could be contacted via their postal address with a request to participate in an Internet-based survey (Dillman *et al.* 2001; Schonlau, Fricker & Elliott 2001; Quigley *et al.* 2000; Nichols & Sedivi 1998; Mehta & Sivdas 1995). By adopting appropriate survey techniques, a satisfactory response rate could be achieved (Couper Traugott & Lamias 2001; Schonlau *et al.* 2001; Quigley *et al.* 2000). Executing the survey on the Internet itself will define the target population as the Internet population. Respondents who participate in the survey would all be Internet-users, making the sample optimal for collecting probability data for Internet-based products and services.

A sub-optimal sample could be generated by implementing the same questionnaire in a mail survey on another set of respondents randomly selected from the general population. In this case, both Internet and non-Internet users would complete the survey, making the sample not as optimal as the former one. The questionnaire version administered could include Internet related questions to identify Internet-users in the sample. Forecasts made using the Juster Scale could be compared between the two samples for accuracy.

If forecasts obtained in the two samples were within the margins of standard error, then the argument raised in the current review of literature (that the nature of sample influences forecasts made on the Juster Scale) could be ruled out. In which case, there would be serious concerns about the reliability of the Juster Scale. On the contrary, if results were different, then some of the variation in the accuracy of the forecasts could be attributed to the sample nature. Forecasts would then have to be validated to establish the sample that produces the best results as far as the Juster Scale is concerned. A

comparative study of this type would clearly explain the sampling requirements for future Juster Scale studies.

3.3 Contextual Background

In survey research, investigators use questionnaires to ask subjects questions. Subjects in turn answer the questions by completing the questionnaire. Thus, questionnaires facilitate communication between researchers and respondents (Hair *et al.* 2000). Apart from knowing the language, both parties must be familiar with the context of the subject. Researchers must ensure that the questions are presented in the appropriate context to elicit accurate information (Sudman & Bradburn 1982; Schuman & Presser 1981).

The literature on context revealed various factors that deflected the context of survey questions. Very often researchers are unaware when this happens. Consequently, studies that appear similar were not comparable (Schuman *et al.* 1983; Sudman & Bradburn 1982; Schuman *et al.* 1981; Schuman & Presser 1981; Duncan & Schuman 1980). In this section, the contextual literature is first reviewed to set the background of the issue raised. This will be followed by a review of Juster Scale studies in the light of the knowledge gained from the contextual literature.

3.3.1 Context of Survey Questions

Duncan, Schuman & Duncan (1973) raised the issue of contextual effects on survey responses after comparing results of two surveys on religion conducted in 1958 and 1959 that were thought to be similar (Detroit Area Surveys). The surveys asked respondents the following question about changes in religious interest:

> *"All things considered, do you think you are more interested, about as interested or less interested in religion than you were 10 or 15 years ago?" (1958)*

> *"All things considered, has your interest in religion grown, remained the same, or decreased over the last 10 or 15 years?" (1959)*

Responses obtained for these questions were markedly different in the two surveys. The authors of the study found it hard to believe that such a drastic shift in religious interest had occurred within a gap of 12 months. The wording of the questions was slightly different, but not so different as to account for the large variation observed. On closer examination, Duncan *et al.* (1973) found that the contents of the two questionnaires were different. The authors suspected this to have altered the context of the question, resulting in different response distributions.

Turner & Krauss (1978) raised context as the cause for incongruity in the reporting of confidence indicators by two research syndicates. The authors looked at indicators obtained from surveys carried out by the National Opinion Research Centre (NORC) and by Louis Harris and Associates (Harris) between 1973 and 1977. Comparisons were made on the level of confidence in nine national institutions collected by the two surveys. Both surveys used the following question for this purpose:

> *"As far as the people running [institution] are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in them?"* (p. 458).

The time series data from 1973 to 1976 of the two syndicates provided 45 comparisons on the above question. In 27 of these comparisons, difference in estimates were more than five percent and in 10 comparisons the differences in estimates were more than ten percent. These differences may not be large enough to raise concerns but the year-to-year differences in the two time series caused the authors to subject the data to further scrutiny. This led to the discovery of the following differences:

- Harris reported higher confidence in 13 comparisons, whereas the NORC reported higher confidence in 27 comparisons.
- The stability of the indicators was seen to vary over time for the two series. For example, confidence in the Supreme Court reported by NORC was quite stable (+2%, -2%, +4%, 0%), while that by Harris exhibited considerable variations (+7%, -12%, -6%, +7%).
- The year-to-year directions of trends were reversed quite often in the two series. For example, confidence in organised religion exhibited "divergent

trends" between 1973 and 1976 (NORC versus Harris: +9% versus -4%; -20% versus 0%; +6% versus -8%).

Attempts to attribute seasonality, demographic make-up, and sampling error as possible reasons for these discrepancies failed. A further attempt to explain the variations by comparing the rank orders of the year-to-year indicators also proved futile. The authors then used a multivariate analysis of variance technique (Goodman 1972; Thiel 1970) to explain the variation between the two time series. For this technique, the dataset was viewed as having one dependent variable ("whether or not the respondent said she or he had a great deal of confidence") and two independent variables ("year of the survey" and "research syndicate"). Indicators obtained for the nine institutions functioned as separate measurements of the independent variables. The following four hierarchical models were formulated to carry out this test:

> *"**Model 1**: No change across time and consistent indicators. This model posits that within the limits of sampling error, public confidence was constant across time and equivalent between the Harris and NORC series (No main effects interactions).*
>
> ***Model 2**: Systematic change across time measured by consistent indicators. It posits that the year-to-year changes in public confidence are too large to be attributed to sampling error, and that the two indicator series show no significant inconsistencies (Main effect for year only).*
>
> ***Model 3:** Systematic changes across time measured by indicators that have a constant bias. This model also posits significant changes across time, but it allows the indicators to have a significant constant bias. While acceptance of this model comprises simple across-house comparisons of the level of public confidence, it would indicate that year-to-year trends in the two series were consistent in magnitude and direction.*
>
> ***Model 4:** Significant and inconsistent biases in indicators. If we reject Model 3, we are forced to explain the pattern of changes reflected in these indicators by positing significant biases that vary from year to year and house to house. Acceptance of this model compromises all comparisons (Significant Year x House interactions)"* (Turner & Krauss 1987, p 463-464)

Analyses were run for each institution to find out how best the two time series fitted the four models. Measure of fitness between responses estimated by the model and those observed in the time series was assessed using the Chi-square statistic. In almost all the Chi-square tests, Models 1, 2 and 3 were rejected. In 16 of the 18 tests, the Chi-square tests returned significant levels resulting in accepting Model 4 ($p < 0.05$). From this the authors concluded that the indicator varied across the years and syndicates.

The above observations led to examination of the surveys individually rather than examining them across the years or syndicates. This revealed that in 1976 and 1977, estimates of all institutions made by Harris were lower than those made by NORC. No obvious reason could be identified for this pattern over the two years (1976 and 1977).

Examining the questionnaire used in each survey (across years and research syndicates) revealed that they were different except for the question on confidence. The authors raised the possibility of respondents interpreting the questions according to contexts based on the contents of the questionnaire. This appears to be a logical reason for the different levels of confidence reported by the two syndicates. The above reasoning has the support of studies done in this area of question wording (Gendall 1998; Belson 1981; Converse & Presser 1986). These studies cited showed that interpretations influenced the response distributions.

Turner & Krauss (1978) pursued the contextual issue further by implementing a planned study in the 1976 NORC survey. The following question was presented in separate questionnaire versions in different order:

> *"Do you consider the amount of federal income tax, which you have to pay as too high, about right, or too low"* (p 466).

The above question was asked immediately after, and immediately before the following question in separate questionnaire versions:

> "*We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount*" (p 466-467).

Comparison of responses revealed that the question on federal spending sensitised the responses given to the question on federal tax. When the latter followed the question on

federal spending, 52% stated the federal tax was too high compared to 63% in the alternative version. The question on federal programs asked before the federal tax question caused a reduction in the number of individuals who felt that the tax was too high.

Based on the above observation, Turner & Krauss (1978) suggested implementing the questionnaire in its entirety to produce comparable results. Further support to make this suggestion was derived from a study by Stouffer (cited in Turner & Krauss 1978). Stouffer (1955) compared response on communism, conformity and civil liberties obtained in separate surveys by the American Institute of Public Opinion (Gallup poll) and NORC. The battery of questions was implemented in its entirety in the two surveys. Comparing the responses revealed that there was no difference between the first two items (communism and conformity); difference on civil liberties was marginal (-3%). Similar results were obtained for questions that enquired about attitudes towards surveys and government (cited in Turner & Krauss 1978). The U.S. Bureau of the Census and Survey Research Center, University of Michigan implemented the questionnaire in separate surveys. Discrepancies between the two surveys for individual items were marginal (differences ranged from –6% to 1%).

The observation made by Duncan *et al.* (1973) on contextual effect discussed earlier triggered further interest. Duncan & Schuman (1980) conducted a study to find out whether the differences between the 1958 and 1959 surveys reported by Duncan *et al.* (1973) were truly because of the context of the question. They implemented a 2x2 factorial design to investigate the effect of context and question wording on responses collected on the two questions of the 1958 and 1959 surveys. The first level of the design allowed the comparison of the context. This was made different by having a set of questions on beliefs before the question on religious interest in one questionnaire version. These questions were placed after the question on religion in the alternative questionnaire version. The second level of the design compared the wordings of the question. The wordings used in the 1958 and 1959 were implemented in separate treatments.

The questionnaire developed by Duncan & Schuman (1980) commenced with questions that had very little connection with religion. There were two questions that made

reference to religion in the questionnaire. The questionnaire was so developed to make it identical to the 1958 and 1959 questionnaires that followed the above structure. Analysing the interactions between variables revealed that question wording had very little effect on the response, whereas, the context showed significant effect. The authors' advice from this experiment was to maintain the context constant to produce comparable results.

Schuman *et al.* (1981) made a similar discovery, which was more an accidental one. A general question (see Table 3.4) on abortion used in the General Social Survey of the National Opinion Research Centre (NORC), Michigan, USA (1978) was included in the questionnaire implemented by the Survey Research Centre (SRC) in 1979. Favourable responses to the question received in the SRC survey were 18 percentage points higher than those received in the NORC survey (See Table 3.4).

Table 3.4 Discrepancy between NORC and SRC Results (Schuman *et al.* 1981 p. 218)

"Do you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children?"

	NORC-78	SRC-79	Adjusted SRC *
Yes	40.3%	58.4%	54.9%
No	59.5%	41.6%	45.1%

Chi sq = 48.34, df=1, $p<0.001$
*SRC results standardised on NORC education distribution

The two surveys were different on three counts, sample make-up[3], survey mode[4], and questions that preceded the abortion question. In the NORC questionnaire, the abortion question was followed by a specific question on abortion in the case of a defective child. In contrast, the abortion question was the only one included on that topic in the questionnaire used by the SRC.

Responses received to the above question in the two surveys were very different. Differences persisted even after the samples were made similar in representation (see

[3] In this sample, individuals designated as low education were underrepresented.
[4] The SRC survey was a telephone survey whereas the NORC was a face-to-face survey.

Adjusted SRC in Table 3.4). Some of the difference could be attributed to the survey modes used (Peter & Wilson 1992). This, however, was not the view shared by Schuman *et al.* (1981), instead they attributed the differences to the context created by the question on a defective child included before the question on abortion in the NORC survey, raising the concern of these results being in-comparable.

To verify the above observation, Schuman *et al.* (1981) tested the two questions (see Table 3.5) in a split-ballot study. They compared the order in which the general and specific questions were asked. The question order was altered ("general/specific" versus "specific/general") in separate questionnaire versions (June 1979).

Table 3.5 General and Specific Question used by Schuman *et al.* 1981

General question
Do you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children?
Specific Question
Do you think it should be possible for a pregnant woman to obtain a legal abortion if there is a strong chance of serious defect in the baby?

The proportion of respondents favouring abortion was higher by 13 % when the general question was asked first. Proportions of those favouring abortion in the case of a defective child were very much the same for the two orders (84% when asked first and 83% when asked after the general question). The order of questions significantly affected the general question (abortion), but had no affect on the specific question (defective child).

The experiment was replicated two months later (August 1979) with the questionnaire including questions on religious belief and a question on ambivalence. The order effect was stronger this time with respondents favouring abortion by an excess of 17% points when the question appeared first. Once again there was no difference in the responses received to the question on a defective child.

The response pattern exhibited in the two studies suggests that agreement to the general question implied agreement to the specific question; the reverse, however, was not true. The authors called this the "part-whole effect". It appears that when the general question on abortion was asked first, respondents spoke out their mind, but when it was placed after the specific question (defective child) the context of the general question was modified, resulting in a different response distribution.

McFarland (1981) tested the part-whole effect on topics such as the economy, energy, politics, and religion (see Table 3.6). The questionnaire included one general question to gauge the general attitude of respondents and a series of specific questions on each topic. Two versions of the questionnaire were administered in separate treatments. In one the general question preceded the specific questions and in the other the reverse was followed.

Table 3.6 General Questions on Energy, Economy, Politics and Religion used by McFarland (1981)

1) How would you describe the current problem in the United States?
   i. Extremely serious
   ii. Somewhat serious
   iii. Not serious at all

2) During the next year, do you think the economy
   i. Will get better
   ii. Will get worse
   iii. Stay the same

3) In general, how interested would you say you are in politics:
   i. Very interested
   ii. Somewhat interested
   iii. Not very interested\\

4) In general, how interested would you say you are in religion:
   i. Very Interested
   ii. Somewhat interested
   iii. Not very interested

Interest expressed in politics ($p = 0.001$) and religion ($p = 0.025$) was significantly more when the general question was placed after the specific questions. As for economy and

energy, the order did not alter the responses received. Questions on politics and religion asked for respondents' interest, whereas questions on economy and energy asked for respondents' judgement on the respective topics. The authors noted that questions that ask respondents to express interest were affected more by question order than those that ask respondents to make judgements. There was no question order effect in the responses of the 17 specific questions.

This study also compared the strength of correlation between the specific and general questions when the orders were altered. Overall, the effect of question order on the correlation between the general and specific questions was very minimal. Of the 17 comparisons of correlations, two were significantly different. In both cases, correlations were more when the general question was placed after the specific question. It was not possible to establish the order that produced the true response. The author, more by logical reasoning, suggests placing general questions before the specific ones to capture the actual feelings of respondents.

Smith (1979) asked three specific questions on happiness, with marital happiness being the last of the three before asking a general evaluative question on happiness. Respondents, who indicated greater marital happiness that was last of the three questions, also indicated greater general happiness. Respondents seem to think that the last of the specific questions (marital happiness) was part of the general question (general happiness). In this instance, the last of the specific question set the context of the general question.

In the three studies reviewed above, the number of specific questions asked before the general question varied. Schuman *et al.* (1981) had just one specific question, while Smith (1979) and McFarland (1981) had multiple questions. Answers to the general questions had an assimilative effect from the specific questions. This was evident in the two correlations (between the general and specific questions) that exhibited significant difference (McFarland 1981). As mentioned earlier, in both instances, correlation was high when the general question was asked after the specific question. When only one specific question was asked before the general question (Schuman *et al.* 1981), depending on the nature of the specific question the effect was either assimilative or

negative. All three studies reported that there was no order effect on responses given to specific questions.

Schuman et al. (1983) compared the effect of context on responses given to questions placed at different point in the questionnaire as opposed to one after the other. They used the following question in their study:

> ***Communist reporters***: *"Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?"* (p 112)
>
> ***American reporters***: *"Do you think a Communists country like Russia should let American newspaper reporters come in and send back to America the news as they see it?"* (p 113)

Hyman and Sheatsley (1950) first tested the above two questions and observed that the order influenced the responses collected. They found that when the question on American reporters was asked first, respondents answered the question on Communist reporters favourably. But when the order was reversed respondents answered the question on American reporters less favourably. Schuman & Presser (1981) repeated the test after a gap of 30 years and reported the same results. The sensitivity of the topic during that period could have set a certain context in the minds of respondents. The order of the questions in the study would have added to that context, promoting respondents to express in a reciprocating manner (Schuman *et al.*, 1983).

Seeing these results, Schuman *et al.* (1983) decided to use the two questions in a test to find out if the response differed when questions were together as opposed to when separated by other questions in between. Three questionnaire-versions were used to present the two questions in different orders (Communist/American (A); American/Communist (B); American/17 items/Communist(C)).

Table 3.7 Percentage of "Yes" Responses to the Question on Communist Reporter by Context and Contiguity (Schuman *et al.* 1983 p. 113)

	Com/Amer (A)	Amer/Com (B)	Amer/17items/Com (C)
Percent	44.4	70.1	66.4
Sample size	117	107	107
A vs. B: Chi sq = 15.20, p<0.001			
B vs. C: Chi sq = 0.34, ns			
A vs. C: Chi sq = 10.95, p<0.001			

The proportion of respondents favouring Communist reporters working in America obtained when the item on American reporters was placed first (Amer/Com) was similar to when the two items were placed separate but in the same order (Amer/17 items/Com) (see Table 3.7). The proportion of respondents favouring Communist reporters working in America when the item on Communist reporters was placed first (Com/Amer) was different to that obtained in the two versions (Amer/Com and Amer/17 items/Com) that had a different order (see Table 3.7). The authors concluded that the apparent contextual effect did not diminish by placing questions separately in the questionnaire.

Sigelman (1981) made similar observations about the following presidential popularity question used in the Gallup polls:

> *"Do you approve or disapprove of the way that _____is handling his job as President?"* (p 199)

A split-ballot study compared the responses to this question when placed at the start of the questionnaire (the very first opinion question) versus when placed further down in the order (33rd question). The questionnaire included 48 questions on topics such as social problems, energy problems, gasoline costs, home heating costs, nuclear power, pollution, drugs, and bribery by politicians. None of these questions made direct reference to the president. They could "politically charge" respondents to respond either favourably or unfavourably to the above question (Darcy & Schramm 1979). Results showed that there was no difference in the responses received to the above question in the two versions. Irrespective of where the question was placed in the questionnaire, it produced the same result. Difference between the versions was the numbers of

responses received. When the question was asked first 80 % offered their opinion, whereas about 89% offered their opinion when it was placed further below in the order. Studies reviewed so far revealed that context of questions had an important part in collecting valid responses. Results were seen to significantly change when the context of questions was modified. Evidence gathered from the contextual literature suggests two ways by which context can modify results. One was caused by the questions (topics) placed before the main question. Hence, the order of questions must be carefully decided when designing the questionnaire. The other was the interpretation of the question. This is not limited to questionnaire design but includes factors such as respondents' understanding and the question topic.

While the studies examined above demonstrated the problem of context in survey questions, they failed to elucidate how to secure the context to produce valid results. This was because of the types of question used in these studies that were mostly opinion questions (Schuman *et al.* 1981; McFarland 1981). Responses to such questions cannot be validated and hence there was no way to establish the context that produced valid responses. This, however, need not be the case for behavioural questions, as responses can be validated against the actual behaviours. As such the contextual requirement of behavioural questions can be investigated in greater depth. It, however, requires a systematic investigation to isolate the factors that define their contexts. Following which methods could be developed to keep them constant for everyone in the sample.

The approach of using certain questions before the actual question to modify its context was successfully employed in the contextual literature (Schuman *et al.* 1983, 1981; Schuman & Presser 1981; McFarland 1981). All the same, no evidence was found to suggest any particular types of question (attitudinal or opinion question) to do the task. Labaw (1980) suggested asking attitudinal and opinion questions before behavioural questions to set the context of the former. Juster (1966) suggested the inclusion of certain demographic questions before the Juster Scale to get respondents to give more accurate information about their purchase behaviour. Others have suggested the inclusion of a series of general questions before the behavioural ones to elicit accurate information (Bradburn & Sudman 1988). All these suggestions remain open to future investigation.

3.3.2 Context of the Juster Scale

The Juster Scale was designed for implementation in questionnaires. Hence, as seen in the contextual literature, question order and interpretations would influence the context of Juster Scale. In addition to these two factors, certain survey practice employed in many studies was seen as a threat to modify the original context of the Juster Scale. If these factors did modify the context then the scale got implemented in contexts different to those originally intended. Results of such studies may not be comparable because of differences in the context. This could be one reason for the variation in the scale's performance across different studies. While references have been made to the contextual requirements of the Juster Scale, scientific investigations were limited. In this section the earliest references to the context of the Juster Scale will be uncovered first. Following this, the three factors that appear to modify the context of the Juster Scale are discussed. In the final part of this section, studies that addressed the context of the Juster Scale are reviewed and recommendations for future research are made.

Earliest References to the Context of the Juster Scale

Considering the complexity of decision-making, merely getting respondents to give scores on the Juster Scale may not be sufficient to convey their future behaviour. Juster (1966) in recognising this stated that

> *"...a survey which, prior to asking about probabilities, contains questions on the household's income, income prospects, asset holdings, stock market participation, etc., may obtain more accurate judgement than a survey which does not"*(p38).

Juster was of the view that asking such demographics questions before the Juster Scale question could encourage respondents to give more precise purchase probability data. He, however, did not test to see if the addition of such questions improved the performance of the Juster Scale.

Pickering & Isherwood (1974) argued in the same vein, saying that purchase probability data alone may not be a good measure of purchase behaviour. They pointed out that using purchase probability scores, as the only predictor variable would have

> *"to assume that consumers are implicitly taking into account a range of other*

factors including their confidence, income, wealth, the state of the stock of existing durables, other financial commitments and wants as well as their desire or plans to acquire a particular item" (p 204).

Getting respondents to consider certain factors alone may not be sufficient, as a host of factors interplayed with one another before a behaviour is formed (Pickering & Isherwood 1974). The weighting each of these factors has towards forming the behaviour is subjective and varies from individual to individual. If respondents fail to give the required consideration to them then their purchase probability scores would not be reflective of their actual behaviour. Gabor & Granger (1972) reported one such factor that caused respondents to indicate purchase probability scores that were not reflective of their purchase behaviour. They observed that respondents who gave zero probability scores accounted for 60% of the actual purchases. These respondents were asked to indicate the purposes of their purchases, to which 90% stated that it was made to replace their existing product. Obviously they had not considered whether their existing product needed replacing when giving their purchase probability score. Consequently, their purchase probability scores failed to convey their actual purchase behaviour.

Suggestions made by the above authors (Pickering & Isherwood 1974; Juster 1966) expressed the importance of getting respondents to keep contextual relevance while indicating their purchase probability score on the Juster Scale. Questions that Juster (1966) suggested to include before the Juster Scale were to encourage respondents to interpret the purchase probability question in the context of their personal situations. Subsequent investigators, however, failed to recognise this and have included the Juster Scale by itself in questionnaires, without worrying about this form of context.

After much of the developmental work was completed, Brennan (1995) drew attention to the context of the Juster Scale from a questionnaire designing point of view. He attributed the context to the irrational forecasts of an innovation over a period of two years. Following this observation, Hoek & Gendall (1997b) tested two methods of providing context to the Juster Scale by including opinion questions prior to the Juster Scale. These were the two direct references to the context of the Juster Scale cited in the literature; one being an observation (Brennan 1995) and the other an empirical

examination (Hoek & Gendall 1997b). These two studies will be discussed in detail below under the relevant sections.

Factors that modify the context of the Juster Scale

Question Order

In many of the Juster Scale studies, investigators tended to include the scale in omnibus surveys (Brennan 1995; Brennan & Esslemont 1994; Brennan *et al.* 1994; Gan *et al.* 1986) or "piggyback" on other studies (QIS studies by Juster 1966; Byrnes 1964). Consequently, very little control existed over where the question appeared in the questionnaire. The placement of the question was not the matter of concern (Schuman *et al.* 1983; Sigelman 1981). However, the questions that appeared before the Juster Scale were a concern as they could modify its context and influence the response distributions (Schuman *et al.* 1983; Smith 1982; McFarland 1981; Schuman & Presser 1981).

Brennan (1995) reported the above contextual influence to have occurred to forecasts made in two successive omnibus surveys in 1994 and 1995 (the Palmerston North Household Omnibus Survey by the Department of Marketing, Massey University). The Juster Scale was included in the two surveys to collect purchase probability data of a laser disc player. Being an innovation introduced in 1994, its adoption rate was expected to increase in the following year (1995). This, however, was not reflected by the forecasts of the two years (forecast of 1994 was higher than that of 1995). The author noticed that in the 1994 questionnaire the question about the laser disk (Juster Scale) occurred after a set of questions about video and video stores, whereas in the 1995 questionnaire there was only one question about videos (*"whether the respondent had ever rented a video tape, on their own or with others"*, p 56). In the 1994 questionnaire, the section before the laser disc player question was focused on video and video stores. The video usage context of this section appears to have spilled over, modifying the purchasing context of the laser disc player question. In the 1995 questionnaire, the laser disc player question (Juster Scale) did not suffer such contextual influence. Brennan (1995) suspects the contexts of the Juster Scale in the two questionnaires to be responsible for the reverse trend in adoption of this product. It

would be dangerous to make conclusions about the performance of the scale based on these two studies alone.

Respondents' Interpretations

The ideal environment to collect purchase probability data is a purchasing environment. In such an environment, respondents would have the opportunity to weigh all factors that influenced their purchase behaviour before stating their purchase probability scores. Such an environment is almost impossible in a survey situation. This is because most respondents are caught unaware by invitations to participate in surveys. Further, information available to respondents while participating in the survey is limited. This was observed in a Juster Scale study that attempted to forecast the adoption of an innovation (McDonald & Alpert 2001). The level of accuracy obtained in this study was not very impressive. The authors attributed the poor performance to a good number of respondents not having sufficient information about the product.

All the same, respondents could be asked to consider themselves to be in a purchasing environment while participating in surveys. Juster (1966) attempted to do this by phrasing the question accompanying the Juster Scale as,

> *"Taking everything into account, what are the prospects that some member of your family will buy either a new or old car sometime within the next 6 months; between now and next January?"* (Juster 1966, p 57).

The above question appears to be appropriate to use with the Juster Scale and has been employed in subsequent studies (Parackal & Brennan 1999; Brennan *et al.* 1995a, 1995b; Brennan 1995; Gendall *et al.* 1991). Whether it fostered the desired purchasing environment requires investigation.

The literature has records of other forms of questions that were used with the Juster Scale. Juster, himself, used two versions in his study. The question mentioned above was used to collect purchase probability data for automobiles. The question used to collect purchase probability data of appliances further down in the questionnaire was slightly different. It was worded as follows:

> *"What are the prospects that some member of your family will buy either a*
> *< > sometime within the next 6 months; between now and next*

January?" (Juster 1966, p 58)

Juster reported that the probability data collected using the above questions exhibited certain differences. In particular, he observed that purchase probability collected for automobiles distinctly differentiated respondents into those who purchased and those who did not purchase. The differentiation was not that distinct for appliances. The poor differentiation for appliances could be because of the omission of the phrase "taking everything into account" as this was the only difference in the way probability data were collected for the two product categories.

There were still other question versions that accompanied the Juster Scale in the literature. Some of them extracted from the original manuscript are listed below:

- Urban *et al.* (1996, p 53)

  "From what you know about the Impact, what is the probability that you would purchase it within the next three years?"

- Brennan *et al.* (1995a, p 244)

  "We would like to know what the chances are of you buying certain products during the next 4 weeks. The answers you may give are provided on the Juster Scale that is printed on the back of the letter we sent you. The answers are arranged on a scale a bit like a thermometer. If you are certain, or, practically certain that you will purchase a product then you would choose the answer '10'. If you think there is no chance, or almost no chance of purchasing, the best answer would be 'zero'. If you are uncertain about the chances, choose another answer as close to 'zero' or '10' as you think it should be

  Taking everything into account, what are the chances that you, or anyone else in your household, will buy the following product within the next four weeks, that is between now and two weeks before Xmas. Please give me an answer between zero and ten from the Juster Scale on the sheet we sent you."

- Day (1987, p 65)

  "We would like to know something about your spending plans. Please think carefully about the things, which might affect what you and others in your household will buy during the next three months – that is between now and the end of December.

Now, please indicate exactly what you think the chances are that you, or anyone else in your household, will buy these items during the next THREE (3) months – that is, between now and the end of December:"

- Gan *et al.* (1986 p 42)

  "I want you to say exactly what you think the chances are that someone in the household will actually buy the items during the next three months – we're going back to three months again [Respondents at this point were required to indicate their chance for 3 months period]. (SHOW CARD B) This card shows the chances, labelled from 1 to 10. For example, if you were absolutely sure that someone would buy the item, you would choose"10"; if you thought, for example, that there was a 40% chance that it would be bought, you would choose"4", and so on.

  Now I'd like to go through the same list again, but this time this will be for the next 6 months period [Respondents at this point were required to indicate their chance for 6 months period]. (READ LIST)"

All the studies mentioned above employed the same approach of collecting purchase probability data using the Juster Scale. The only difference between them was the wording of the question and instructions accompanying the scale. Forecasting errors across these studies ranged from –17% to 245%. Some of the variations could be because of the product categories and also the survey mode used; all the same, the forecasting error range observed was quite large. A major concern about these versions was the difference in interpretations that could result in context variations, which in turn could influence the response distributions (Turner & Krauss 1978). There was no direct comparison of Juster Scale estimates between these versions to confirm the above observation. The literature on question wording, however, provided explicit support to the above reasoning of different interpretation, resulting in different response distributions (Gendall 1998; Converse & Presser 1986; Belson 1981). Similar observation was also made regarding the interpretations of verbal descriptors of scales (Worcester & Burns 1975). While standardised pre-tests were available to test questions to see if respondents interpreted them correctly (Belson 1981), no studies have been done to verify this about the question and instruction accompanying the Juster Scale.

Survey Practices

In many Juster Scale studies, investigators have tested the scale concurrently on different product categories (Brennan *et al.* 1995a, 1995, 1995bc; Brennan & Esslemont 1994a, 1994b; Seymour *et al.* 1994; Hamilton-Gibbs *et al.* 1992; Day *et al.* 1991; Gan *et al.* 1986; Gabor & Granger 1972; Clawson 1971; Gruber 1970; Juster 1966). While there is no contention regarding the intention of these investigators to obtain comparisons of Juster Scale across different product categories, this practice could seriously modify the context of the Juster Scale.

The contextual literature drew attention to the fact that contexts were altered by the order in which questions were asked (Schuman & Presser 1981). This was seen in the response pattern obtained in studies that compared alternative orders of questions on patriotism (Schuman *et al.* 1983; Schuman & Presser 1981; Hyman & Sheatsley, 1950) and ethics (Schuman *et al.* 1981). Similar order-effects could be present in studies that collected probability data of different product categories concurrently.

The relevance of each product to sample members is bound to be different. This was seen in the three studies discussed under Sub-optimal sample in Section 3.2 (Day *et al.* 1991; Gan *et al.* 1986; Clawson 1971). At the respondents' level, perceived relevance towards the first product could set the context and the tone of response to the subsequent products. This could result in respondents giving probability scores to products listed further down in the order that fail to reflect their actual behaviour. There is no study in the literature that examined the order effect of test products on the forecasts made on the Juster Scale. However, studies that collected probability data of a single item have produced satisfactory results with the absolute forecasting errors as low as 5% (Urban *et al.* 1997).

3.3.3 Methods of Providing Context for the Juster Scale

Hoek & Gendall (1997) compared two methods to provide context to the Juster Scale. Theirs was a polling study designed to investigate whether contextual questions placed before the voting question reduced the proportion of undecided voters (undecided voters were a potential source of forecasting error) and improved the accuracy of the forecast.

A two by three experimental design was used to compare two methods of providing context (reduced and developed context - see Table 3.8) across three methods of collecting polling data of which one used the Juster Scale (forced-choice questioning method, the Juster Scale and a secret ballot response mechanism).

Table 3.8 Reduced and Developed Contexts (Extracted from the questionnaire in Hoek 1996, p 267 and 273)

Reduced Context

1. Are you enrolled to vote in this year's General Election?
2. Thinking back to the 1990 General Election, which party did you vote for?

Developed Context

1. There have been several issues affecting New Zealanders discussed in the news recently, I'd like to ask you now which party you think has the best policies for dealing with some of these issues.
    a. Which party do you think has the best healthcare policies?
    b. Which party do you think has the best education policies?
    c. Which party do you think has the best policies for dealing with unemployment?
    d. Which party do you think has the best policies for dealing with crime and violence?
2. Now, I'd like to ask you some questions about your feelings towards the political parties. Generally speaking, do you think of yourself as being <read the names of parties> or what?
3. And how strong<party stated in the previous question> do you feel?
4. Are you currently a member of a political party?
5. Which party are you a member of?
6. Are you enrolled to vote in this year's General Election?
7. Now, thinking back to the 1990 General Election, which party did you vote for?

Results of Hoek & Gendall's study showed that the proportion of undecided voters was considerably reduced in the version that employed the developed context across the three methods. The reduction, however, did not translate into more accurate forecasts.

As for the accuracy of forecast, the reduced context produced more accurate forecasts in all the three methods. The reduced context comprised of questions on registration and past voting behaviour. These were equivalent and antecedent behaviours to the one in question and were seen as being more helpful to respondents to convey their future behaviour. The developed context included the two questions on registration and past voting behaviour and a series of attitudinal questions on party policies. According to the authors the poor association of attitude with future behaviour (Driver & Foxall 1986) must have flawed respondents' judgement of their future behaviour. This study suggests that using specific behavioural questions before the actual behavioural question was sufficient to encourage respondents to give accurate information about their behaviour. This was true across the three methods tested by Hoek & Gendall (1997).

Another method that researchers have successfully used to set the context of the Juster Scale was by getting respondents to pass through a purchasing environment before indicating their purchase probability scores (Urban *et al.* 1997; Urban *et al.* 1996; Hauser *et al.* 1993; Bruck 1988; 1985). This approach was based on the premise that consumers engaged in certain pre-purchase deliberation prior to making a purchase (Punj & Staelin 1983; Kiel & Layton 1981; Claxon, Fry & Portis 1974; Newman & Staelin 1972; Katona & Muller 1955). General Motors (GM) used such an approach to forecast sales for their electric vehicle. In this approach a simulation of future market conditions was used to provide a purchasing environment. Respondents after experiencing the simulation were asked to indicate their purchase probability on the Juster Scale. The simulation provided respondents with contextual information relevant to the test product. A virtual representation called Information Acceleration (IA) developed in Macromedia Director was used to facilitate the simulation of a purchasing environment. The simulation comprised of the following information sources set up in a virtual reality:

Showroom visit: Respondents viewed the product in a showroom setting. In the case of the electrical vehicle (Urban *et al.* 1996), respondents could virtually walk around the vehicle, open the doors, view the interior, and ask questions of a sales person.

Television advertisements: Respondents viewed footage of television ads produced for the purpose of this study.

Word of mouth communication: Respondents viewed video recording of

actors who posed as consumers belonging to different segments.

Magazine article: Respondents read articles written for the purpose of this study.

Based on the forecast made, GM decided not to go ahead with their plan to launch the product, hence, the forecast made for this product could not be validated. To give assurance to GM about their decision, Urban *et al.* (1996) included another product (Toyota Celica) in the study for which the company had sales figures. The mean absolute forecasting error for this product was 10%. A subsequent study employed the same virtual representation (IA) to collect purchase probability data for a new camera (Urban *et al.* 1997). In this study, after adjusting for differences in the marketing plan a mean absolute forecasting error of 5% was reported. In comparison with the previous studies discussed these (Urban *et al.* 1996; 1997) reported more accurate forecasts for the Juster Scale.

In Urban *et al.*'s approach, respondents were engaged in antecedent activities that might better serve in forming the desired context than asking antecedent behavioural questions as advocated by Hoek & Gendall (1996). The overall satisfactory performance of this approach makes it a potential candidate against which the context of the Juster Scale could be validated. Such an investigation would help in establishing whether the Juster Scale required additional contextual inputs to collect accurate purchase probability data.

Urban *et al.* (1997) compared virtual and real representations of different information sources (real showroom versus virtual showroom and real salesperson versus virtual salesperson) and found that one was just as good as the other. This finding provides support for using virtual representation to facilitate antecedent activities required for the investigation in this thesis. The investigation could be carried out using the Internet technology for easy deployment of Urban *et al*'s approach.

After standardising the contextual requirement of the Juster Scale, investigations could be done to address the contextual issues raised in the review of literature. These could be in the form of separate investigations designed to find out whether question order, respondents' interpretation and the practice of concurrent testing of the Juster Scale on test items deflect the context of the Juster Scale. Such a stream of investigation would

help in accounting for some of the variations in forecasting accuracy of the Juster Scale. Results of these investigations would also help suggesting survey procedures for the production of comparable results.

To conclude the discussion on the topic, investigations required to address the contextual issues of the Juster Scale raised in the preceding discussion are outlined below:

- Standardisation of the contextual requirements of the Juster Scale by investigating whether the scale required additional inputs to collect purchase probability data in a purchasing context.
- Investigate whether the question asked before the Juster Scale deflects its intended context.
- Investigate whether the practice of collecting purchase probability data for different products concurrently is appropriate from the context point of view.
- Investigate whether the question form accompanying the Juster Scale fosters the purchasing context.

3.4 Probability Scales

Literature contains records of other probability scales that were used with satisfactory results (Pickering & Isherwwod 1974; Stapel 1968; Byrnes 1964; Ferber & Piskie 1965); these were discussed in Chapter Two. All the same, the Juster Scale received most of the subsequent research attention. Reviewing the literature on the Juster Scale and other probability scale brought to light two issues that might interfere with their performance. These will be discussed under two headings namely, "Scale descriptors" and "Response distribution" in this section. Following this, studies that compared the Juster Scale with other probability scales will be reviewed to gain direction for subsequent investigation in this area.

3.4.1 Scale Descriptors

Juster (1966) was of the view that probability scales with just quantitative descriptors would be more accurate than those with both adjectival and quantitative descriptors.

Juster, however, did not endeavour to verify his claim. Pickering and Isherwood (1974) used an 11-point probability scale with just two adjectival descriptors placed at the two terminal ends ("Completely certain" alongside 10 and "No chance" alongside 0 on the scale). The adjectival descriptors used on this scale were very much reduced when compared with most other scales. Forecasting error reported on this scale was comparatively lower (4%) than those of most other probability scales (-5% to +577% reported in Reibe 2000). The low forecasting error obtained on Pickering and Isherwood's scale may be because of fewer adjectival descriptors, as this was the only difference between their scale and other probability scales (e.g. Juster Scale, Byrnes' Scale, Ferber & Piskie's Scale discussed in Chapter Two). Observation about Pickering & Isherwood's scale offers some support to Juster's claim of probability scales with quantitative descriptors being more accurate than those with both adjectival and quantitative descriptors.

Juster's apprehension about adjectival descriptors became evident when studies on verbal description of scales and question wording were reviewed. Worcester and Burns (1975) observed considerable variation in the meanings that respondents gave to adjectival descriptors of a Likert scale. The variation in the meanings translated into variation in the response distribution obtained on the scale. Likewise, Laswad (1997) observed considerable variation in the meanings that accountants gave to probability expressions. Similar inconsistency could exist in the way numerical descriptors were quantified; however, no study was found to confirm this.

If meanings given to scale-descriptors varied across respondents, they are bound to influence the response distribution collected on the scale. Worcester and Burns' (1975) study was the only experimental study cited that showed the confounding effect that meanings of scale descriptors had on the response distribution. There are, however, a number of studies that shows such effect arising from variation in the interpretation of question wordings (Gendall 1998; Gendall & Hoek 1990; Belson 1986; Converse & Presser 1986; Kalton & Schuman 1982; Schuman & Presser 1981). Gendall & Hoek (1990) concluded the following about question wording, which could also be said of the Juster Scale:

> *"The problem for questionnaire designers is knowing when wording variations have changed a question and when they have not. Only by systematically*

researching the effect of question wording variations will this distinction become apparent" (p. 36).

As the Juster Scale has both adjectival and numerical descriptors, it is more vulnerable to biases caused by variation in meanings of scale descriptors than probability scales with fewer descriptors. It may not be practical to test the meaning of each descriptor on the scale. What may be possible is a systematic comparison of the accuracy of forecasts made on scales with different combinations of scale descriptors. Such an experimental study would identify the probability scale with the set of descriptors that collects the most accurate purchase probability data.

3.4.2 Response Distribution

A distinct point of difference between the Juster Scale and other probability scales in the literature was with their response distributions. This was clearly seen in the response distributions of durables. All probability scales, including the Juster Scale, produced an "inverse J" shape response distribution curve (Gendall *et al.* 1991; Gan *et al.* 1986; Pickering & Isherwood 1974; Stapel 1972; Byrnes 1964; Juster 1966; Ferber & Piskie 1965). The Juster Scale, however, consistently produced a smooth curve, whereas all other scales had a distinct blip at the mid-point (at point 5) of the curve.

Gan *et al.* (1986) collected purchase probability data for three product categories (durables, services and fast moving products) on the Juster Scale. Smooth "inverse J" shape response distribution curves were observed for all durables, as seen in the earlier studies. In the cases of services and fast moving products, "U-shape" response distribution curves were observed. The response distributions were rather uneven with distinct blips observed at scale-point 5 for services and at scale-points 3 and 5 for fast moving products. There was no study in the literature that compared the response distributions of the Juster Scale with that of other probability scales for these categories (services and fast moving products). From the experience with durables, it is possible that the other probability scales also produced the "U-shape" curves. Whether the blips would also be reproduced must be investigated.

To understand the causes for these distinct blips on the response distribution, the

literature on measurement scales was examined. According to this literature, respondents tended to choose the mid-point when they were unsure of an answer on the scale (Peter 1979; Garner 1960). This could be one reason for the blip at the mid-point observed for probability scales other than the Juster Scale. The smooth response distribution of durables produced by the Juster Scale suggests that this scale discriminated purchase behaviour of respondents satisfactorily. Whether that translated into better forecasts in terms of accuracy needs to be investigated.

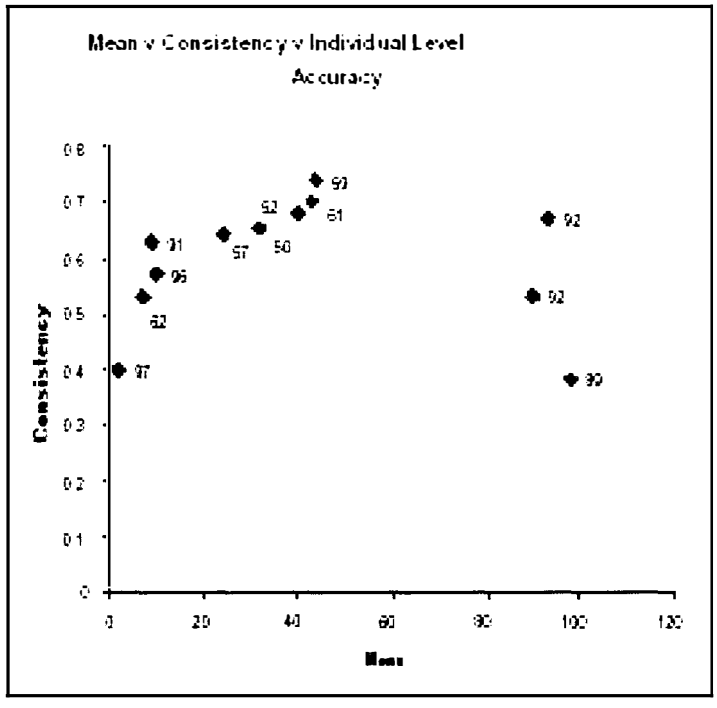
Researchers at the Marketing Science Centre, University of South Australia (Adelaide) have tried to use the Dirichlet model[5] to explain the differences between probability scales (Reibe *et al.* 1999). The beta distribution assumptions underlying the model were found to be true across most situations (Ehrenberg 1988). This provided Reibe *et al.* (1999) a standard to compare the response distribution of probability scales. Two hundred and eighty data sets obtained from 25 studies (conducted between 1995 and 1997) were examined for beta distribution fitness. Data that distributed to form the "inverse J" shape curve exhibited the best fit and data that aggregated at the mid-point of the scale had the least fit. The authors did not disclose the scales or the products of the dataset; hence, the better fit could not be related to any particular scale. The "inverse J" shape curve was obtained for durables in the earlier studies (Gendall *et al.* 1991; Gan *et al.* 1986; Pickering & Isherwood 1974; Stapel 1972; Byrnes 1964; Juster 1966; Ferber & Piskie 1965) and aggregation of scores at the mid-point for services and fast moving products (Gan *et al.* 1986). Whether this was what the beta distribution fitness reflected in Reibe *et al.*'s study needs further investigation. All the same, the study showed that variations occur, even in the beta fitness across data sets.

Reibe *et al.* (2000) investigated to see if the shape of the distribution reflected the accuracy of the forecast. Twelve data sets, for which forecasts and actual behaviours

[5] The Dirichlet model comprises of two distributions: the Negative Binominal Distribution (NBD) and the Dirichlet Multinomial Distribution (DMD) (Goodhardt *et al.* 1984). The NBD is a Poisson distribution depicting the rate of purchase and the DMD is a Multinomial distribution depicting the selection of brands. The model mixes the two distributions at the two ends with gamma of NBD depicting choices between multi-brands and beta of DMD depicting choice between two brands (Ehrenberg & Uncles 1998; Ehrenberg 1988). As the beta of DMD depicted choices between two alternative brands it was similar to the responses obtained on probability scales (choice between two alternative behaviours). The "inverse J" and "U" shaped distribution so far observed for probability scales were typical of the Beta Binominal Distributions (Ehrenberg & Uncles 1998; Ehrenberg 1988). Hence the two distributions could be thought of capturing similar types of responses.

were available, were used in the investigation. Effect of aggregate means and consistency (a measure of the shape of the distribution based on Riebe *et al.* (1999) on the accuracy of forecasts at the individual level) was examined (see Figure 3.1; numbers in the graph are the percentage of correct classification at the individual level). The authors report that accuracy at the individual and aggregate levels were best for data sets with low and high means (see Figure 3.1). In all but one dataset, the percentages of correct classification at the individual level were over 90%. Accuracy was comparatively less when the mean was in the 50s. The percentage of correct classification at the individual level ranged from 50% to 61% for means that centred in the 50s.

Figure 3.1 Correlations between Means, Distribution Shapes and Individual Level Accuracy (Reibe *et al.* 2000, p 1065)



According to Reibe *et al.* (2000), accuracy at the aggregate and individual levels improved as the shape of the distribution becomes more polarised (J shape and U shape distribution curve). It was not possible to ascertain the scale to which these shapes corresponded as those details were not disclosed in the article. Nonetheless the article revealed that there was association between the shape of the distribution and accuracy.

This might be a factor that could be used to distinguish probability scales with better accuracy.

3.4.3 Comparisons of probability scales

The Juster Scale was devised to smooth the blips observed in the response distributions produced by the earlier scales (Ferber & Piskie 1965, Byrnes 1964). The scale was successful in achieving that objective and was received as an improvement over the earlier scales. No attempt whatsoever was made by Juster (1966) to establish the accuracy of his scale in relation with the other probability scales. It was in the 1990s that researchers started to compare the Juster Scale with other probability scales for forecasting accuracy. Gendall *et al.* (1991) undertook a study that compared the Juster Scale with a Non-Verbal Probability Only Scale. For seven of the ten products tested the Juster Scale produced more accurate forecasts. The authors observed that the Juster Scale was user-friendly and respondents conveyed their purchase behaviour without any difficulty. In the case of the Non-Verbal Probability Only Scale, a small number of respondents (seven) created an additional point at the lower end of the scale to convey absolutely no chance of purchasing.

Brennan *et al.* (1995a) compared the Juster Scale with a Verbal Probability Scale to find out which one was suitable in telephone surveys. The Juster Scale was mailed out to respondents with instructions to keep it handy when an interviewer telephoned. Subsequently, interviewers contacted respondents and obtained probability data on the Juster Scale for the test products. The Verbal Probability Scale was implemented over the telephone by asking respondents to give a number between zero and ten that represented their chances for purchasing the same test products. Probability data was obtained for nine different fast moving consumer products. For all nine products the accuracy of the scales was similar. Comparing the individual errors however showed that for eight of the nine products, the Verbal Probability Scale produced less error. This latter observation and the fact that it was more cost effective led the authors to recommend the Verbal Probability Scale over the Juster Scale in telephone surveys.

Parackal & Brennan (1999) compared a Web-based pull down Juster Scale with a printed form of the Verbal Probability Scale (Brennan *et al.* 1995a) in an effort to

develop a version that could be used in Internet-based surveys. The scales were implemented on separate groups selected from the clientele of an Internet service provider. In both groups respondents were asked to indicate their probability to subscribe to five new payment plans, using the respective scales. The study found that the mean probability scores were similar for both scales across the payment plans.

Of the three studies discussed above the Juster Scale clearly produced better results in one (Gendall *et al.* 1991). The same could not be said of the Juster Scale in the other two studies (Parackal & Brennan 1999; Brennan *et al.* 1995a). The three studies used different survey modes and hence their comparability was a concern. To verify the comparability of these studies, literature that compared different survey modes was examined. McFarlane & Garland (1994) observed that data collected from health professionals by implementing the questionnaire using interviewers and via mail produced similar responses. This study compared the two modes for responses received to open-ended questions, acquiescence (the tendency to overstate agreement or favourability) and social desirability bias and found no significant difference on any of these factors. Studies that compared self-completion questionnaires administered by post and via the Internet also produced similar results (Chatman 2002; Burr, Levin & Becher 2001; Daly, Thomson & Cross 2000). Based on the literature cited and the fact that the three Juster Scale studies (Parackal & Brennan 1999; Gendall *et al.* 1991; Brennan *et al.* 1995a) have used similar approaches, their results may be comparable. In which case, the results were somewhat inconsistent.

The sections under Scale descriptors and Response distribution highlighted factors that might be connected with the accuracy of probability scales. Studies (Parackal & Brennan 1999; Brennan *et al.* 1995a; Gendall *et al.* 1991) that compared the Juster Scale with other probability scales were not conclusive in establishing any one scale as being better than the other. Hence, a study that compares the accuracy of forecasts made using these probability scales appears essential. Probability scales could be implemented in separate treatments and forecasts made could be validated against actual behaviour. Such a study would establish the scale that is best in terms of forecasting accuracy. The following probability scales were identified in the literature, which could be included in the comparison:

- The Juster Scale (scale with verbal and numerical probability description) (Juster, 1966)
- Scale with reduced or no verbal description (Pickering & Isherwood 1974; Ferber & Piskie 1965)
- Non-verbal probability only scale (scale with only numerical descriptors) (Gendall *et al.* 1991)
- Verbal Scale (scale that requires respondents to write or verbally give a probability score) (Parackal & Brennan 1999; Brennan *et al.* 1995a)

3.5 Mutually Exclusive Behaviours

Forecasting mutually exclusive behaviours, such as the purchase behaviours of competing brands and election results, has immense value to marketers. The Juster Scale has been used for this purpose (Fannelly, Fannelly & McLeado Jr 2000 a; Fannelly, Fannelly & McLeado Jr 2000b; 1999; Parackal & Brennan 1999, Hoek & Gendall 1997; Seymour *et al.* 1994; Hoek & Gendall 1993). Probability data of mutually exclusive behaviour collected on the Juster Scale could be used to understand switching behaviour between competing brands, market shares and customer preferences.

When making a choice from a set of mutually exclusive items, the choice is very much dependent on the relative influence of the alternatives. If the probability data were to convey the purchase behaviour of individuals and sample logically, the relative influence must be accounted for in the data. For example, consider collecting purchase probability data for four mutually exclusive brands. If a respondent stated seven probabilities, chances or odds out of ten to purchase the first brand, then the respondent is left with three probabilities that must be distributed across the remaining three brands. Accordingly, the respondent might assign two probabilities to the second brand, one to the third brand and zero to the fourth product. The probabilities assigned to the brands this way will add up to ten (7+2+1+0=10) and purchase behaviour in terms of probability could be interpreted as 70%, 20%, 10% and 0% probability to purchase the respective brands. Probabilities assigned this way logically reflect the purchase behaviour of the respondent towards each brand. In the sample, the aggregate probability of brands could be interpreted as the proportion of the sample that would

purchase them.

Studies reviewed showed that respondents, by and large, failed to understand probability in the way explained above (Fannelly *et al.* 2000a; 2000b; 1999; Parackal & Brennan 1999; Hoek & Gendall 1996; 1993). Many respondents in the studies cited have assigned probability scores treating the mutually exclusive items as being independent. Consequently, probability scores of items did not add up to one or ten and failed to convey the behaviour of the individual or the sample. In the literature two approaches were employed to handle this problem; these will be discussed in the following sections.

3.5.1 Weighting of Probability Scores

One approach used in the literature to rectify the problem caused by irrational assigning of probability scores to mutually exclusive items was by a weighting process. This process comprised of dividing probability scores assigned to each item by the total probability score across the mutually exclusive items. Weighting was done to the probability scores of each respondent before calculating the mean. Weighted probability scores added up to one and the means could be interpreted as the proportion of sample that favoured each item. Table 3.9 illustrates the weighting process using a hypothetical example.

Table 3.9 Example of the Weighting Processes Applied to the Raw Probability Scores

	Raw probability scores	**Weighted probability scores**
Brand A	0.5	(0.5/**2.3**)*1 = 0.22
Brand B	0.8	(0.8/**2.3**)*1 = 0.35
Brand C	0.6	(0.6/**2.3**)*1 = 0.26
Other Brands	0.4	(0.4/**2.3**)*1 = 0.17
Total	**2.3**	**1.0**

The above weighting process was employed first by Hoek & Gendall (1993) to correct the irrational assigning of voting probabilities. In their study, Hoek & Gendall (1993)

compared the Juster Scale with a conventional method (forced-choice question) to see which produced more accurate election forecasts. Respondents assigned to use the Juster Scale were asked to indicate their probability of voting for each of the participating parties on the scale. Respondents assigned to use the conventional methods indicated their absolute choice of party or candidate on the following question:

> *"If a general election had been held yesterday, which party would you have voted for?"* (Hoek & Gendall 1993, p.366)

The overall accuracy of forecasts obtained using the Juster Scale was better than that obtained using the conventional method. Absolute average difference between forecast and actual results across the parties was 4.7% for the Juster Scale and 7.2% for the conventional method. Everyone who used the Juster Scale expressed their voting behaviour in probability terms, whereas 22% of respondents in the conventional method stated that they were undecided. The above study (Hoek & Gendall 1993) also collected the probability for voting in the general election on the Juster Scale. The mean probability score for this was 82.7% and it closely matched the actual turnout in the electorate (84.6%) where the survey was carried out.

The overall performance of the Juster Scale was better than the conventional method. The investigators of the study (Hoek & Gendall 1993), however, observed that for many respondents their probability scores failed to reflect their voting behaviour. The probability scores given by these individuals to the competing parties or candidates failed to add up to one. The weighting process illustrated in Table 3.8 was used and the weighted probability scores logically conveyed the voting behaviour of individuals and the sample.

Flannelly, Flannelly & McLeod (1998) compared the Verbal Probability Scale (Brennan *et al.* 1995a) with a conventional method to forecast election results in three separate polling experiments. Forecasts, obtained using the Verbal Probability Scale, were within 1.4% and 4.4% of the actual results whereas those, obtained using the conventional method, were within 0.4% and 6.7% of the actual results (see Table 3.10). The chi-square test for goodness-of-fit showed that the differences between the forecast and actual results were not significant for both methods. Nevertheless forecasts made on the Verbal Probability Scale were more accurate than those made using the conventional

method.

Table 3.10 Percent of Actual Votes Compared with Forecasts Made On Verbal Probability Scale and Forced-Choice Method (Flannelly *et al.* 1998, p. 343)

Experiment	Candidate	Actual Results (%)	Predicted Results (%)	
			Verbal Probability Scale	Forced-Choice
One	Incumbent	62.1	60.1	-
	Challenger	31.3	26.9	-
Two	Incumbent	47.0	50.1	53.7
	Challenger	45.9	43.4	46.3
Three	Incumbent	54.1	57.1	59.5
	Challenger	43.4	44.8	40.5

In the above study (Flannelly *et al.* 1998), the sum of probability scores assigned to candidates did not add up to ten. The investigators, however, did not apply the weighting process (Hoek & Gendall 1993). In spite of that, forecasts made with the Verbal Probability Scale were better than those made with the conventional method. This raises the question of whether the weighting process was required at all in the first instance. The elections on which Flannelly *et al.* (1998) did their three experiments had just two candidates (Incumbent and Challenger). It is possible that the discrepancies caused were comparatively less in magnitude and hence the weighting process was not applied. As the authors (Flannelly *et al.* 1998) did not provide details of how many respondents gave scores that were logical and not logical, there was no way to verify the above assumption.

Flannelly *et al.* (2000a) repeated the above study in three telephone surveys implemented prior to three different elections (seven, fourteen and twenty eight weeks before the Election Days). The specific objectives of this study were to investigate whether the Verbal Probability Scale reduced the number of undecided respondents and to compare its accuracy with the conventional method. To achieve these objectives, the questionnaire presented the forced-choice question followed by the question that asked

for voting probability on the Verbal Probability Scale. The questions were structured to investigate whether respondents who stated "Don't Know" on the forced-choice question were able to express a probability on the Verbal probability scale. Respondents who refused to answer the voting probability questions and those who stated that they were undecided were coded as "Don't Know".

In all three surveys, the proportion of "Don't Know" responses received on the Verbal probability scale was relatively low. The average proportion across the surveys was 6.4% for the Verbal Probability Scale and 25% for the forced-choice question. In all three surveys, the Verbal Probability Scale produced better forecasts than the forced-choice question (% of error: -1.1 versus +12.3; +2.0 versus –11.6; +0.8 versus –9.8) (p = 0.001). All three elections covered in this study were two-candidate elections. The investigators did not employ the weighting process on the raw probability scores collected. Results of this study suggest that the weighting process may not be necessary although this may be specific to instances when there are only two competing parties.

Flannelly *et al.* (1999) (cited in Flannelly *et al.* 2000b) reported the observations from comparisons made between two-candidate and multi-candidate elections. The authors reported that forecasts made on the Verbal Probability Scale for two-candidate elections were more accurate than those made for a multi-candidate election. They also observed that respondents were able to give probability scores that were complementary to the two contesting candidates, and the probability scores added up to ten. This explained why the weighting process was not used in the studies done on elections that had only two candidates. In the case of the multi-candidates election, however, probability scores assigned to candidates did not add up to one. The scores had to be weighted to reflect the voting behaviour logically.

Flannelly *et al.* (2000b) compared the accuracy of forecasts between weighted and raw probability scores. Data for this study were obtained from 23 polling surveys implemented over the telephone on random samples. Thirteen of them were done on elections that had two candidates and the remaining ten were done on elections that had three or four candidates. All 23 surveys collected voting probabilities on the Verbal Probability Scale. The raw voting probability data for the multi-candidate elections failed to reflect the voting behaviour of the sample logically, confirming the earlier

observation of Hoek & Gendall (1993). Discrepancies in the voting probability data of two-candidate elections were not as severe as those in the voting probability data of the multi-candidate elections.

Comparing the mean error and absolute error of the raw and weighted probability scores revealed that differences were marginal in the two-candidate elections. The mean errors obtained for the raw and weighted probability scores in the two-candidate elections (-1.9 and –2.6) were in the same direction and the absolute mean errors were the same (4.9) (see Table 3.11).

Table 3.11 Mean Errors and Mean Absolute Errors of Weighed and Raw Probability Scores Obtained in Two and Multi-Candidates Elections
(Flannelly *et al.* 2000b, p 235)

	No of candidates	Raw probability scores	Weighted probability scores
Mean error	2	-2.6	-1.9
	3 or 4	+3.0	-4.0
Mean absolute error	2	4.9	4.9
	3 or 4	6.8	5.0

In the multi-candidate elections, error differences between the raw and weighted probability scores were substantial. The directions of mean errors were in the opposite directions, with the raw probability scores overestimating the election results by an average of 3% while the weighted probability scores underestimated by an average of 4%. The mean absolute error was less for weighted probability scores (5.0 versus 6.8), suggesting that the weighting process produced more accurate forecasts in multi-candidate elections.

The weighting process adopted in the above studies was based on the total probability score obtained by adding probability scores across the alternatives ("2.3" in the example in Table 3.9, page 75). This was done for each respondent separately; hence, the weighted probability scores could be thought as reflecting the behaviours after taking into account the relative influence of the alternatives. While there is no contention in

accepting it as a practical way to rectify the irrational assigning of probability scores to mutually exclusive items, the pattern of accuracy in Flannelly *et al.* (2000b) raised a certain concern. To explain the concern clearly the hypothetical example of Table 3.9 is extended in Table 3.12; this latter table includes a column representing hypothetical actual behaviours.

Table 3.12 Forecasting Errors of Raw and Weighted Probability Scores

	Absolute Mean Scores				Actual adoption
	Raw	***Error %**	**Weighted**	***Error %**	
Brand A	0.5	**400**	0.22	**120**	.1
Brand B	0.8	**300**	0.35	**75**	.2
Brand C	0.6	**200**	0.26	**30**	.2
Other Brands	0.4	**20**	0.17	**66**	.5
Total	**2.3**		**1.0**		**1.0**

*Error = [(probability score – actual adoption)/actual adoption] x 100 (Day et al, 1991)

Forecasting error is the difference between the forecast and actual behaviour; the same is expressed in percentage terms for the sake of comparison (column titled "Error %" in Table 3.10). Comparisons of the raw and weighted scores with the actual adoption showed that for three brands (Brand A, Brand B and Brand C) forecasts were overestimated while for one (Other Brands) it was underestimated. The error percentages of the weighted probability scores were less when the forecasts were overestimated ("Brand A", "Brand B" and "Brand C") and more when the forecast was underestimated (Other Brands) (see Table 3.10).

The weighting process is seen to reduce the magnitude of the probability scores for all brands. For example, the probability score of Brand A was reduced from "0.5" to "0.22" on weighting. The movement was always downward, irrespective of the error being positive or negative. On validation, forecasts that were over estimated (positive error) moved towards the actual behaviour reducing the error (see Figure 3.2). In the case of the one item that was under estimated (negative error), the movement was away from

the actual behaviour thereby increasing the error (see Figure 3.3).

Figure 3.2 Error Difference when the Forecast is an Over Estimation (Brand A)

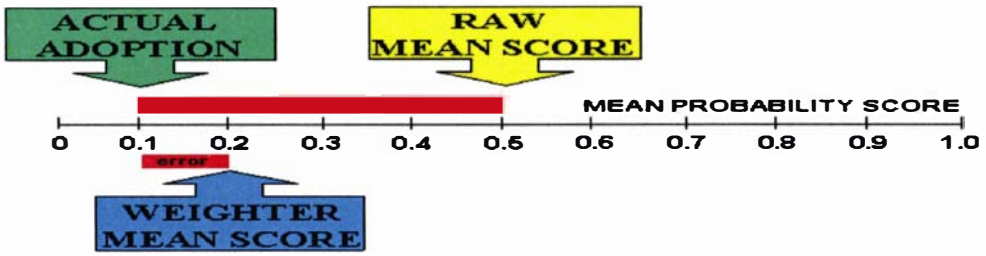
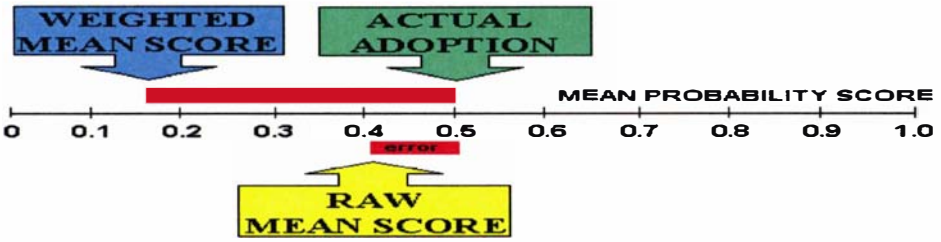


Figure 3.3 Error Difference when the Forecast is an Under Estimation (Other Brands)



The accuracy gained by the weighting process in the above example appears to be dependent on the forecast being either over or under estimated. Flannelly *et al.*'s (2000b) data seem to follow the pattern illustrated by the brands for which forecasts were over estimated in the above example (Figure 3.2). No study was found to confirm the reverse pattern of accuracy when forecasts were underestimated. One could, however, envisage the two patterns of accuracy illustrated in Table 3.12 as being possible. The illustration suggests that this way of handling the irrational assigning of probability scores to mutually exclusive behaviours needs more testing.

In all of Flannelly *et al.*'s (2000b) surveys, forecasts were over estimated. The improvement in the forecasting error achieved by applying the weighting process was evidently because of the forecasts being over estimated. This was clearly demonstrated by the example in Table 3.12 and Figure 3.2. No study was found to confirm whether the same improvement occurred when forecasts were under estimated as show in Figure 3.3. This may be a hard task to investigate, as it is difficult to tell beforehand whether a forecast would be over or under estimated. One could choose products based on previous experience, but there is always the possibility of forecasts turning out contrary

to what was previously seen. An alternative suggestion would be to test this method against one that does not require the weighting process, that is, a method that collects probability data for mutually exclusive items that adds up to ten.

3.5.2 Constant Sum Scale

Another approach used in the literature to collect probability data of mutually exclusive items was the Constant Sum Method. The method comprised of getting respondents to distribute a fixed number of points or units (10 or 100) and hence was called the "Constant Sum Method". The first record of this method dates back to 1947, when Milton Metefessel proposed it to collect comparative judgements on two or more items. The method comprised of getting respondents to assign 100 units (*"pennies, chips, points, marks, or material that can be counted"* Metefessel 1947, p. 230) to mutually exclusive items to convey their judgement of value for them. Values conveyed by assigning the units were expressed in terms of percentages. Metefessel (1947) suggested the method for collecting ranking and rating of mutually exclusive items. He also suggested that it be used for forecasting election results. By getting respondents to assign points to convey their rankings or ratings, distance or difference between respondents for the items can be known. Hence, data collected using the Constant Sum Method had the distance property activated. This permitted the use of advanced statistical techniques to analysis the data collected.

Alexrod (1964) compared sixteen scales used in forecasting purchase behaviour on three dimensions, namely sensitivity, stability and predictivity. The Constant Sum Scale was one of the scales. The scale however was implemented slightly differently to Metefessel (1947). Alexrod required respondents to distribute 11 cards across the items instead of 100 points. The following were the instructions used to explain the task required:

> *"Here's a sheet on which I have listed several brands of [product class]. Next to each brand is a pocket. Here are 11 cards. I would like you to put these cards in the pockets next to the brands to indicate how likely it is that you would buy each brand. You can put as many cards as you want in front of any brand or you can put no cards in front of a brand"*(p. 4).

Alexrod's (1964) study showed that the Constant Sum Scale was most accurate in forecasting repeat purchase. It was also more successful than other scales to distinguish respondents who intended to re-purchase from those who did not. The scale distributed respondents on a purchasing continuum. This allowed the data to be analysed using statistical techniques such as correlation and analysis of variance.

Reibstein (1978) compared the Constant Sum Method with two methods, one measuring preference (Dollar metric model) and the other attitude (Multi-attribute attitude model). The three methods were compared to find out which one produced the best forecast for brand choices made by respondents. The Constant Sum Method collected probabilities to purchase each brand. Respondents were asked to distribute 100 points across the alternative to convey their probability to purchase the brands. The Dollar metric model and Multi-attribute model collected preference and attitude as scale values. These values were converted into probabilities of choice for each brand using Luce's Choice Axiom (Luce 1959). The Constant Sum Method produced significantly more accurate estimates than the two alternative methods ($p < 0.01$). The Constant Sum Method was also successful in estimating the most frequently purchased brands correctly for 65% of the respondents followed by the Dollar-metric model (52%) and the Multi-attribute model (22%).

Hamilton-Gibbs *et al.* (1992) used the principle of the Constant Sum Method to collect probability data to forecast purchase levels of fast moving products. A household is bound to purchase several units of fast moving products over a certain period. As such, forecasting whether the household would purchase or not (purchase rate) has very little value. A useful forecast would be the number of units that would be purchased over the period (purchase level). Hamilton-Gibbs *et al.* developed a Constant Sum Scale to forecast the purchase level of fast moving products. The scale comprised of a grid printed on a flash card and 10 tokens (see Figure 2.9 in Chapter Two). Each column on the grid represents the number of units, starting from 0 to 12 units. Each token represented 0.1 probability or one in ten chances or odds. Respondents were required to distribute the tokens across the columns to convey their probability of purchasing different units of the product for the period in question. For example, if a respondent thought that there were five, three and two chances out of ten to purchase one, two and three units, then that respondent conveyed the chances by placing five, three and two

tokens in columns representing the respective units. The Constant Sum Scale was found to be more accurate than the Multiple Question Approach to forecast purchase levels of fast moving products (Hamilton-Gibbs *et al.* 1992). Brennan *et al.* (1995b) tested the Constant Sum Scale to forecast purchases of branded products (Coca-Cola, Campbell's canned soup). This study compared the Constant Sum Scale with the Multiple Question Approach. For both brands, the Constant Sum Scale produced better forecasts. The forecast of Campbell's canned soup was heavily overestimated by both methods (+ 158% on the Multiple question method and +102% on the Constant Sum Scale). Nevertheless the forecast made on the Constant Sum Scale was better. Seymour *et al.* (1994) confirmed the previous results; in particular they found the method advocated by Brennan *et al.* (1995b) (Constant Sum Scale that required respondents to lay the token flat against the units) as being more accurate (these studies were discussed in Chapter Two).

Over and above the satisfactory performance of the Constant Sum Scale, it collected probability data of mutually exclusive items that added up to ten. This prompted investigators to test this scale to collect probability data of mutually exclusive behaviours. Hoek & Gendall (1997) compared the Constant Sum Scale with a conventional forced-choice method to forecast election results in New Zealand. Under the MMP system in New Zealand, voters have two votes: the candidate and the party vote. The two methods were compared to forecast the outcome of the candidate and party vote.

The scale listed the names of parties and candidates and along the side of each party was ten cells numbered 1 to 10 (as shown in Figure 2.9, page 29). Interviewers gave respondents ten stickers with instructions to distribute them across the list of parties to indicate their chances. This was the same way in which Brennan *et al.* (1995b) implemented the scale. The following instructions were included to convey the task required:

> *"If you are certain, or practically certain to vote for a party you would put all 10 stickers beside it. If you thought there was no chance, or almost no chance of you voting for that party, you would not put any stickers next to it. If you were uncertain about voting for that party you would place as many stickers next to it as you think there should be. Could you please use these stickers to indicate the*

chances that you will vote for each of the parties listed in this grid" (p. 20).

The Constant Sum Scale was compared with the conventional method that enquired respondents' absolute choice of candidate or party. The following question was used in the conventional method:

> *"Which of these candidates [parties] do you plan to vote for in this year's General Election?"*(p. 8)

The mean absolute error obtained for the candidate vote on the Constant Sum Scale was slightly lower than the conventional method (3.1% against 3.8%). For the party vote, however, the Constant Sum Scale clearly produced the better forecast. The mean absolute error obtained on the scale was 1.8%, compared to 3.0% on the conventional method. This study showed that the Constant Sum Scale produced more accurate forecasts for election results.

The two approaches (Weighting process and Constant Sum Scale) discussed so far have produced satisfactory results in separate studies. All the same their relative accuracy of forecasts remains unknown. In the sections that discussed the weighting process suggestion was made to test the approach against a proven one for accuracy. The satisfactory performance of the Constant Sum Scale in all the studies reviewed makes it a potential candidate against which the weighting process approach could be compared. Such a comparative study would establish the method that is best for forecasting behaviour of mutually exclusive behaviours.

A draw back of the Constant Sum Scale is that it requires the assistance of an interviewer for implementation. The studies discussed above have all implemented the scale in face-to-face surveys. The advancement of the Internet technology and scripting languages, however, offers the opportunity to implement the scale in similar fashion in a virtual environment. The study suggested above could be carried out using Internet-based surveys so as to allow the easy implementation of the Constant Sum Scale.

Table 3.13 Summary Table of Literature Reviewed

Reference & Type of Study	Key findings & Comments
Sub-Optimal Samples	
Clawson (1971) Conceptual	Forecasting errors reported for high frequency purchases were comparatively lower than those reported for low frequency purchases. Sampling method - cluster sample.
Gan *et al.* (1986) Empirical	Probability scores were distributed evenly across the scale for high frequency purchases. Probability scores were concentrated to the lower end of the scale for durables. Sampling method - cluster sample.
Day *et al.* (1991) Empirical	In this study, the Juster Scale was tested concurrently on categories that were similar to the previous two studies. Rank ordering the absolute mean errors showed the levels of forecasting accuracy were similar to the previous ones. The sample for this study was selected from a panel list.
Kingsley & Anderson (1998) Conceptual	This article argues the requirement for defining the target population for forecasting studies.
Urban *et al.* (1996) Conceptual	In this study, a screening technique was used to identify prospective respondents. The study employed this to make the sample relevant to the test product. The forecasting error reported in this study was 10%.
Urban *et al.* 1997) Conceptual	This study also used a similar screening technique to identify prospective respondents. Forecasting error reported was even lower at five percent.
Brennan *et al.* (1994a) Empirical	Purchase probability data was collected for fast moving items from a random sample. Forecasting errors reported ranged between –8% and +30%.
Contextual literature	
Duncan *et al.* (1973) Conceptual	The authors raised the issue of context of survey question after comparing the results of two surveys on religion conducted in 1958 and 1959. The questions used in the two surveys were similar but

	the contents of the two questionnaires were different. This might have altered the context of the questions, resulting in different response distributions.
Turner & Krauss (1978) Conceptual and Empirical	This article observed substantial differences in confidence indicators of nine national institutes reported by the National Opinion Research Centre and Louis Harris and Associates between 1973 and 1977. A systematic investigation of the two time series identified the context of the questionnaires as the cause for the observed differences. The article also reported the results of a planned study on context, implemented in the 1976 NORC survey. The results revealed that general questions asked before specific questions sensitised the responses to specific questions. Suggestion was made to use the questionnaire in its entirety to produce comparable results.
Duncan & Schuman (1980) Empirical	To confirm the observation of Duncan *et al.* (1971), this study carried out a 2x2 factorial design investigating context and question wording on responses collected on the two questions on religion used in the 1958 and 1959 surveys. Results showed that question wording had very little effect on the response, but the contextual effect was significant.
Schuman *et al.* (1981) Conceptual & Empirical	This article compared the results of a question on abortion used by the General Social Survey of the National Opinion Research Centre (NORC) in 1978 and the Survey Research Centre (SRC) in 1979. Responses obtained in the SRC survey were 18 percentage points higher than in the NORC survey. The difference was attributed to a question about defective children that appeared before the abortion question in the NORC survey. This article also compared the order of the general and specific question on abortion by altering its order. Question order had significant effect on the general question (abortion), but no effect was seen for the specific question (defective child). A second experiment done two months later (August 1979) confirmed this result. The article introduced the idea of "part-whole effect", that is, agreement to general questions implies agreement to the specific question; the reverse, however, was not true.

McFarland (1981) Empirical	The author tested the part-whole effect on topics such as the economy, energy, politics, and religion. Comparisons were secured by alternating the order of general and specific questions in separate questionnaires. Part-whole effect was confirmed for questions on politics and religion. For economy and energy the effect was not seen. Questions on politics and religion asked respondents to indicate their interest, whereas questions on economy and energy asked respondents for their judgement.
Smith (1979) Empirical	In this study, three specific questions on happiness, with marital happiness being the last, were asked before asking a general evaluative question on happiness. Respondents, who indicated greater marital happiness, were seen to indicate greater general happiness. The last specific question (marital happiness) appears to have defined the context of the general question.
Schuman *et al.* (1983) Empirical	This study investigated whether the placement of question had any effect on its context. Results showed that the placement of question had no effect on the response.
Sigelman (1981) Empirical	Sigelman repeated the above investigation using a question on presidential popularity used in the Gallup polls. This study also found that question placement had no effect on the response
Context of the Juster Scale	
Juster (1966) Conceptual & Empirical	Juster held the view that asking questions on household income, income prospects, asset holdings, stock market participation, before the probability question could improve the accuracy of the Juster Scale.
Pickering & Isherwood (1974) Conceptual	Pickering & Isherwood argued that purchase probability data alone may not be sufficient to forecast purchase behaviour.
Gabor & Granger (1972) Conceptual	This article showed that respondents who gave zero probability scores accounted for 60% of the actual purchases. When these respondents were asked to state their purpose for purchasing, 90% stated that it was to replace their existing products. These respondents seem to have not seen the need for replacement when giving their purchase

	probabilities. When this happens, the resulting probability scores fail to convey the actual purchase behaviour.
Brennan (1995) Observation	Brennan was the first to draw attention to the context of the Juster Scale from a questionnaire-designing point of view. He blamed the context for the irrational forecasts of an innovation obtained over a two-year period.
Hoek & Gendall (1997b) Empirical	This study tested two methods of providing context to the Juster Scale. Results showed that asking questions about past behaviour prior to the Juster Scale helps respondents to give more accurate probability scores.
McDonald & Alpert (2001) Conceptual	In this study, the Juster Scale was used to forecast the adoption of an innovation. The performance of the scale was not satisfactory and was attributed to respondents not having sufficient information about the product.
Urban *et al.* (1996) Brennan *et al. (1995)* Day (1987) Gan *et al.* (1986)	These studies used different versions of questions with the Juster Scale. The forecasting error ranged from –17% to 245% across the studies. Some of the variation in forecasting error could be because of difference in question versions.
Brennan *et al.* (1995a), (1995b), (1995c) Brennan & Esslemont (1994a), (1994b) Seymour *et al.* (1994) Hamilton-Gibbs *et al.* (1992) Day *et al.* (1991) Gan *et al.* (1986) Gabor & Granger (1972) Clawson (1971) Gruber (1970) Juster (1966).	The literature review drew attention to a survey practice that could offset the context of the Juster Scale. In the studies listed, the Juster Scale was tested concurrently on a set of product categories. As seen in the contextual literature, the order of presentation could influence the responses given for items listed lower in the order. This needs to be investigated to find out whether such practice is appropriate for the Juster Scale.
Hauser *et al.* (1993) Conceptual	The authors developed a method for providing context by getting respondents to pass through a simulation of a purchasing environment. Studies that adopted this approach have reported a comparatively low absolute forecasting error (5% to 10%).

Bruck (1988), (1985) Conceptual	The articles listed suggest the use of a search engine to provide contextual information. The author argues that using search engines will negate the need for presenting the items in any particular order.

Probability Scale

Juster (1966) Observation	Juster was of the view that probability scales with just quantitative descriptors would be more accurate than those with both adjectival and quantitative descriptors.
Pickering & Isherwood (1974) Observation	Pickering and Isherwood used an 11-point probability scale with just two adjectival descriptors placed at the two terminal ends ("Completely certain" alongside 10 and "No chance" alongside 0 on the scale). Forecasting error reported for this scale was 4%.
Reibe (2000) Empirical	This article reported forecasting errors across different probability scales ranging from -5% to +577%.
Worcester & Burns (1975) Empirical	This article showed that considerable variations existed in the meanings that respondents gave to adjectival descriptors of Likert scales.
Laswad (1997) Observation	Observed variation in the meanings that accountants gave to probability expressions.
Gendall (1998) Gendall & Hoek (1990) Belson (1986) Converse & Presser (1986) Kalton & Schuman (1982) Schuman & Presser (1981)	The literature on question wording listed suggests that interpretations of question wordings could give rise to variation in responses collected.
Gendall *et al.* 1991 Gan *et al.* (1986) Pickering & Isherwood 1974 Stapel (1972) Byrnes (1964) Juster (1966) Ferber & Piskie (1965)	These studies concur with the general "inverse J" shape response distribution curve produced by probability scales. In the studies listed, the Juster Scale consistently produced a smooth curve, whereas other probability scales exhibited a distinct blip at the mid-point (at point 5).

Gan *et al.* (1986)	This study showed that the response distribution for services and fast moving products obtained using the Juster Scale emulated into U-shaped curves.
Peter (1979) Garner (1960) Conceptual	According to these authors, respondents tend to choose the mid-point when they are unsure of their responses.
Reibe *et al.* (1999)	In this article, the Dirichlet model was used to explain the performance of probability scales. The results suggest that scales that form the "inverse J" shape distribution curve exhibit the best fit. Scales that aggregated data at the mid-point had the least fit.
Reibe *et al.* (2000)	In this article investigation was done to find out whether the shape of the distribution reflected the accuracy of the forecast. Results suggest that accuracy at the individual and aggregate levels was best for data sets that had low and high means. Accuracy was comparatively less for data sets that had their means around the 50s.
Gendall *et al.* (1991)	In this study, the Juster Scale was compared with a Non-Verbal Probability Only Scale. For seven of the ten products tested the Juster Scale produced more accurate forecasts.
Brennan *et al.* (1995a) Empirical	This study compared the Juster Scale with a Verbal Probability Scale for telephone surveys. The two scales produced similar levels of forecasting accuracy. Comparing the individual errors showed that for eight of the nine products, the Verbal Probability Scale produced less error. Because of this, the Verbal Probability Scale was recommended for telephone surveys.
Parackal & Brennan (1999) Empirical	In this study, the Juster Scale was tested for implementation in Internet-based surveys. Comparisons were made between a Web-based pull down Juster Scale with a printed form of the Verbal Probability Scale (Brennan *et al.* 1995a). Results of the study showed that mean probability scores produced by the two scales were similar.

McFarlane & Garland (1994) Empirical	This study compared two survey modes, face-to-face interviewers and mail survey, for responses to open-ended questions, acquiescence and social desirability bias. Results suggest that there was no significant difference between the two survey modes for any of the test items.
Chatman (2002) Burr, Levin & Becher (2001) Daly, Thomson & Cross (2000) Empirical	These studies compared self-completion questionnaires administered by post and via the Internet. Comparison revealed that responses obtained by the two modes were similar.

Mutually Exclusive Behaviours

Hoek & Gendall (1993) Empirical	This article employed the weighting process to correct the irrational assigning of probabilities collected for mutually exclusive behaviour. The study reported, compared the Juster Scale with a conventional force-choice question for accuracy to forecast election results. The Juster Scale was seen to produce more accurate forecasts than the conventional method.
Flannelly, Flannelly & McLeod (1998) Empirical	In this study, the Verbal Probability Scale (Brennan *et al.* 1995a) was compared with a conventional method to forecast election results. Forecasts made using the Verbal Probability Scale were more accurate than those made using the conventional method. The authors reported discrepancy in the assignment of scores to candidates. All the same, the weighting process was not employed; still the scale performed better than the conventional method. As the elections covered in this study were two party elections (Incumbent and Challenger), the discrepancies caused were minimal.
Flannelly *et al.* (1999) Empirical	Reported the results of comparisons between two-candidate and multi-candidate elections. The Verbal Probability Scale produced better results in two-candidate elections than in multi-candidate election. Respondents were able to give probability scores that were complementary to the two contesting candidates. In the multi-candidates' election, however, probability scores assigned to candidates did not add up to one. The scores had to be weighted to reflect the voting behaviour logically.
Flannelly *et al.* (2000a)	This study investigated whether the Verbal

	Probability Scale could reduce the number of undecided voters. The study included three surveys and the questionnaire used presented the forced-choice question followed by the Verbal Probability Scale. In all three surveys, proportions of undecided voters were comparatively less for the Verbal Probability Scale. The scale also produced more accurate forecasts. The weighting process was not applied on the raw probability scores collected. It could be because the three elections covered were two-candidate elections and the discrepancies caused were minimal.
Flannelly *et al.* (2000b) Empirical	In this article, accuracy of forecasts was compared between weighted and raw probability scores. Data were obtained from 23 polling surveys implemented over the telephone. Thirteen of them were on two-candidate elections and the remaining ten were on three or four-candidate elections. All 23 surveys collected voting probabilities on the Verbal Probability Scale. The raw voting probability data for the multi-candidate elections failed to reflect the voting behaviour of the sample logically. Discrepancies in the two-candidate elections were not as severe as seen in the multi-candidate elections. Differences in the mean error and absolute error for raw and weighted probability scores were marginal in the two-candidate elections. In the multi-candidate elections, forecasting error differences between the raw and weighted probability scores were substantial. The mean absolute errors were less for weighted probability scores. In all the surveys, forecasts were over estimated. The improvement in the forecasting error achieved by applying the weighting process could be because of the forecasts being over estimated.
Metefessel (1947) Conceptual	Metefessel proposed to collect comparative judgements of two or more items using the Constant Sum Method. The method comprised of getting respondents to assign 100 units to mutually exclusive items. The method was suggested for ranking and rating of mutually exclusive items.

Alexrod (1964) Empirical	Alexrod compared sixteen forecasting scales for sensitivity, stability and predictivity and the Constant Sum Scale was one of the scales. This scale produced the best results for all three factors compared.
Reibstein (1978) Empirical	Reibstein compared the Constant Sum Method with a method measuring preference (Dollar metric model) and attitude (Multi-attribute attitude model) to forecast brand choice. The Constant Sum Method produced more accurate estimates than the two alternative methods. The Constant Sum Method was also successful in estimating the most frequently purchased brands correctly for 65% of the respondents followed by the Dollar-metric model (52%) and the Multi-attribute model (22%).
Hamilton-Gibbs *et al.* (1992) Empirical	Hamilton-Gibbs *et al.* employed the principle of the Constant Sum Method to collect purchase probability data for fast moving products to forecast their purchase levels. The scale was made into a grid, printed on a flash card with 10 tokens. Each column on the grid represents the number of units, from 0 to 12 units and each token represented 0.1 probability or one in ten chances or odds. Respondents were required to distribute the tokens across the columns to convey their probability of purchasing different units of the product for the period in question. The scale was compared against a Multiple Question Approach and the results showed it was more accurate than the latter to forecast purchase levels of fast moving products.
Brennan *et al.* (1995b) Empirical	In this study the Constant Sum Scale was tested for forecasting purchase levels of branded products (Coca-Cola, Campbell's canned soup). The scale was compared against the Multiple Question Approach. For both brands tested, the Constant Sum Scale produced better forecasts. The forecast of Campbell's canned soup was heavily overestimated by both methods. Nevertheless the forecast made on the Constant Sum Scale was better.
Hoek & Gendall (1997) Empirical	Hoek & Gendall (1997) used the Constant Sum Scale to forecast election results in New Zealand. The scale was compared against the absolute choice method to forecast the outcome of the candidate and party vote. The scale listed the names of parties and

	candidates and beside each party were ten cells numbered 1 to 10. Interviewers gave respondents ten stickers with instructions to distribute them across the list of parties to indicate their chances. The mean absolute error obtained for the candidate vote on the Constant Sum Scale was slightly lower than for the conventional method. For the party vote, the Constant Sum Scale clearly produced the better forecast. This study showed that the Constant Sum Scale was a better scale to forecast election results.

3.6 Chapter Summary

In this chapter some factors causing variations in the forecasting accuracy of the Juster Scale were discussed. From the relevant academic literature reviewed, two issues were identified as causing some of the variations. The first was the sampling methods used in the Juster Scale studies. The review revealed that in many studies the investigators failed to ensure the suitability of the sample to provide probability data for the test products included. This was made evident by an examination of the response distributions and forecasting accuracy of three studies that tested the Juster Scale on three similar product categories (Day *et al.* 1991; Gan *et al.* 1986; Clawson 1971). Observations from these studies suggested the possibility of the sample make-up causing some of the variations in the forecasting accuracy observed across the product categories. While there were theoretical reasons to believe this to be true, there was no study done to confirm this for Juster Scale studies. Suggestion was made to carry out a study that would establish the sample nature required for Juster Scale studies.

The second issue raised was about the context of the Juster Scale. The contextual literature revealed that a question asked in different contexts resulted in different response distributions. Reviewing the Juster Scale studies in the light of the contextual literature suggests that some of the variations in accuracy of the Juster Scale could be because of inadequate control of context. Question order, respondents' interpretation of the question and the practice of testing the Juster Scale concurrently on test products were identified as factors deflecting the original context of the Juster Scale. If these factors truly deflected the context of the Juster Scale then the scale would get implemented in a context different to what was originally intended. This would result in

such studies not being comparable. The scale would require fresh testing with controlled context to produce comparable results. Seeing the seriousness of the matter, a stream of research, commencing by verifying the contextual requirement of the Juster Scale, was suggested. Once the context was standardised, investigation could be directed at finding out whether question order, respondents' interpretation of the question and the practice of testing of the scale concurrent on test product deflected the context of the Juster Scale. If these factors were found to deflect the context then investigation could be continued to develop appropriate control measures to preserve the standardised context.

The latter part of this chapter discussed two areas of development for the Juster Scale. The first was to establish its accuracy in comparison with other probability scales. In the review of literature carried out four other probability scales that produced satisfactory results were identified. There were very limited attempts in the literature to establish the relative accuracy of these scales. Suggestion was made to compare the probability scales for forecasting accuracy and overall performance by implementing them simultaneously in separate treatments.

The next developmental work discussed was addressing a problem of the Juster Scale when used to collect probability data of mutually exclusive items. Review of academic literature revealed that respondents in general failed to understand how to express their behaviour towards mutually exclusive items in terms of probability. Respondents tended to be irrational in assigning probability scores to the alternatives. Consequently, mean probability scores of the alternatives failed to reflect the purchase behaviour of the individuals and sample. Investigators used a weighting process to correct the discrepancy caused by the irrational assigning of probability scores to mutually exclusive items. While the weighting process corrected the discrepancy, a concern was raised in the review of literature about the forecasting accuracy of the method. A hypothetical illustration was used to show that the forecasting accuracy of this method was dependent on the forecast either being an over or under estimation of the actual behaviour. While the method was practical, it requires more testing before being recommended.

Another method used by investigators to collect probability data of mutually exclusive items was the Constant Sum Scale. This scale was able to collect scores for mutually

exclusive items that added up to ten. It was also found to produce satisfactory results. The satisfactory performance of the Constant Sum Scale made it a suitable standard against which the weighting process could be compared. Suggestion was made to execute a comparative study over the Internet to permit the easy implementation of the Constant Sum Scale. The two methods could be implemented in separate treatments and the probability data collected could be compared for accuracy of forecasts.

All four issues raised in this chapter require research attention. It was, however, beyond the scope of the current research in terms of time and resource to address them all. In Chapter Five, the above topics are prioritised and the objectives set out in Chapter One are reiterated. Following this, that Chapter discusses the methodology adopted to achieve the objective of the research.

4. INTERNET-BASED SURVEYS

4.1 Introduction

The purpose of this chapter is to delineate the rationale for the Internet-based survey approach employed to carry out the data collection for this thesis. The first half of the chapter focuses on literature pertinent to the adoption of the Internet and related technologies in New Zealand. The latter half covers literature describing probability survey approaches implemented over the Internet.

4.2 Adoption of the Internet

The adoption of the Internet has been phenomenal in most countries during the last decade (Bankston, 1996; Killen, 1996; Rutkowshi, 1997). Over 498 million people around the world had Internet access from their homes by the end of 2001 (see Table 4.1). The percentage growth of Internet access from homes over the third quarter of 2001 was 24 million (Nielsen/NetRatings 2001; see Table 4.1). This works out to 5% of the total number of people with Internet access in 2001, which indicates the growth rate of home based Internet access.

Table 4.1 People with Internet Access via Their Home PCs in 2001

(Nielsen//NetRatings 2001)

	Number of People with Internet Access (in Millions)	Growth over the third quarter of 2001 (%)	Internet Population by Region in 2001 (%)
US/Canada	191.7	6.1	39
Europe/Middle East/ Africa	134.7	6.3	27
Asia Pacific	110.1	5.8	22
Latin America	20.7	0.7	4
Rest of World	41.0	5.1	8
Total	498.2	24.0	100

Nielsen//NetRatings, a subsidiary of A.C. Nielsen, operates as the Internet watchdog for many European and Asian countries. The company provides regular updates of Internet adoptions for these countries. Country wise figures for 2001 reveal that Internet adoption by households has crossed the 50% mark in many European and Asia Pacific

countries (see Table 4.3). On that list, New Zealand ranks fifth after Singapore, South Korea, Sweden and Hong Kong with 52% of households having Internet access in 2001 (see Table 4.2). This is over half of the target population that most marketing surveys are directed at. The observation suggests that New Zealand stands comparative above most other countries for adoption of Internet by households.

Table 4.2 Households with Internet Access and Telephone Connection (Nielsen//NetRatings 2002)

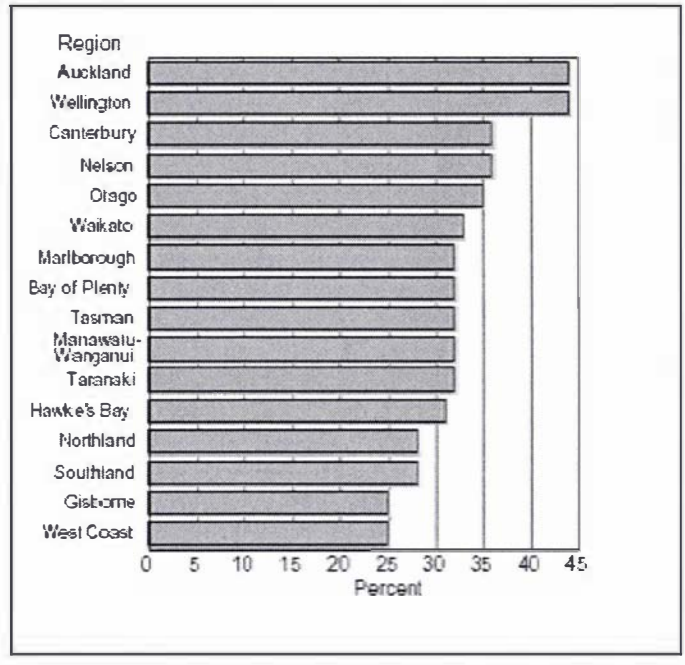
Country	Households with Internet Access via Home PC (%)	Households with Telephone Connection (%)
Singapore	60	89
South Korea	58	83
Sweden	57	87
Hong Kong	56	90
New Zealand	52	84
Netherlands	52	82
Denmark	51	82
Australia	51	77
Taiwan		83
Norway	47	78
Switzerland	43	78
Finland	42	81
United Kingdom	38	78
Austria	38	70
Israel	35	61
Germany	35	72
Italy	34	80
Ireland	34	76
Belgium/Luxembourg	32	68
Brazil	21	77
France	20	53
Argentina	20	55
Spain	18	48
South Africa	17	59
Mexico	14	56
India	7	66

Census 2001 of New Zealand reported a figure of 37% households having Internet access (Statistic New Zealand 2001). This figure is comparatively lower than the one reported by the Nielsen//NetRatings. The difference could be because of the timing of the two surveys. The census survey was held on 6th March 2001, hence, the Census 2001 figure reflects New Zealand's adoption rate at the start of that year.

Nielsen/NetRatings collect data for every quarter of the year and the figure reported in Table 4.2 was for the fourth quarter of 2001. The Ministry of Economic Development in its report on Information Technology of New Zealand for 2001 has used the Nielsen//NetRatings (2002) figure.

What was worth noting in the Census 2001 was the spread of the adoption rate across the nation; this ranged from 25% for Gisborne and East Coast to over 45% for Auckland and Wellington (Statistics New Zealand 2001) (see Figure 4.1).

Figure 4.1 Region Wise Household Accesses to the Internet in New Zealand (Statistic New Zealand 2001)



The comparatively high adoption rate reported by Nielsen//NetRatings (see Table 4.2) and the satisfactory coverage of the Internet in New Zealand seen in the Census 2001 (see Figure 4.1) prompted the use of Internet-based surveys for collecting data required to achieve the current research objectives.

4.3 Adoption of Computer and Browser Technology

An obvious trend in the technology sector is the increasing number of individuals using the Microsoft Windows operating system. According to Access New Zealand[1], 92% of visitors accessing its site from New Zealand used a Windows operating system (AcessNZ 1999). The corresponding figure for overseas visitors was 85% (see Table 4.3).

Table 4.3 Operating Systems and Web Browsers of Visitors to the AccessNZ Site (AccessNZ, 1999)

	NZ (%)	Overseas (%)
Operating system		
Windows	92.0	85.0
Macintosh	6.5	4.6
Unix	0.3	0.4
Others	1.4	10.1
Web browser		
Internet explorer	72	62
Netscape	27	30
Others	1	8

In the case of Web browsers, a large majority of those who accessed the AccessNZ site used Internet Explorer (see Table 4.3). Seventy two percent of those who accessed from New Zealand used Internet Explorer, 27% used Netscape and just one percent used other browsers. Internet Explorer was also the dominant browser used by overseas visitors (62%); all the same its usage was less in comparison to New Zealand visitors. Proportion of overseas visitors using Netscape was slight higher than New Zealand visitors (30% versus 27%), and that of other browsers was considerably higher (8% versus 1%).

Concerning browser versions, about 72% of all visitors to the AccessNZ site used version four and above of either Internet Explorer or Netscape (AccessNZ 1999). The figure (72%) was the same for both New Zealand and overseas visitors. This

[1] Access New Zealand (AcessNZ) provides a searchable directory of Web sites of New Zealand companies and organisations. AccessNZ is one of the high traffic sites in New Zealand with monthly average of 250,000 pages being viewed (http://www.accessnz.co.nz/statistics/).

observation suggests that many individuals used reasonably advanced computers that had the capability to run the recent browser versions. These browsers supported Web programming languages such as ASP (Active Serve Page), HTML (Hypertext Mark-up Language) 4.01, and ECMASript that were used to develop the Internet-based surveys implemented in the current research.

4.4 The Internet Technology

The remarkable adoption rate of the Microsoft Windows operating system and the latest version of Web browsers in New Zealand led to choosing Window-based programs to develop the Internet-based surveys implemented in the current research. Microsoft's Active Serve Page (ASP) was used as the platform for developing the Internet-based surveys implemented in this thesis. The survey interfaces were developed in ASP and powered by HTML and ECMAScript. In the remainder of this section these technologies will be described.

4.4.1 Active Server Page

Active Server Page (ASP) is an in-house technology developed by Microsoft to combine scripting languages and databases technology for Web applications (Microsoft, 2002). It has been in use since 1996 and has greatly improved the quality of Web sites. ASP codes are placed in between ASP delimiters ("<%" and "%>") usually before the opening tag (<html>) of a HTML page. Windows NT Internet Information Server (IIS), Windows NT Workstation and Windows 95 Personal Web Server recognise the delimiters and process the codes before sending the page to the remote computer. When these Microsoft Web servers receive a request for a Web page with ASP codes, the codes get compiled and executed before forwarding the page to the client. This way of compiling ASP codes each time the page is called allows changes to be made whenever required. The changes get compiled and executed when the Web page is requested of the server the next time.

ASP can initiate and maintain database connectivity between the server and remote computers. This feature allows the transfer of client-inputs into a database on the server. Based on the inputs, ASP can customise subsequent Web pages that clients receive.

This valuable feature was successfully used in managing and running the different functions of the Internet-based surveys implemented in the current research. ASP technology can use Extensible Mark-up Language (XML)[2], Component Object Model (COM)[3], HTML, and ECMAScript to create dynamic Web sites. These were successfully combined to write scripts that executed and managed the Internet-based survey implemented in this research.

4.4.2 Hypertext Mark-Up Language

The World Wide Web or WWW or Web for short is that part of the Internet that can be accessed using Web browsers such as Netscape Navigator, Microsoft Internet Explorer or Opera. The World Wide Web is made up of Web sites (e.g. Google.com, Microsoft.com and Massey.ac.nz) with a large number of Web pages linked one with another. Hypertext Mark-up Language (HTML) is the standardised scripting language developed for the Web.

In 1990, Tim Berners-Lee pioneered HTML that was made popular by the Mosaic browser of the National Centre for Supercomputing Applications (NCSA)[4]. To make the language usable on browsers around the world, HTML 2.0 was standardised by Berners-Lee & Connolly (1995). The subsequent years saw a team under Dave Raggett work at the World Wide Web Consortium[5] to improve the language. The team came up

---

[2]Extensible Mark-up Language (XML) is a standard format for sharing data on the Web, intranets, and across platforms and devices (Gatt & O'Dwyer 2002; Tamas et. al, 2000). It transforms structured data (spreadsheets, address books, configuration parameters, financial transactions, technical drawings etc) into a text file, which than can be moved around via the Internet.

[3] Component Object Model (COM) is the framework for developing and supporting program components and objects (Microsoft 1999). It enables building features such as interface negotiation, when components or object should be removed from a system, renewing and expiring of software license.

[4]The National Centre for Supercomputing Applications (NCSA) is a platform that brings the advancements made in software and hardware technologies into a single system. The centre collaborates with various American institutions and industries to ensure that these advancements are available to everyone. It commenced operation in January 1986 at the University of Illinois at Urbana-Champaign, USA as one of the five centres initiated by the National Science Foundation's Supercomputer Centres Program. The centre is recognised internationally for its high-performance computing and networking and software applications. The NCSA Telnet in 1987 was instrumental in making supercomputing and the Internet popular. The Mosaic, the first graphical Web browser was introduced on this network in 1992.

[5]The World Wide Web Consortium (W3C) was formed in October 1994 to develop common protocols to direct the evolution of the World Wide Web and ensure its interoperability. http://www.w3.org/Consortium/ provides information about the consortium.

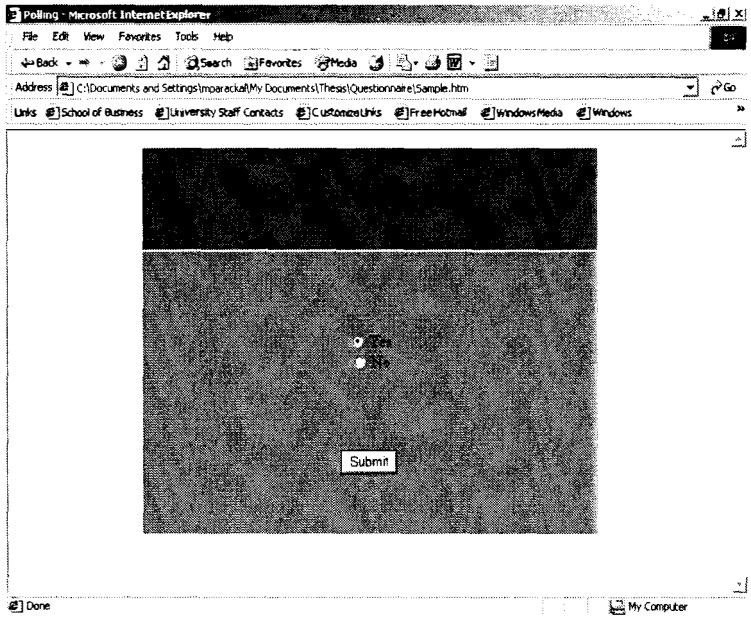
with HTML 3.0 that was an improvement over HTML 2.0. This version (HTML 3.0), however, failed to receive the acceptance of standard setters at the World Wide Web Consortium. Work to improve the language continued, leading to the successful standardisation of HTML 3.2 (Raggett 1997). In the following year, HTML 4.0 with more capabilities was standardised (Raggett, Le Hors & Ian 1998). The new capabilities include style sheets, frames, embedding objects, improved support for right, left and mixed direction text, richer tables, enhancements to forms and improved accessibility for people with disabilities. HTML 4.01 is the current version developed after correcting certain errors of HTML 4.0 (Raggett, Le Hors, & Ian 1999).

HTML files can be composed and edited on any operating system (Windows, Mac, and UNIX), using Notepad or an HTML editor. Files are saved as HTML files that have "htm", "html", or "shtml" as its extensions. It comprises of mark-up[6] symbols or codes that instruct Web browsers to display contents, following a logical structural sequence. Mark-up codes are elements usually referred to as tags and identified by the specific opening and closing tags. Figure 4.2 shows the mark-up codes compiled together using Notepad.

Figure 4.2 Html Mark-Up Codes Compiled using Notepad

When the file is requested by a client, the mark-up codes are transferred via the Internet to the remote computer. Web browsers such as Internet Explorer or Netscape decode the file to display it as a Web page as shown in Figure 4.3.

Figure 4.3 The Html Mark-Up Codes Displayed as a Web Page on Microsoft Internet Explorer



The two Internet-based surveys (Vodafone survey and New Zealand survey) implemented in the current research were developed in HTML 4.01. The questionnaire was formatted as Web pages that presented questions either one at a time as shown in Figure 4.3 or in groups according to the topic on respondents' computers. The hyperlink feature allowed respondents to navigate through the questionnaire.

4.4.3 ECMAScript

ECMAScript is an object-based scripting language developed to write client and server-side applications. ECMAScript codes can be embedded in between HTML or ASP tags. The technology was developed by Netscape Communications and standardised by the European Body for Standardisation of Information and Communication Systems in 1996. The standardised edition is contained in the ECMA 262 language specification

[6] Mark-up refers to the sequence of characters or other symbols insert in a HTML file to indicate how the page should look when displayed. The mark-up indicators are often called "tags."

(ECMAScript Edition 3). The ECMA 262 Standard describing the language is available from the ECMA Website (www.ecma-international.org) for anyone to adopt and further develop. Companies such as Microsoft and Netscape Communications have developed their own versions (Jscript and JavaScript respectively) based on the ECMA 262 Standard.

Web browsers such as the Microsoft Internet Explorer, Netscape Navigator, and Opera are developed to recognise and execute ECMAScript codes embedded in Html or ASP files. When a client requests a file containing ECMAScript codes, the server forwards the codes over the Internet to the remote computer. The Web browser opens the file to display the Web page according to the HTML or ASP codes and executes the client-side ECMAScript codes. ECMAScript helps to create dynamic Web pages that processing client inputs to effect communication between the remote computer and the host server. In the current research ECMAScript was used to develop the electronic version of the Constant Sum Scale that was used in the investigation of the second objective.

4.5 Internet-Based Surveys

The advantages of the Internet to collect and disseminate information were persuasive in employing it in survey research. This led researchers to develop the Internet for survey purposes (Aoki & Elasmar 2000; Askew, Craighill & Zukin, 2000; Bradley 1999; Rae & Brennan 1999; Vasja, Zenel, & Katja 1999; Kottler 1998, 1997a, 1997b; Schillewaert, Langerak, & Duhamel, 1998; Bublitz, 1997; Knoth 1997; Pitkow & Recker 1994a; Pitkow & Recker 1994b; Brucks 1988, 1985). All the same, researchers (Aoki & Elasmar 2000; Askew, Craighill & Zukin, 2000; Bradley 1999; Batagelj, Vehovar, & Katja 1999; Batagelj & Vehovar 1998; Bublitz 1997; Knoth 1997) are aware of the difficulty of using probability samples in Internet-based surveys. While non-probability samples may be used to address certain methodological issues of the Internet, they were viewed as being inappropriate for the current research because of reasons raised in the literature reviewed under Sub-optimal samples in Chapter Three. That review (Optimal samples in Chapter Three) drew attention to possible bias introduced by non-probability sampling technique on the Juster Scale forecasts. To control this, the current research collected data from a probability sample.

Researchers in the past have implemented probability surveys over the Internet with reasonable success (Deutskens, Ruyter, Wetzels & Oosterveld 2003; Couper 2001; Aoki & Elasmar 2000; Askew, Craighill & Zukin 2000; Bradley 1999; Jones & Pitt 1999; Dillman *et al.* 1998). These surveys, however, have been carried out on close populations (example: students of the University of Michigan). Contact information pertaining to members of such populations is usually maintained on records that were used as sampling frames in the studies cited. The response rates obtained ranged from 19%[7] (Jones 1999) to 62%[8] (Couper 2001) and were comparable with most other survey modes. An important point to mention about these studies was that everyone in the target populations had Internet access.

Sampling techniques used to survey close populations via the Internet may not be feasible to survey open populations (e.g. general population). The main problem is the poor coverage of the Internet (Crawford, Couper & Lamias 2001; Solomon 2000; Kay & Johnson 1999). Only those having Internet access will be able to participate in the survey and results obviously would be biased towards these individuals. In countries where the Internet coverage is about 50% (see Table 4.1), only half the population would be covered. Coverage bias would be worse in countries with lower Internet penetration.

In addition to the problem of coverage, the non-availability of suitable sampling frames prevents the Internet from being used in probability surveys on open populations. Further, the non-existence of non-list assisted sampling methods (similar to the random digit dialling) to select Internet users makes random sampling altogether impossible for Internet-based surveys (Berrens *et al.* 2001; Aoki & Elasmar 2000). Even if such a method existed, the strong repulsion towards unsolicited emails (spamming) would result in the method being less successful in obtaining a satisfactory response rate. A number of ways to overcome these problems were observed in the literature, these are discussed in the reminder of this section.

[7] Received for a health survey done on staff at ten universities in the USA

[8] Survey on drug and alcohol done on students of the University of Michigan

Research syndicates have developed their own methods to overcome sampling problems of the Internet (lack of coverage and sampling frame). Gordon Black of Harris Interactive (HI) developed a large panel of willing respondents from which samples were drawn for Internet-based surveys (Berrens *et al.* 2001; Rademacher & Smith, 2001; RFL Communications 2000). Panel members were recruited by running advertisements and sweepstakes, the Harris/Excite poll, telephone surveys, and product registrations on the Excite and Netscape Web sites (Taylor *et al.* 2001). According to Berrens *et al.* (2001), the Harris Interactivity panel consisted of about seven million adults from which samples are drawn for survey purposes.

Harris Interactivity came into the limelight after accurately predicting the closely contested 2000 presidential election between George W Bush and Al Gore (RFL Communications 2000). The National Council on Public Polling (NCPP) placed the Harris Interactivity Internet polls ahead of the traditional telephone polls and an innovative method[9] that used automated telephones (Rademacher and Smith 2001); the Rasmussen Research another research syndicate pioneered the latter method.

Knowledge Networks (KN) founded by political scientists Norman Nie and Douglas Rivers in 1998, pioneered an approach that generated a panel of randomly selected willing respondents (Berrens *et al.* 2001; Huggins & Eyerman 2001; Krotki & Dennis 2001). Telephone numbers were selected using the RDD method from geographical areas[10] covered by the Web-TV network. Postal addresses of the selected telephone numbers were identified from a list and advance letters together with incentives ranging from $5 to $10 were sent prior to contacting the households. Up to 15 attempts were made to contact an adult in the household before abandoning a telephone number. Individuals who agreed to participate as panel members were provided with free Web-TV units, Internet access, email addresses and ongoing technical support. Panel members also received separate incentives for participating in surveys and the occasional reward for remaining on the panel. In return, panel members were required to participate in at least one survey every week for a period of two to three years.

[9] This method used an automated telephone managed by a computer. The interviews were conducted using recorded voices.

[10] Eighty four percent of households in the USA are located in geographical regions that have Web-TV coverage.

Survey notices were emailed to panel members who met the screening criteria of each survey. The email message included a hyperlink that when clicked triggered a multimedia questionnaire on the TV screen. Panel members returned the completed questionnaires via the Internet and the responses were collated into a database on the host server. The KN-panel consisted of 250,000 members at the end of 2001 (Berrens *et al.* 2001).

Krotki & Dennis (2001) compared a random sample selected from the KN-panel with the US population. The authors observed that the sample was representative of the US population on gender, race, and Hispanic ethnicity. The sample, however, under-represented the elderly and low-income households (see Table 4.4). According to Krotki & Dennis (2001), Knowledge Networks rectified such discrepancies by including additional panel members who matched the required criteria in the sample, thereby, making the sample representative of the target population. Knowledge Networks also regularly updated the panel to match the current US population by age, gender, region, race, ethnicity, and education.

Table 4.4 Comparisons between the KN Panel and the US Population (Krotki & Dennis 2001)

	KN (%)	US (%)
Female	51	51
Black	11	12
Hispanic	09	11
Over 55 years	19	28
Low income (<$25,000)	16	28

Berrens *et al.* (2001) compared the demographic make-up of samples selected from such a proprietary database of willing respondents with a standard probability sample. The authors provided results of comparisons made between a telephone survey that used the RDD method and three Internet-based surveys. Of the three Internet-based surveys, two drew samples from the panel maintained by Harris Interactive (HI) and one from the panel maintained by Knowledge Networks (KN). The telephone survey and the first Internet-based surveys (HI_1) were carried out at the same time in January 2000. The

second Internet-based survey (HI_2) was carried out in July 2000 and the third (KN) in November 2000. All four surveys fielded the same questionnaire.

The four surveys were similar on mean age and percentage of male respondents. The weighted percentages of respondents with at least a college degree were comparable in the three Internet-based surveys (22%, 23%, 21%). This estimate was considerably higher in the telephone survey (41%). Figures (percentages of respondents with at least a college degree) obtained in the Internet-based surveys, however, were comparable to the Census 2000 figure (23%). The raw estimates obtained for this variable in the two HI surveys were over estimations (44% and 46%), whereas the raw estimate obtained in the KN survey (24%) closely matched the Census 2000 figure (23%). The four surveys under estimated the Hispanic (Telephone: 7%; HI_1: 9%; HI_2: 10%; KN 10%) and African-American (Telephone: 8%; HI_1: 12%; HI_2: 12%; KN 11%) population. The figures obtained in the Internet-based surveys were once again closer to those of the Census 2000 (both percentages being 13% in the Census 2000).

Theoretically, the Harris Interactive panel was self-selected; hence it fails to conform to the sampling theory of marketing research. Thus, the reliability of samples drawn from the Harris Interactive panel to resemble the US population may be questioned (Berrens *et al.* 2001). HI was able to rectified discrepancies in representativeness by making the sample match the target population (Krotki & Dennis 2001). What appears to be an obvious disadvantage is the cost of maintaining such a large panel and in most cases is beyond the scope of academic researchers.

The approach advocated by Knowledge Networks was based on standard sampling theory. The approach, however, was dependent on respondents owning or willing to own Web-TVs. The panel also required regular monitoring and updating to match the target population. One main difficulty of adopting this approach in some other parts of the world would be the low adoption of Web-TV. Further, the cost of setting this approach up may not be justifiable when there are established cost-effective probability surveys (mail, face-to-face, and telephone) available to researchers.

Quigley *et al.* (2000) compared response rates obtained for two approaches that combined two survey modes against a conventional mail survey (Treatment One in

Table 4.2). One of these approaches comprised of a mail survey with an added option of completing the survey via the Web (Treatment Two in Table 4.2). The other approach comprised of an Internet-based survey with an added option of completing a paper version of the questionnaire by post (Treatment Three in Table 4.2). The test was built into the 2000 Information Services Survey of the Defence Manpower Data Centre, USA that surveyed military personnel. Respondents were approached via their postal address with a request to participate in the 2000 Information Services Survey. Results of this study are summarised below in Table 4.5.

Table 4.5 Response Rates of the Three Survey Approaches (Quigley *et al.* 2000, p 127)

		Phase 1		Phase 2		Total
	n	**Mode**	**%**	**Mode**	**%**	**%**
Treatment One	7279	Mail	40	-	-	40
Treatment Two	21805	Mail	77	Web	23	42
Treatment Three	7209	Web	73	Mail	27	37

Overall response rates obtained in the three treatments were comparable (40%, 42% and 37%). In Treatment Two (mail survey with Web option), 77% of respondents chose to complete the survey by mail while the remaining 23% completed via the Web. In Treatment Three (Internet-based survey with mail option), 73% of respondents completed the survey via the Web with 27% choosing to do so by mail. The pattern of response rates obtained in Treatment Two and Treatment Three shows that the survey mode that was offered at the start received the highest response rate. The observation made in this study suggests that the approach employed in Treatment Three (Internet-based survey with mail option) could be thought of as being the best of the three because of the efficiency and cost effectiveness brought about by the Internet-based survey.

Schonlau, Fricker, Jr. & Elliott (2001) reported the results of a survey approach that implemented an Internet-based survey and a mail survey, similar to that of Quigley *et*

al.'s (2000). The approach was used in a survey designed to ascertain intentions of high school graduates to enrol in military services. Respondents were contacted by post via their parents' address with a request to participate in a survey via the Internet. A paper version of the questionnaire was included in the reminder letter, which offered non-respondents the option of completing the survey either via the Internet or by post. The survey produced 2583 valid responses (21% response rate), of which 976 were completed via the Internet (38%) and the remaining 1607 by post (62%). The authors attributed the overall low response rate to the fact that the majority of respondents were not contactable via their parents' address. All the same, they reported a saving of about $2000 brought about by the eliminated cost of editing, data entering and questionnaire printing for those who participated via the Internet.

Dillman *et al.* (2001) provided further support for the approach pioneered by Quigley *et al.* (2000). The authors tested five approaches that provided alternative survey modes to non-respondents to participate in the survey (see Table 4.6). In all five approaches response rates increased when non-respondents were offered an alternative survey mode. The best overall response rate was obtained in the approach that offered the mail and phone options (82.8%). In this case, the response rate obtained in the first phase (75%) was reasonably satisfactory, not warranting the second phase (see Table 4.6). Of the three comparable approaches (first three listed in Table 4.6), the increase in response rate brought about by the alternative survey mode was highest in the approach that offered the Internet and phone options (35%). The Internet-based survey used in this approach produced the lowest response rate of all (12.7%). The alternative mode (telephone survey) offered to non-respondents helped this approach to obtain a satisfactory response rate of 47.7%.

Table 4.6 Response Rates of the Mixed Mode Survey Approaches (Dillman *et al.* 2001)

Original Sample size	Phase 1			Phase 2			Total	Rate of increase
		N	%		N*	%	%	%
2000	Internet	253	12.7	Phone	700	44.9	47.7	35
2000	IVR	569	28.5	Phone	438	35.9	50.4	22
2000	Mail	1499	75	Phone	157	31.7	82.8	7.8
1500	Phone	651	43.4	Mail	1094	66.3	80.4	37
1499	Phone	667	44.4	Mail	1094	66.3	80.4	36

*Includes non respondents and refusals

Studies that compared electronic versions with hard copy versions of questionnaires found no difference in the quality of data collected (Chatman 2002; Burr, Levin & Becher 2001; Kreuels 2001; Daly, Thomson & Cross 2000). Findings of these studies provided the rationale for combining data collected using different survey modes into a single dataset.

4.6 Rationale of the Internet-Based Survey Approach

The research carried out for this thesis (discussed in more detail in Chapter Five) aimed to address two issues concerning the Juster Scale. The first was to standardise the contextual requirement of the Juster Scale. To achieve this objective, the Juster Scale was implemented in separate questionnaire versions. In one, the scale was implemented on its own and in the other two the scale was implemented after respondents had the opportunity to search and view contextual information, similar to when making a purchase. Internet technology was employed in the latter two versions to facilitate information search. In one, a Web interface that presented hyperlinks to information on the Internet was used. Respondents allocated to this version were asked to click on the hyperlinks to view information on their computer (Urban *et al.* 1996). In the other, a Web interface that supported a search engine was used. Respondents allocated to this version were asked to enter key words (e.g. price, brand) into the search engine that produced a list of related hyperlinks on their computer (Brucks 1988, 1985).

The second objective was to further improve the method of forecasting mutually exclusive behaviours. The review of literature revealed that respondents tend to be irrational in their assigning of probability scores when asked to do so for a set of mutually exclusive alternatives. Probability scores assigned this way failed to reflect the purchase behaviour of the individual or the sample. Researchers employed a weighting process to rectify the discrepancies caused by respondents' irrational assigning of probability scores. Weighted probability scores helped in the logical interpretation of the sample's behaviour towards the alternative (Fannelly *et al.* 2000; 1999; 1998; Parackal & Brennan 1999; Fannelly *et al.* 1998; Hoek & Gendall 1993). In the review of literature in Chapter Three attention was drawn to a problem with the forecasting accuracy obtained by the weighting process. The accuracy was dependent on forecasts being either over or under estimated. The weighting process appeared to require more testing before recommending it. Suggestion was made to test it against a method that did not require the weighting process.

The literature also reported of a Constant Sum Method that researchers have used for collecting probability data for mutually exclusive alternatives (Reibstein 1978; Alexrod 1964; Metefessel 1947). This method produced satisfactory results and was also successful in collecting probability data for mutually exclusive alternatives that added up to ten. Hamilton-Gibbs *et al.* (1992) developed and tested a scale (referred to as the Constant Sum Scale) to collect probability data for fast moving consumer products based on the principle of the Constant Sum Method. Hoek & Gendall (1997) successfully adapted the Constant Sum Scale to collect voting behaviour of respondents in a multi-party and multi-candidate election. This scale's satisfactory results and the fact that it got respondents to assign scores that added up to ten made it suitable to be used in the testing of the weighting process method. In this thesis's research, an attempt was made to compare the forecast obtained on the weighting process method with that obtained on the Constant Sum Scale. The browser technology offered the opportunity to implement the Constant Sum Scale over the Internet. To undertake the investigation, an electronic version of the Constant Sum Scale was developed and used to collect probability data for mutually exclusive alternatives from one group of respondents. Probability data from another group was collected without controlling the way respondents assigned scores. These scores were weighted before comparing the forecasts with those made on the Constant Sum Scale.

As the Juster Scale was used to collect quantitative data, it was essential that the above issues be addressed in a quantitative survey. The review in Chapter Three emphasised the importance of testing the Juster Scale using probability samples. On reviewing the probability survey approaches implemented via the Internet, the one pioneered by Quigley *et al.* (2000) was seen as being appropriate for the current research. In the surveys implemented for the current research, respondents were approached via their postal address with a request to participate in an Internet-based survey. In the letters mailed out, Internet users were requested to complete the survey on the Internet and non-Internet users were requested to return a post card included in the mail out to receive a paper version of the questionnaire. The latter step was different to that of Quigley *et al.*'s (2000) and was so designed to maximise Internet participation. Quigley *et al.* (2000) included the hard copy with the reminder letters; this could prompt respondent who other wise could complete the survey via the Internet to choose the paper version option as it was readily available in their hand. As for the method adopted in the current research, respondents had to return the card by post and wait to receive the paper version. It was envisaged that this task would discourage Internet users from opting that option and would chose to complete the survey via the Internet.

The comparatively large Internet population in New Zealand allowed the adoption of the Quigley *et al.* (2000) approach to collect data required to achieve the objectives of this thesis. The wide spread of households with Internet access across the country suggests that Internet users are normally distributed or tending to be normally distributed within the general. With 52% of New Zealand householders having access to the Internet (Nielsen//NetRatings 2002), over half of any random sample selected from the general population would be from homes with Internet access. As respondents were part of the randomised selection, the final sample obtained via the Internet had the qualities of a random sample. The approach helped in defining the target population and ensured that a satisfactory cross-section of respondents was present in the sample. It also meant that random errors were naturally removed when measurements were averaged in the sample (mean probability scores in the case of the current research). As an additional precaution to ensure sufficient number of respondents in the treatments, it was decided to contact a comparatively larger number of potential respondents in both the surveys implemented (Vodafone study N=3400; New Zealand study N=3000). The

sample sizes were decided based on the research budget. The approach was used in part (without the hard copy) in the Vodafone survey and in full in the New Zealand survey.

4.7 Chapter Summary

The comparatively large Internet population in New Zealand provided the opportunity to carry out the research for this thesis over the Internet. Analysis of a leading New Zealand Web site (AcessNZ) revealed that a large majority of New Zealand visitors used Microsoft based operating system. Internet Explorer was the most popular Web browser among these visitors. These observations led to the decision of using Microsoft based technology to develop the surveys for this research.

The Active Server Page of Microsoft was chosen as the operating environment in which the surveys were developed. This technology allowed the combining of scripting languages and databases technology for Web applications. Scripting languages such as the Hypertext Mark-up Language (HTML) version 4.01 and ECMAScript were used to develop the survey interfaces.

The review of literature on Internet-based surveys carried out, identified a probability survey approach that could be used on an open population (Quigley *et al.* 2000). The approach comprised of selecting a random sample from an appropriate list of the target population. Potential respondents were contacted via their postal address with a request to participate in an Internet-based survey. The letter offered non-Internet users the option of participating in the survey by filling in a paper version of the questionnaire and returning it by post.

The comparatively large Internet population in New Zealand provided the opportunity to use the Internet in this research. As respondents who participated via the Internet were part of the randomised selection, the final sample also closely resembled the target population. This was important in this research to ensure that the sample included a satisfactory cross-section of respondents. The approach will be explained in detail in the Chapter on Methodology (Chapter Six).

5. CONCEPTUAL FRAMEWORK AND HYPOTHESES

5.1 Introduction

Reservations about intention scales to forecast purchase behaviour accurately led researchers to test probability scales. The pioneering study by Juster (1966) compared an 11-point probability scale with a five-point intention scale and found the former forecast automobile purchases more accurately. Subsequent studies confirmed Juster's results for other durables, services and fast moving products (Day *et al.* 1991; Gan *et al.* 1986). The Juster Scale, as it became popularly known received much of the subsequent research attention. It has been extensively tested for various applications (Garland 2002; Parackal & Brennan 1999; Danenberg & Sharp 1996; 1999; Brennan *et al.* 1995a; 1995b; Brennan & Esslemont 1994a; 1994b; Seymour *et al.* 1994; Hamilton-Gibbs *et al.* 1992; Gendall *et al.* 1991; Day *et al.* 1991; Gan *et al.* 1986; Gabor & Granger 1972; Clawson 1971; Gruber 1970; Heald 1970; Juster 1966) and found to produce satisfactory results.

The studies cited above were successful in meeting their respective objectives. Nevertheless, attention was drawn in the review of literature in Chapter Three to variations in the accuracy of forecasts made on the Juster Scale. The review of literature identified two issues that were believed to cause some of the variations and two development works for the Juster Scale. All four topics require research attention; however, it was not possible to address them all in the research done for this thesis because of time and resource constraints. In this chapter, these topics are prioritised, and the research purpose and objectives restated.

5.2 Research Priorities

5.2.1 Accuracy of the Juster Scale

In the review of literature, two issues were raised as causing some of the variations in the accuracy of forecasts made on the Juster Scale. The first was the sample make-up of Juster Scale studies. In the review of literature in Chapter Three, raw probability data

from three Juster Scale studies (Day *et al.* 1991; Gan *et al.* 1986; Clawson 1971) that tested the scale on similar product categories were examined. In all these studies, accuracy of the scale varied considerably across the product categories. Observations made on the response distributions collected for these product categories suggest that the samples used may be responsible for some of the variations. The review in that chapter concluded by outlining an investigation to find out whether samples influenced the forecasts made on the Juster Scale. The investigation comprised of comparing forecasts for Internet-based services and products made on the Juster Scale in an optimal sample and sub-optimal sample. The section also explained the method to generate these samples (optimal and sub-optimal samples).

The recommended research would empirically explain the sampling requirements for Juster Scale studies. All the same, the theoretical reasoning behind using samples relevant to survey topics (optimal samples) is well established. Apart from seeing whether the theory holds good for Juster Scale studies, there was no pressing need to confirm it again. Hence this issue was not addressed in the current research.

The second issue raised was the context of the Juster Scale. This issue appears to have serious ramifications on the comparability of Juster Scale studies. The contextual literature reviewed revealed the detrimental effect that altering context had on survey response distributions (Schuman *et al.* 1983; Sudman & Bradburn 1982; Schuman *et al.* 1981; Schuman & Presser 1981; Duncan & Schuman 1980). Juster Scale investigators, in general, expected the scale to collect purchase probability data in the context of making a purchase. The question accompanying the scale was formulated to do this. Nevertheless, a number of factors were discussed in Chapter Three that appeared to alter the context of the Juster Scale. If these factors did alter the context of the Juster Scale then results would be incomparable. In which case, the Juster Scale would require further testing to generate sufficient comparable studies so that its reliability could be established.

While there was agreement among Juster Scale investigators on the adverse effect of context on forecasts, very limited research has been directed at the issue. Given the seriousness of this issue, the current research aimed to verify, through empirical

examination, whether the Juster Scale required additional input to collect purchase probability data in the context of making a purchase. On establishing this, further investigation exploring whether the context is maintained in the questionnaire, and while administering the scale, could be carried out.

5.2.2 Scale Development

The review of literature in Chapter Two uncovered two development works for the Juster Scale. The first was to isolate a set of scale descriptors that produces the best forecasts. To achieve this, suggestion was made to compare probability scales with different scale descriptors for forecasting accuracy. The review of literature in Chapter Two identified four probability scales (Pickering & Isherwood 1974; Stapel 1968; Byrnes 1964; Ferber & Piskie 1965) that used different sets of scale descriptors. The suggestion was to compare these four probability scales for forecasting accuracy by implementing them simultaneously in separate treatments.

Examining the studies (Pickering & Isherwood 1974; Stapel 1968; Byrnes 1964; Ferber & Piskie 1965) done on the four probability scales revealed that they produced satisfactory forecasts. All four scales served their purpose well and no alarming issues were found to undermine any one of them. While it would be good to establish the relative forecasting accuracy of these scales, there was no pressing urgency to carry out such a comparative research. Hence this research was not pursued further in the current thesis.

An application of the Juster Scale that has considerable practical use to marketers and researchers is in the forecasting of mutually exclusive behaviour. Probability data collected for such behaviour could be used to forecast market shares, switching behaviours between competing products or behaviours and election results. The challenging part of this Juster Scale application was to get respondents to give probability scores in relation to the available alternatives. Literature examined on this topic revealed that respondents tended to give probability scores treating alternatives as being independent. Consequently, probability scores failed to reflect the purchase behaviour of the sample logically. To rectify the discrepancies caused, a weighting

process was applied on the raw probability scores (Flannelly *et al.* 2000, 1999; Parackal & Brennan 1999; Seymour *et al.* 1994; Hoek & Gendall 1993). This helped investigators to explain the purchase behaviour for the sample logically. In the review of literature, attention was drawn to a problem with this method, warranting more research. Suggestion was made to test the method against one that did not require the weighting process.

The review of literature also covered a Constant Sum Method used for collecting probability data (Reibstein 1978; Alexrod 1964; Metefessel 1947). This method forced respondents to give probability scores that added up to a constant number (usually 100 or 10). Investigators of the Juster Scale developed and tested a Constant Sum Scale based on this method to collect purchase probability data for fast moving consumer products (Brennan *et al.* 1995b; Seymour *et al.* 1994; Hamilton-Gibbs *et al.* 1992). The scale was implemented in a face-to-face setting, in which interviewers got respondents to distribute ten tokens across a set of mutually exclusive alternatives. Investigators found the scale easy to administer and the task required of respondents was simple to perform (Hamilton-Gibbs *et al.* 1992). Hoek & Gendall (1997) used the Constant Sum Scale to collect voting probabilities. Forecasts based on the Constant Sum Scale were more accurate than those based on the traditional forced-choice questioning method.

The two methods mentioned above have produced satisfactory results in separate studies. There was, however, no study in the literature that investigated the relative accuracy of the two methods. The fact that the Constant Sum Scale collected probability data for mutually exclusive alternatives that added up to ten made it a suitable scale against which the weighting process could be compared. In the current research the two methods were compared to establish the one that was most suitable to collect probability data for mutually exclusive alternatives. This objective also aimed at resolving the problem investigators faced when collecting probability data of mutually exclusive items.

5.3 Research Objectives and Hypotheses

To address the two topics chosen in the preceding sections, the following objectives were set:

- To investigate whether the Juster Scale requires additional contextual information to collect purchase probability data in a purchasing context.

- To investigate whether forecasts of mutually exclusive behaviours based on probability scores that did not add up to ten (weighted probability scores) are more accurate than those based on scores that added up to ten (Constant Sum Scale).

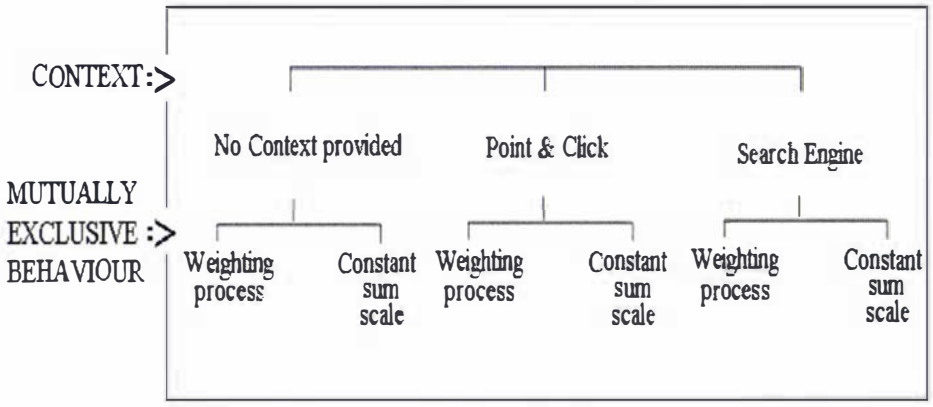
To investigate the two issues statistically, the above objectives were formulated into the following hypotheses:

- H1: Mean probability scores obtained in the treatment that implemented the Juster Scale on its own without providing contextual information will be lower than those obtained in treatments that implemented the scale after contextual information was provided (i.e. at least one of the means were different).

- H2: Forecasting accuracy of mutually exclusive items based on the weighted probability scores will be less accurate than those made on the Constant Sum Scale ($\hat{x}_1 \neq \hat{x}_2$).

To test the above hypotheses, probability data were collected using separate questionnaire versions. Mean probability scores obtained in the versions were compared for statistical differences. The comparisons used to test the above hypotheses are shown in Figure 5.1.

Figure 5.1 Research Design



Two Internet-based surveys were implemented in this research to collect data to make the necessary comparisons. The first survey was implemented on Vodafone clients and secured data to achieve the first objective (Context). This survey will be referred to as "Vodafone survey" for the sake of simplicity. The second survey was implemented on the national population; this survey secured data to make the comparisons to achieve the first (Context) and the second (Mutually exclusive behaviour) objectives. This survey will be referred to as "New Zealand survey" for the sake of simplicity. The two surveys provided comparisons of the treatments on different samples. In the following chapter (Chapter Six), the survey approach, treatments, test products and scripts employed in these two surveys are explained. Following which, the results of the two surveys are reported.

6. METHODOLOGY

6.1 Introduction

In the previous chapter, two issues were chosen to be address in the research done for this thesis. The two issues were, one the contextual requirement of the Juster Scale and the other, the problem Juster Scale investigators encountered while collecting purchase probability data for mutually exclusive behaviour. Data required for testing the hypotheses outlined in the previous chapter were collected by implementing two Internet-based surveys. The first survey was implemented on Vodafone clients and the data collected was used to achieve the first objective (Context). This survey will be referred to as "Vodafone survey" for the sake of simplicity. The second survey was implemented on the national population; this survey secured data to make the comparisons to achieve the first (Context) and the second (Mutually exclusive behaviour) objectives. This survey will be referred to as "New Zealand survey" for the sake of simplicity. The two surveys provided comparisons of the treatments on different samples. In the following sections, the survey approach, treatments, test products and scripts employed in these two surveys are explained.

6.2 Survey Approach

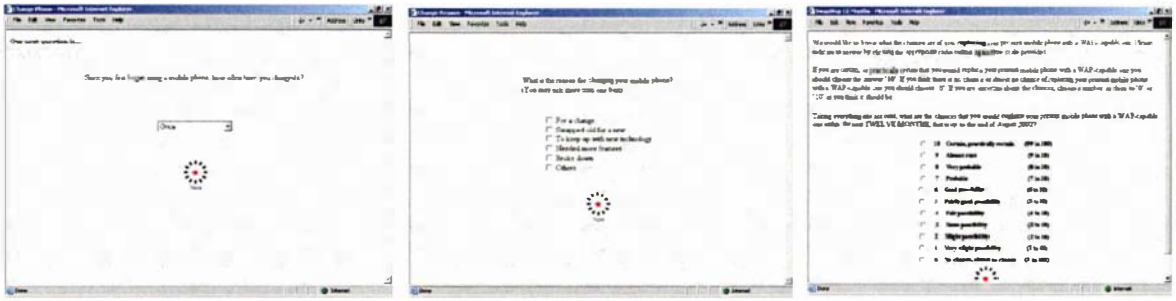
6.2.1 Vodafone Survey

The Vodafone survey was implemented on the clientele of Vodafone New Zealand. The company provided the contact details of 3400 individuals randomly selected from its client list. The sample size was decided by the managers of the company and was based on what they were prepared to provide for this research. Four respondents were flagged as using WAP-capable mobile phones in the sample. They were removed from the sample; consequently the original sample size was 3396. After removing incomplete addresses, 3388 respondents were contacted via their postal address with a request to participate in an Internet-based survey about WAP-capable mobile phones. The cover letter provided each

respondent a unique login name, access code and the URL of the survey site. Respondents were required to use the login name and access code to gain access into the survey site.

Restrictions placed by Vodafone management on this survey permitted the implementation of a short questionnaire. Hence, the questionnaire included some questions placed by Vodafone management and those pertaining to the first objective. Company policy also restricted the number of contacts with respondents to one. Hence, no effort was made to increase the response rate (no reminder letters or incentives were used). This also meant that the survey approach suggested by Quigley *et al.* (2000) could not be implemented in its entirety in this survey. The survey was kept open for 30 days.

Figure 6.1 Web Format of the Questionnaire



The questionnaire was formatted into separate Web pages that presented questions one by one on respondent's computer (see Figure 6.1). A "Next" button was provided on the Web page that enabled respondents to move from one question to the next. When respondents hit the "Next" button, scripts that were in place transferred the answer(s) from the Web page into corresponding field(s) in an Access database table on the server.

6.2.2 New Zealand Survey

The New Zealand survey was implemented on 3000 respondents randomly selected from the New Zealand electoral roll. Respondents were contacted via their postal address with a request to participate in an Internet-based survey. As in the Vodafone survey, each respondent was provided with a login name and an access code. These and the URL of the

survey site were communicated to respondents in the letters mailed out. The format of the questionnaire was the same as that shown in Figure 6.1.

Two reminder letters were used to increase the response rate. The initial contact letters were mailed out on September 12th, 2001 (September 11th in the Northern Hemisphere). To distance respondents from the events of September 11th, the first reminder letters were sent after a gap of one month. The second reminder letter was sent three weeks after the first one was sent.

In this survey, Quigley *et al.*'s (2000) approach was implemented in its entirety. The aim of the approach was to maximise Internet participation and get non-Internet users to complete a paper version of the questionnaire. With this in view, non-Internet users were required to ask for the paper version. To facilitate this, a reply paid request card was included with the letters. On receipt of the request card, a paper version of the questionnaire was mailed out to respondents. Comparisons required to achieve the objectives were made using the data collected via the Internet. Data collected using the paper version allowed the monitoring of non-response bias arising from the non-inclusion of non-Internet users in the analyses.

6.3 Treatments

6.3.1 Context of the Juster Scale

The first objective was achieved by comparing the mean probability scores obtained in three separate treatments. In the first treatment, the Juster Scale was implemented on its own. This treatment is called "Standard" in this thesis. In the other two treatments, the Juster Scale was implemented after respondents had viewed contextual information about the test product. In one of these, the approach pioneered by Urban *et al.* (1997, 1996) was implemented after adapting it to suit the Internet and the test product used. Information items were arranged as hyperlinks on the computer screen. Respondents were asked to use their mouse and click on the hyperlinks to view information. This treatment is called "Point & Click" in this thesis. In the other treatment, contextual information was provided via a

search engine (Brucks 1988; 1985). This latter treatment was included to find out if the order of presenting information had any bearing on the forecast made on the Juster Scale. This treatment is called "Search Engine" in this thesis.

All respondents in the Vodafone survey were mobile phone users. They were asked to indicate probability scores on the Juster Scale to replace their present mobile phones with WAP-capable ones. The following question accompanied the Juster Scale to collect twelve and six months-probability data:

- Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next **TWELVE MONTHS**, that is up to the end of < >? (see Figure 6.2)

- Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next **SIX MONTHS**, that is up to the end of < >? (see Figure 6.3)

Figure 6.2 Web Pages Showing the Juster Scale and the Twelve Months-Probability Question

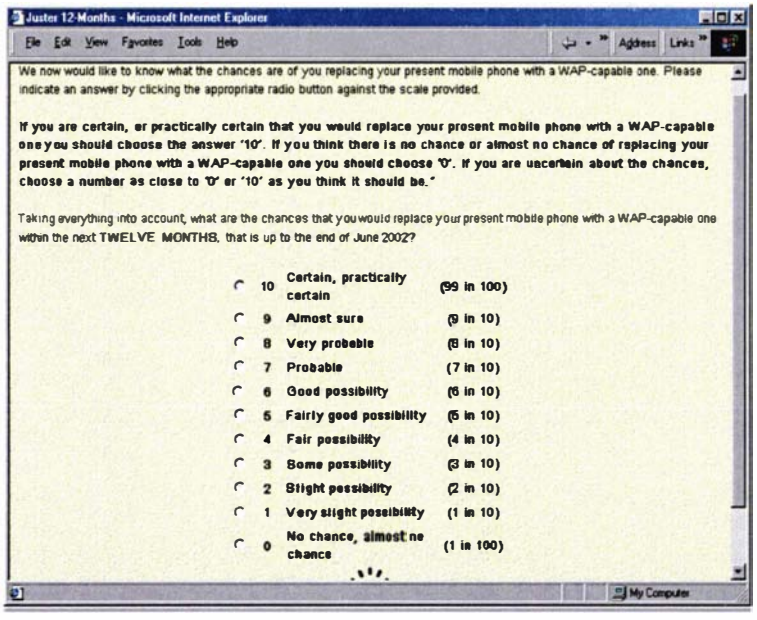
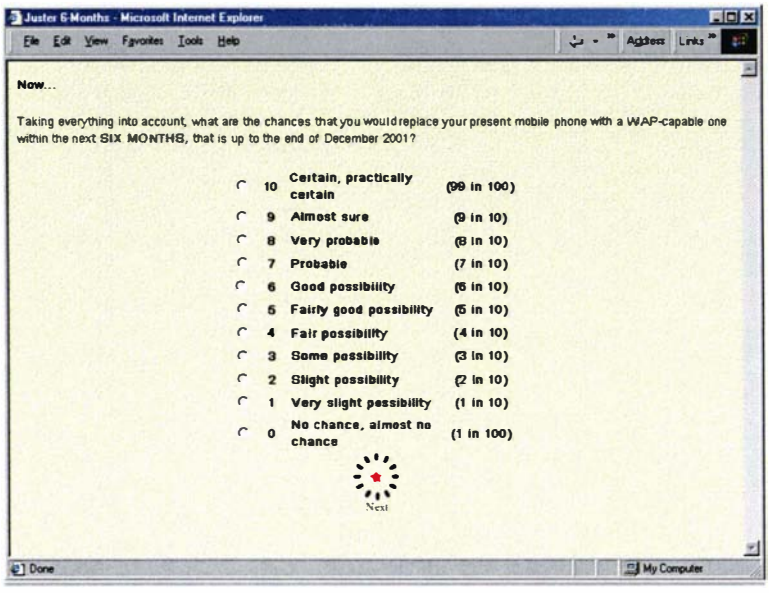


Figure 6.3 Web Pages Showing the Juster Scale and the Six Months-Probability Question



The New Zealand survey sample included respondents who were mobile phone users and non-mobile phone users, providing separate groups to compare the treatments. Mobile phone users were asked the same probability questions as shown above (see Figure 6.2 and 6.3). Non-mobile phone users were asked to indicate probability scores on the Juster Scale to purchase WAP-capable mobile phones. The wordings of the questions asked to the latter group were as follows:

- Taking everything into account, what are the chances that you would purchase a WAP-capable mobile phone within the next **TWELVE MONTHS**, that is up to the end of < >?
- Taking everything into account, what are the chances that you would purchase a WAP-capable mobile phone within the next **SIX MONTHS**, that is up to the end of < >?

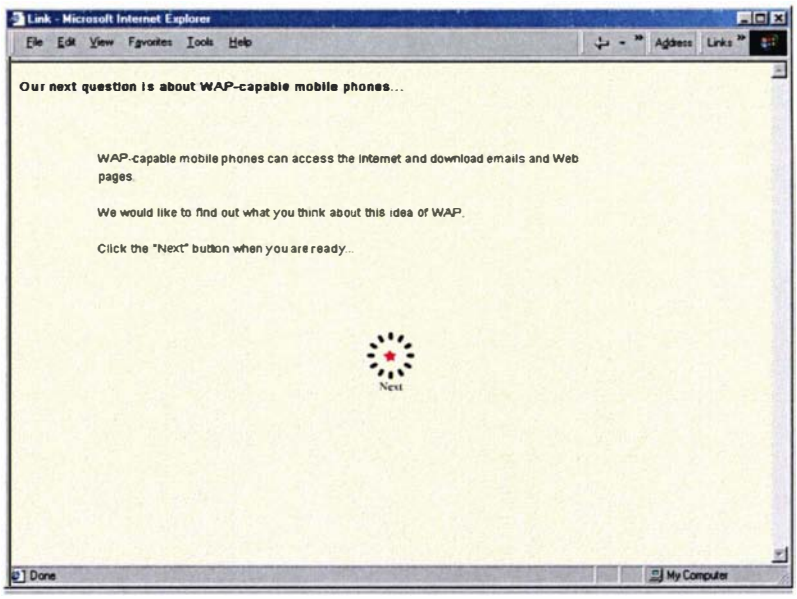
In the following section the treatments are explained with graphic images of the interfaces used.

Standard

Respondents allocated to this treatment were first shown a Web page (see Figure 6.4) that provided a simple description of WAP capable mobile phones before presenting the Juster Scale. Following were the wording of the description used in this treatment:

> *"WAP-capable mobile phones can access the Internet and download emails and Web pages. We would like to find out what you think about this idea of WAP. Click the "Next" button when you are ready..."*

Figure 6.4 Web Pages Showing the Explanation of WAP-Capable Mobile Phones in the Control Treatment



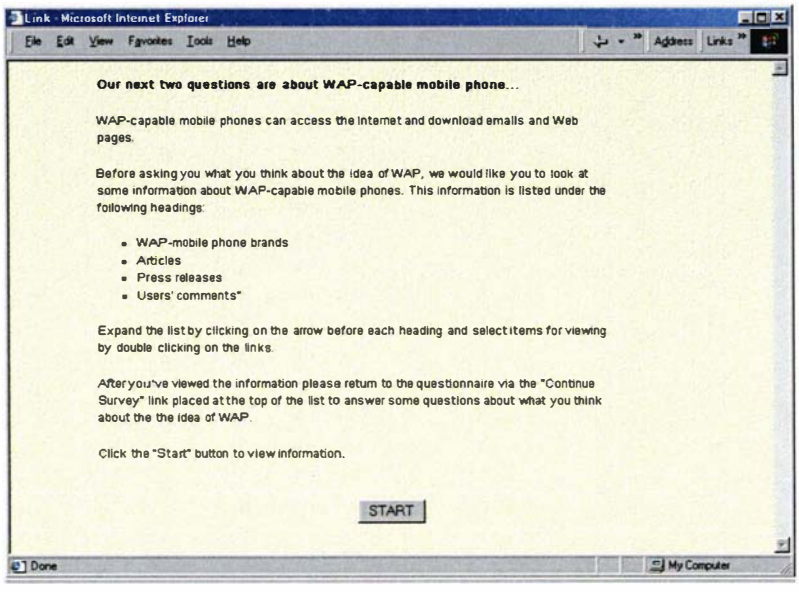
When respondents clicked the "Next" button on the above page, the Juster Scale and its accompanying questions (see Figures 6.2 and 6.3) were presented on their computer.

Point & Click

Respondents allocated to the Point & Click treatment were asked to examine information about WAP-capable mobile phones before indicating their probability scores on the Juster

Scale. Respondents were first shown a Web page with instructions before leading them to the interface that allowed information search (see Figure 6.5).

Figure 6.5 Web Pages Showing the Instructions for Viewing Information in the Point & Click Treatment



When respondents clicked on the "Start" button on the above Web page, they were shown the interface that permitted them to select and view information about WAP-capable mobile phones (see Figure 6.6). Information about WAP-capable mobile phones was organised under four titles (WAP-capable mobile phones, Articles, Press Releases, Users' Comments), using an expandable and collapsible menu. Following were the items listed under the four titles:

- **WAP-capable mobile phones**
    1. Ericsson R380
    2. Philips az@lis268
    3. Philips Xenium 9@9
    4. Philips Fisio
    5. Philips Ozeo 8@8
    6. Motorola Timeport
    7. Nokia 6210

8. Nokia 7110
9. Nokia 9110

- **Articles**

1. "A Billion Internet-Enabled Mobile Phones" by Sam Taylor and Andrew Starling, Web Developers Journal.
2. "Sun Java ™ Technology Powered Mobile Phone to Arrive in Europe" by Maria Villarino and Burson-Marsteller, Geneva, Switzerland, Java.Sun.Com
3. "Lion Nathan finishes WAP trial Mail system package most successful application" by Andrea Malcolm, Auckland, Computer World.
4. "NZA Gold 2000 has online, WAP support" by Rob Clarke, PC World.
5. "Nokia's 7110 a review" by Stephen Ballantyne, PC World.
6. "Surfing on the phone" by PC World Staff.
7. "Visiting WAP guru sketches new landscape" by Russell Brown, Auckland, Computer World.

- **Press Releases**

1. "Leading Mobile Portal Selects Pinpoint to Provide Wireless Web Directory Engine for Lycos Anywhere ™", by Waltham, M. & Durham, N.C., 27th September 2000.
2. "Nokia and Supedo to develop mobile gaming content for WAP enabled phones", 16th May 2001.
3. "Ericsson and Vodafone launch new mobile application", 20th June 2001.
4. "Ericsson launches new, enhanced WAP gateway for business users", 21st March 2001.

- **Users' Comments**

1. Comments featured by Ericsson.

Figure 6.6 Web Page Showing the Information Headings using a Collapsible Menu

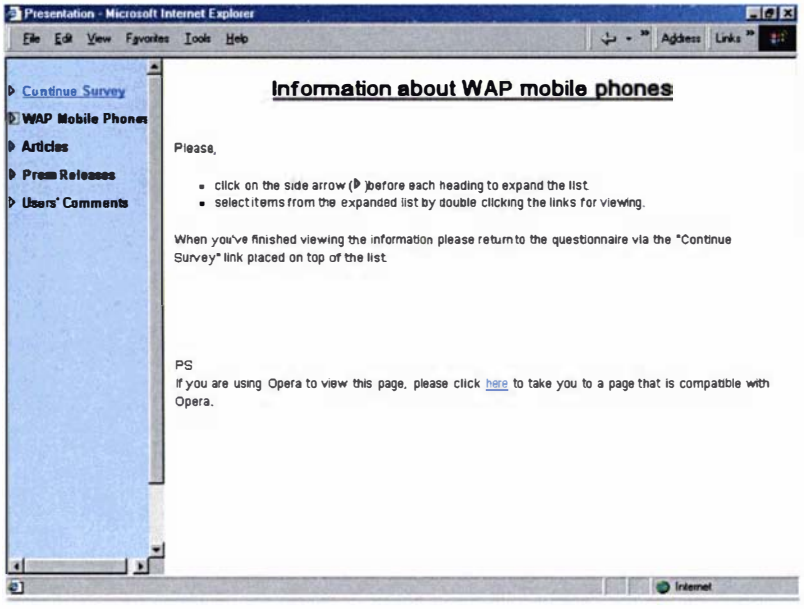
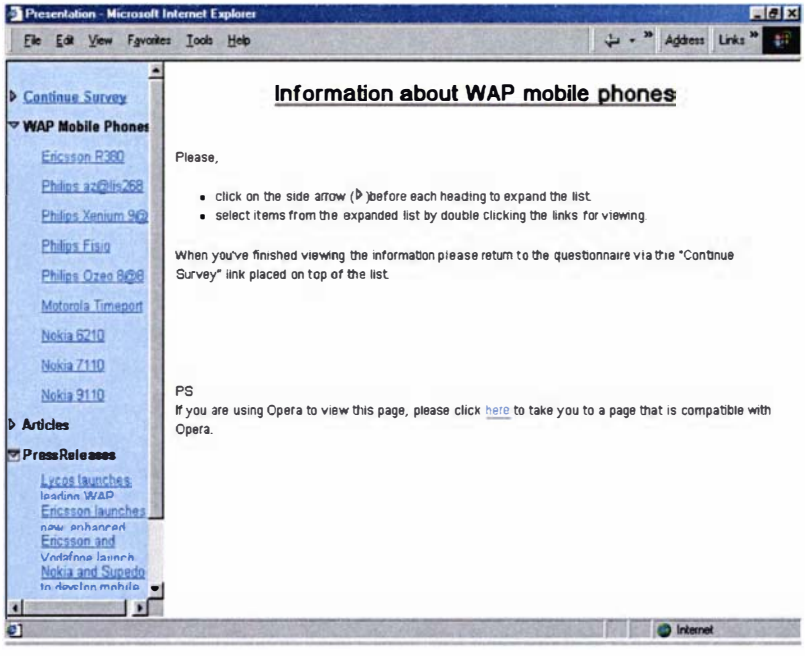
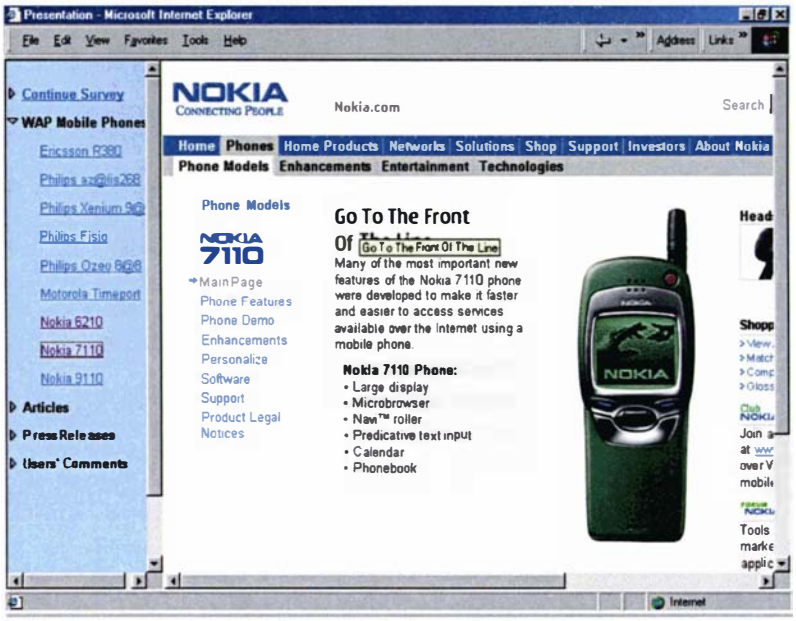


Figure 6.7 Web Page Showing the Menu Expanded



Respondents used their mouse and clicked on the arrows in front of the titles to expand the menu and choose items for viewing (See Figure 6.7). Items were listed as hyperlinks to external Web sites. When a hyperlink was activated, the corresponding Web page loaded in the main frame (see Figure 6.8).

Figure 6.8 Web Pages Showing a Selected Information Item

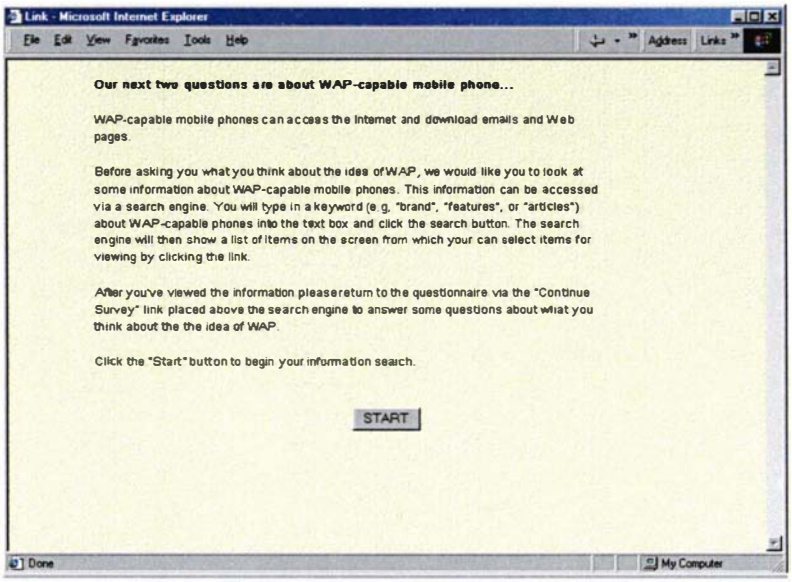


The frame feature of HTML 4.1 held respondents at the survey site while they viewed contents of Web pages of different companies and organisations associated with WAP-capable mobile phone. This feature kept the menu in the grey shaded window visible while respondents viewed the contents of the Web page in the main frame. It allowed respondents to toggle between the frames to select other items for viewing. On completion of searching and viewing information, they were asked to return to the questionnaire via the "Continue Survey" link placed at the top of the menu (see Figures 6.6, 6.7, 6.8). Respondents were then presented with the Juster Scale to indicate their twelve (Figure 6.2) and six (Figure 6.3) months- purchase probability for WAP-capable mobile phones.

Search Engine

Respondents allocated to the Search Engine treatment were provided information about WAP-capable mobile phones via a search engine. To help use the search engine and return to the questionnaire, respondents were first shown a Web page with instructions about that interface (see Figure 6.9).

Figure 6.9 Web Page Showing the Instructions to Use the Search Engine to View Information in the Search Engine Treatment



When respondents clicked on the "Start" button on the above page, they were shown the interface with the search engine (see Figure 6.10). Respondents were asked to enter a search word relating to the information item (e.g. Brands) in the text box and click the search button. The search engine produced a list of items that matched the search word entered (see Figure 6.11). These items were hyperlinked to external Web sites. On activation of the hyperlink, the corresponding Web page loaded in the main frame of the interface (see Figure 6.12).

Figure 6.10 Web Page Showing the Search Engine

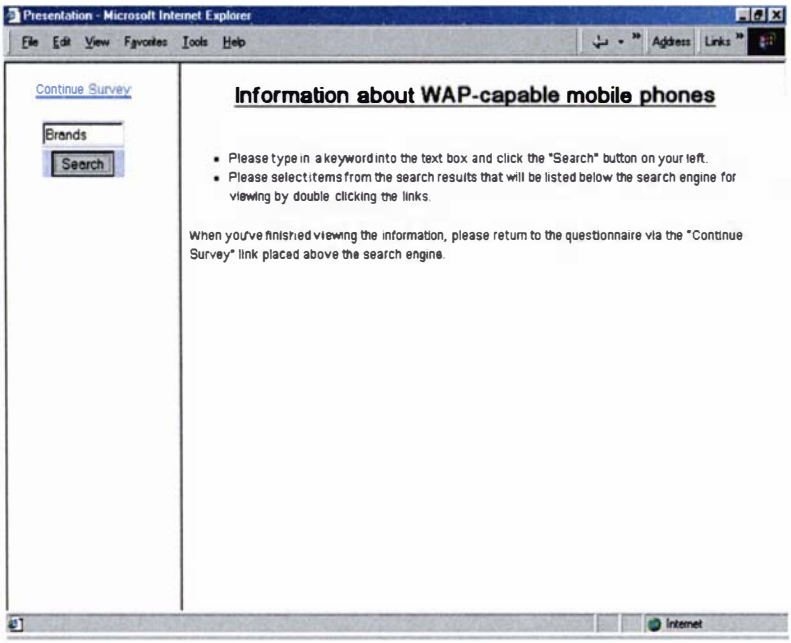


Figure 6.11 Web Page Showing the Search Result

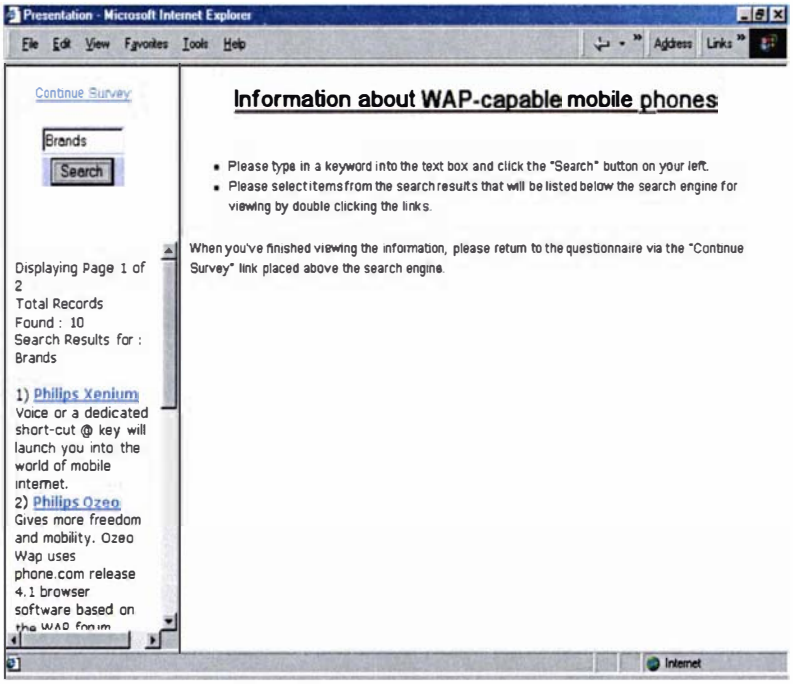
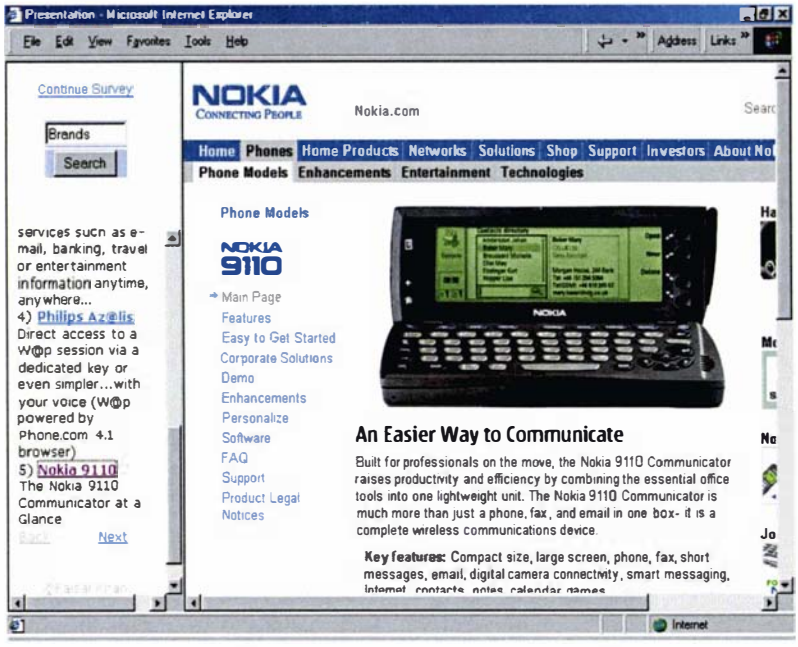


Figure 6.12 Web Page Showing a Selected Information Item



This interface also used the frame feature of HTML 4.1 to hold respondents at the survey site while viewing information on the Web pages of different companies manufacturing WAP-capable mobile phones. The frame feature kept the search engine, search results and the hyperlinked route to the questionnaire visible while respondents viewed the selected Web page in the main frame. Respondents could toggle between the frames to select other items from the search results or perform new searches. When they had finished viewing information, they returned to the questionnaire via the "Continue Survey" link placed over the search engine. On returning to the questionnaire, respondents were presented with the Juster Scale and the twelve and six months-probability questions (Figures 6.2 and 6.3).

6.3.2 Mutually Exclusive Behaviours

Data for the second objective were collected by the New Zealand survey. As the sample for this survey was drawn from the electoral roll, it included respondents who used mobile phones and those who did not use mobile phones. Among mobile phone users, there were those who subscribed to Vodafone and those who subscribed to Telecom. Thus, the sample

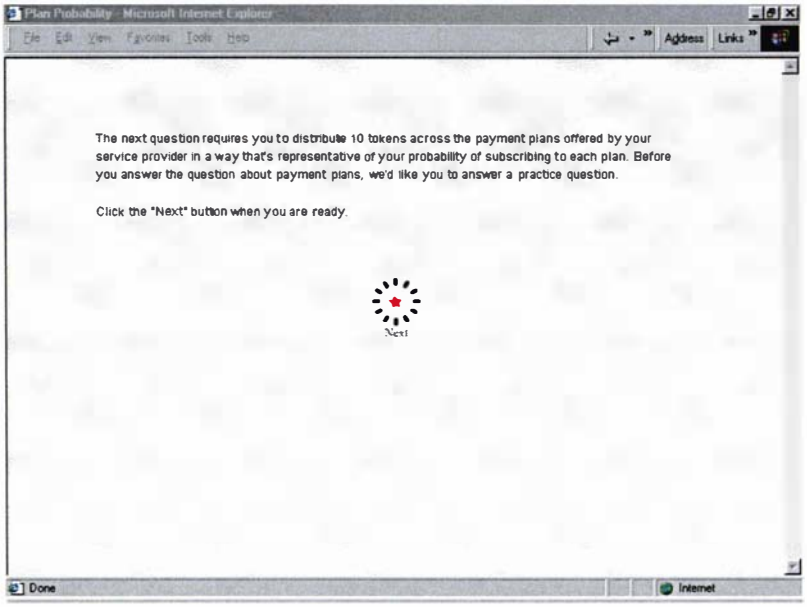
had three separate groups, permitting comparisons to be made on three separate sets of mutually exclusive items, each specific to the groups.

To achieve the second objective, an electronic version of the Constant Sum Scale was developed in EMACscript. This scale was implemented in one of the two treatments to collect probability scores for mutually exclusive items that added up to ten. In the alternative treatment, respondents were asked to indicate their probability scores to each mutually exclusive item separately. In this treatment, scores given to the alternatives were not controlled to add up to ten. Mean probability scores obtained in the two treatments were compared for statistical differences. In the following sections these treatments are explained with graphic images of the interfaces used.

Constant Sum Scale

The electronic Constant Sum Scale followed the same format (rectangular grid and ten tokens) that the earlier investigators have used (Hamilton-Gibbs *et al.* 1992). The tokens were arranged in a row above the grid. Alternatives were listed in rows with ten boxes to accommodate a maximum of ten tokens against them. Respondents allocated to this treatment used this scale to indicate their probabilities to subscribe to payment plans offered by their respective providers. Before being shown the actual question, respondents were asked to answer a practice question to familiarise themselves with the Constant Sum Scale. The sequence in which the practice and actual question were presented to respondents is explained below. This was kept constant for all three groups.

Figure 6.13 Web Page Showing the Instructions for Answering the Practice Question using the Constant Sum Scale

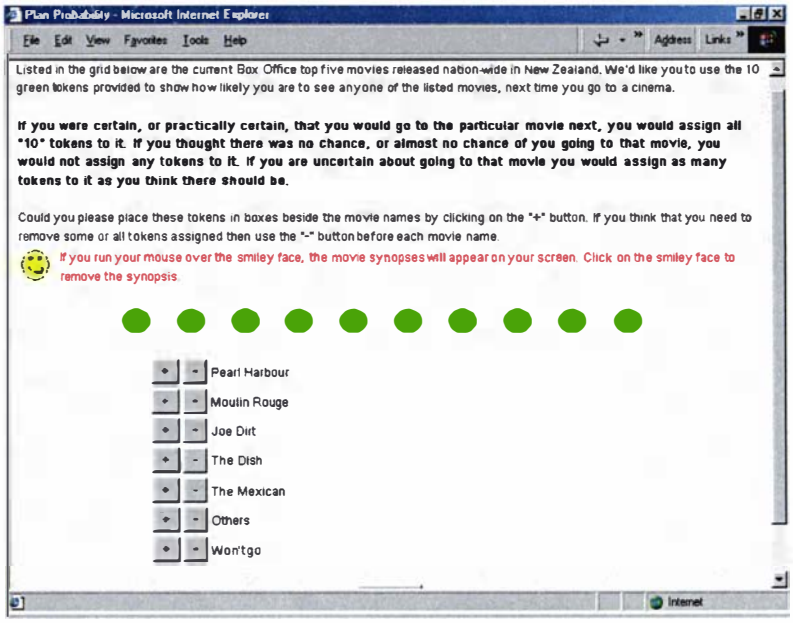


Respondents were first shown the above Web page that explained what was required of them (see Figure 6.13). When respondents clicked the next button on this Web page (Figure 6.13), the Web page with the practice question loaded on their computers. The practice question was about the top five box office movies (Pearl Harbour, Moulin Rouge, Joe Dirt, the Dish, and the Mexican) released nation-wide in New Zealand at the time the two surveys were implemented. To make the list mutually exclusive, two additional options were included, namely, "Others" and "Won't go". The Web page provided an explanation of the topic and what was required of respondents (see Figure 6.14). Instructions used for this purpose were as follows:

> *"Listed in the grid below are the current Box Office top five movies released nation-wide in New Zealand. We'd like you to use the 10 green tokens provided to show how likely you are to see any one of the listed movies, next time you go to a cinema.*
>
> *If you were certain, or practically certain that you would go to the particular movie next, you would assign all "10" tokens to it. If you thought there was no chance, or almost no chance of you going to that movie, you would not assign any tokens to it.*

If you are uncertain about going to that movie you would assign as many tokens to it as you think there should be."

Figure 6.14 Web Page Showing the Practice Question



The Web page also provided the synopses of the movies listed. To read the synopses, respondents were asked to run their mouse over the smiley face (instruction included beside the smiley face in Figure 6.14). This action opened a layer over the Web page, displaying the movie synopses (see Figure 6.15). To remove the layer, respondents had to click on the smiley face using their mouse (instruction included besides the smiley face in Figure 6.14)

The "plus" and "minus" buttons before the movie titles were provided for assigning the tokens to the movies. Respondents clicked on the "plus" button before a movie title to assign tokens to that movie. If they wanted to reallocate the tokens, they used the "minus" button, which removed the token(s) that then could be reassigned. Figure 6.16 shows the Web page with all ten tokens assigned to the movies.

Figure 6.15 Web Page Showing the Movie Synopses

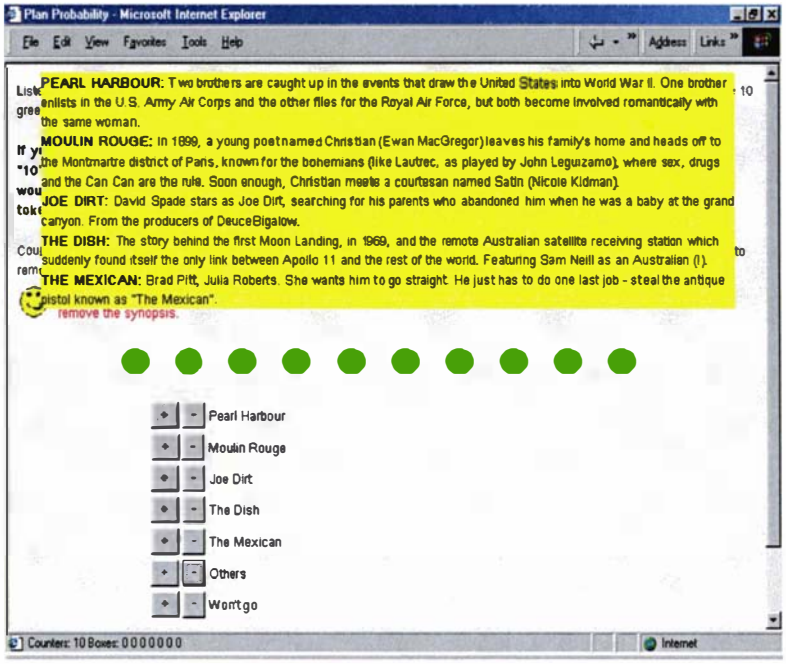
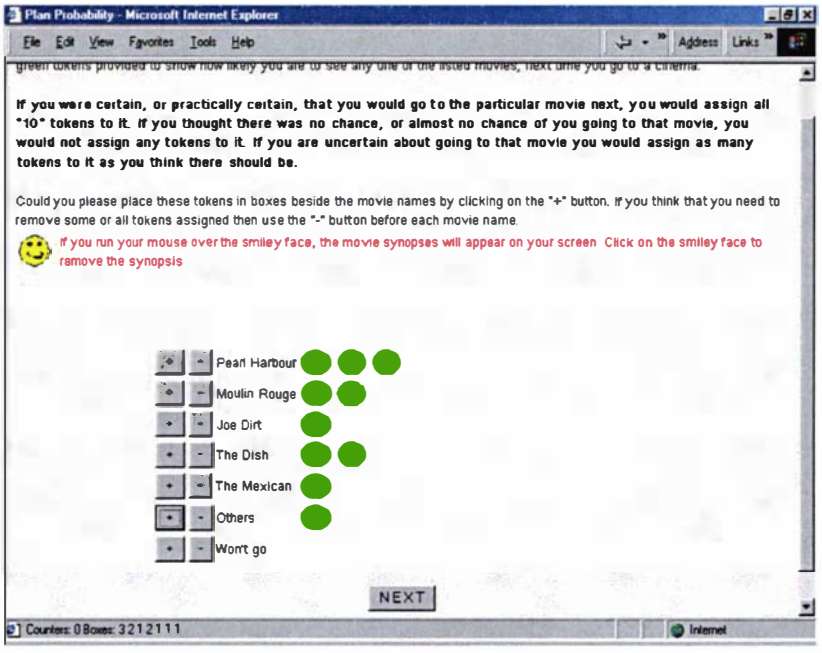


Figure 6.16 Web Page Showing the Tokens Distributed to the Movie Selections



The mutually exclusive nature of the movie selections meant that each individual had some probabilities to see the movies listed. Respondents conveyed the probabilities by assigning the tokens to the movie options. They could either assign all ten tokens to one movie option

or distribute them in some way across the seven movie options. The number of tokens assigned to each option was treated as the probability for choosing that option. Options that had no tokens assigned were given a zero probability. The probabilities assigned to the options added up to ten. After assigning all ten tokens satisfactorily, respondents clicked on the "Next" button on the Web page to move to the next question. This action transferred the number of tokens assigned to each movie option via a form to corresponding fields in the database on the server.

To comply with the ethical requirement for survey completion it was necessary to give respondents the options of answering and not answering the question. To facilitate this, respondents were permitted to proceed to the next question without performing the task required by clicking on the next button. In such instances, scripts used for entering the responses in the database entered zeros against the options; following this, the Web page with the next question was forwarded. Respondents who did this were excluded from the analysis performed. If respondents chose to answer the question, the client side scripts on the Web page held them on the page until all ten tokens were distributed.

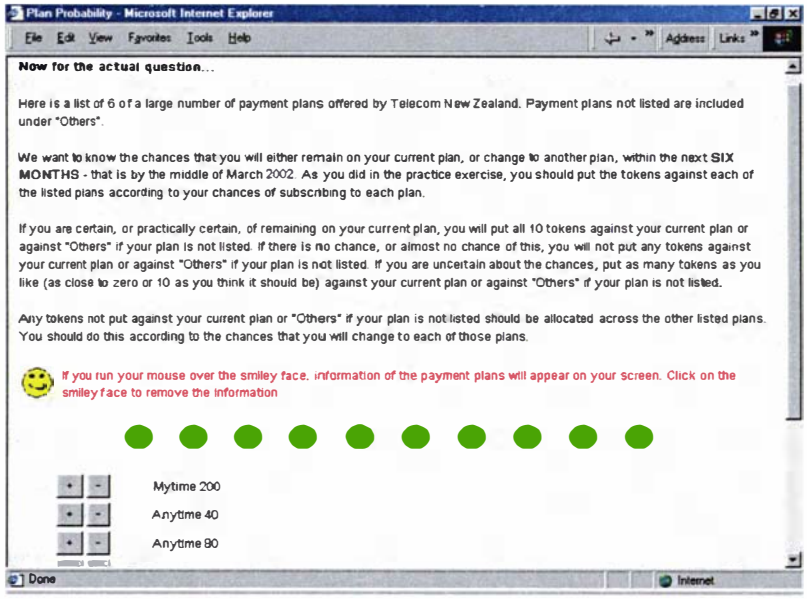
The practice question was followed by the actual question (see Figure 6.17). The Web page that presented the actual question was made in the same format as the one that presented the practice question. The following instructions were included to convey the task required of respondents to answer the actual question:

> *"We want to know the chances that you will either remain on your current plan, or change to another plan, within the next **SIX MONTHS** - that is by the middle of March 2002. As you did in the practice exercise, you should put the tokens against each of the listed plans according to your chances of subscribing to each plan.*
>
> *If you are certain, or practically certain, of remaining on your current plan, you will put all 10 tokens against your current plan or against "Others" if your plan is not listed. If there is no chance, or almost no chance of this, you will not put any tokens against your current plan or against "Others" if your plan is not listed. If you are uncertain about the chances, put as many tokens as you like (as close to zero or*

10 as you think it should be) against your current plan or against "Others" if your plan is not listed.

Any tokens not put against your current plan or "Others" if your plan is not listed should be allocated across the other listed plans. You should do this according to the chances that you will change to each of those plans."

Figure 6.17 Web Page Showing the Actual Question and the Constant Sum Scale



As was the case with the practice question, respondents could view the features of the payment plans by running their mouse over the smiley face (see Figure 6.18). To remove the layer that showed the payment plan features, respondents had to click on the smiley face. Respondents used the "plus" and "minus" buttons to assign the tokens to the payment plans. The number of tokens assigned to each payment plan was treated as the probability to subscribe to that plan. When the "Next" button was clicked, the numbers were transferred via a form to corresponding fields in the database on the server. Respondents who chose not to answer the question were allowed to move on to the next question by clicking on the next button.

Figure 6.18 Web Page Showing the Features of the Payment Plans

Weighted-scores

Respondents allocated to the Weighted-scores treatment were asked to enter a number between "0" and "10" in the text boxes against the payment plans to indicate their probabilities to subscribe to them (see Figure 6.19). The following instructions were included to convey the task required of respondents:

> *"We want to know the chances that you will either remain on your current plan, or change to another plan, within the next **SIX MONTHS** - that is by the middle of March 2002. For each plan, please enter a number between "0" and "10" in the corresponding text box that represents your chances out of "10" for subscribing to that plan.*
>
> *If you are certain, or practically certain, of remaining on your current plan, you will put all 10 tokens against your current plan or against "Others" if your plan is not listed. If there is no chance, or almost no chance of this, you will not put any tokens against your current plan or against "Others" if your plan is not listed. If you are uncertain about the chances, put as many tokens as you like (as close to zero or*

10 as you think it should be) against your current plan or against "Others" if your plan is not listed."

Figure 6.19 Web Page Showing the Conventional Approach used in the Weighted-Scores Treatment

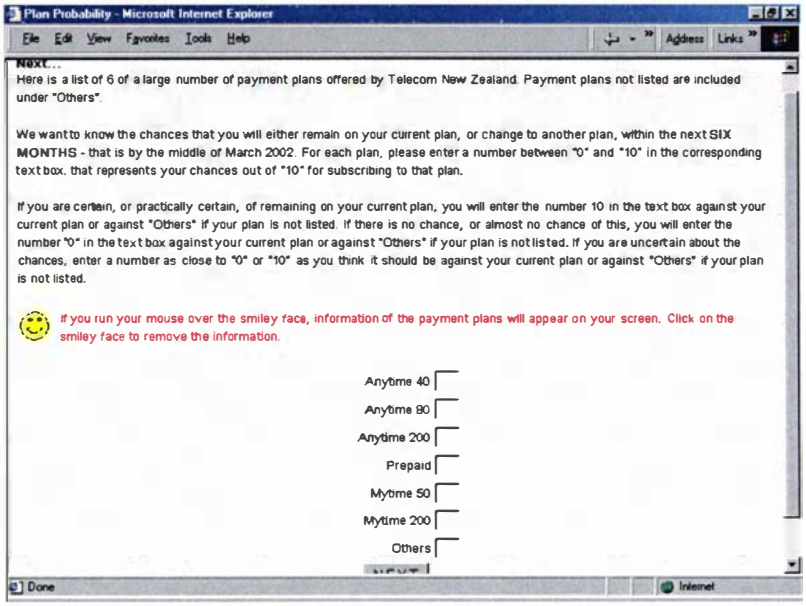
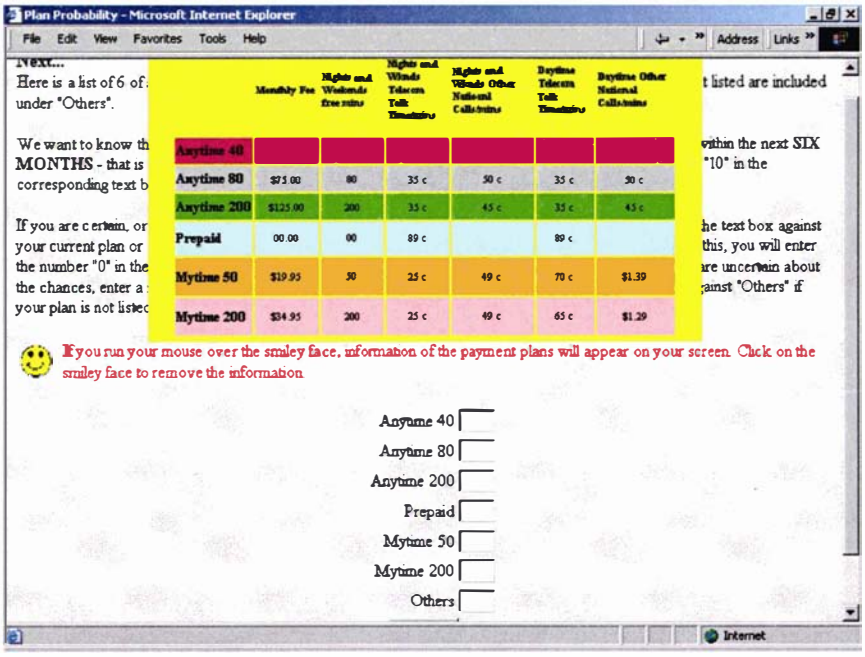


Figure 6.20 Web Page Showing the Features of the Payment Plans



Apart from the way respondents registered their probability scores to the payment plans all other aspects were kept the same as in the previous treatment. When the "Next" button was clicked, the numbers entered in the text boxes against each payment plan were transferred via a form to corresponding fields in the database on the server. Respondents who chose not to answer the question were allowed to move to the next question by clicking on the "Next" button. The scripts entered zeros to all options in such cases and were eventually removed from the analysis.

Respondents in this treatment were not required to give probability scores to all the payment plans. Consequently, there were instances when respondents entered numbers for certain payment plans only. Payment plans for which numbers were not entered were treated as not being included in the consideration set of those respondents. The probability for such respondents to subscribe to those payment plans obviously was zero out of ten. To reflect the mutually exclusive nature of the payment plans in the probability data collected, a transformation was performed that converted the missing entries to zeros.

In this treatment there was no control placed over how respondents distributed their probability scores. As a result, there were respondents whose probability scores across payment plans did not add up to ten. Such scores failed to reflect the subscription behaviour of the respondents and the mean probability scores also failed to explain the behaviour of the sample towards the payment plans. To make the individual probability scores and the mean probability scores logically reflect the subscription behaviour, the raw probability scores were subjected to a weighting process.

The weighting process (see Figure 6.21) was implemented by dividing the probability score give to each payment plan ('p*i*' in equation 1) by the total probability score obtained by adding the probability scores across the payment plans (Ptotal in equation 1). This was done for each respondent separately before calculating the mean probability scores. Each respondent's weighted probability scores added up to one. The mean probability scores also added up to one and they also logically reflected the subscription behaviour of the sample towards each payment plan (equation 2 in Figure 6.21).

Figure 6.21 Equations used in the weighting process

$$p_{ws} = \frac{p_i}{p_{total}} \quad \text{----------equation 1}$$

Where

p_{ws} = weighted probability score

p_i = raw probability score

p_{total} = total probability score obtained by adding probability scores across the mutually exculsive items

$$\text{Mean probability score} = \left(\frac{\sum_i p_{ws}}{n} \right) \quad \text{----------equation 2}$$

As there was no control over how respondents gave their scores, some respondents entered zeros to either one, some, or all the payment plans. Such scores did not adhere with the mutually exclusive nature of the payment plans and were regarded as inappropriate assigning of probability scores. Responses of respondents who gave scores this way were excluded from the analysis. There were respondents who entered numbers outside the range of zero and ten. Such numbers were also treated as inappropriate scores and were removed from the analysis.

6.4 Test Products

6.4.1 WAP-Capable Mobile Phones

Treatments that provided contextual information before getting respondents to indicate their purchase probabilities used Web interfaces to facilitate information search. Respondents were required to choose and view information, as they would do on the Internet, using the Web browsers. To encourage active information search it was decided to use a product that was fairly new to the market (Dahan & Srinivasan 2000; Urban *et al.* 1997; 1996). The Wireless Application Protocol (WAP)-capable mobile phone was being introduced in the New Zealand market at the time of this research. The WAP technology gave Web surfing

capability to mobile phones. The technology is expected to bring about major changes in the way mobile telephony would be used by individuals and businesses. Being an innovation in the market, the WAP-capable mobile phone was chosen as the test product for which probability data were sought by the two surveys implemented in the current research.

6.4.2 Mobile Phone Payment Plans

To meet the second objective, the research collected subscription probabilities for seven payment plans available from mobile phone service providers. Six of the seven plans were ones that providers promoted in their communications. A seventh option (Others) was included to accommodate all other plans, thereby making them mutually exclusive to one another. Respondents used the method implemented in their respective treatment to indicate the probability to subscribe to the payment plans of their providers. Tables 6.1 and 6.2 show the payment plans and its features as shown to respondents in the two treatments.

Table 6.1 Payment Plans Offered by Telecom New Zealand

	Monthly Fee	Nights and Weekends free mins	Nights and Wknds Telecom Talk Time/mins	Nights and Wknds Other Natioanl Calls/mins	Daytime Telecom Talk Time/mins	Daytime Other National Calls/mins
Mytime 200	$34.95	200	25 c	49 c	65 c	$1.29
Anytime 40	$45.00	40	35 c	70 c	35 c	70 c
Anytime 80	$75.00	80	35 c	50 c	35 c	50 c
Anytime 200	$125.00	200	35 c	45 c	35 c	45 c
Prepaid	00.00	00	89 c		89 c	
Mytime 50	$19.95	50	25 c	49 c	70 c	$1.39

Table 6.2 Payment Plans Offered by Vodafone New Zealand

	Monthly Access	Included Min /mth	Daytime rate	Nights and Weekends rate
Get 70	$20.00	70	$1.39	49 c
Get 200	$30.00	200	99 c	49 c
Daytime 40	$35.51	40	79 c	44 c
Daytime 80	$70.00	80	45 c	39 c
Daytime 200	$120.00	200	44 c	37 c
Prepay			89 c	89 c

Day time plans- for calls during working hours(Monday to Friday). Get plans - for calls during the evening and weekends.

Respondents who did not use mobile phones were asked to indicate their probability for signing up with the mobile phone service providers. Options offered to this group were as follows:

- Vodafone
- Telecom
- Other providers
- Will not need the service of a mobile service provider.

6.5 Survey Features

6.5.1 Assigning Respondents to Treatments

Investigators undertaking experimental survey research have to decide beforehand the questionnaire version that each respondent will complete. This requires the generation of treatment groups, production of separate questionnaire versions and adoption of suitable coding procedures to track down the version that each respondent has completed. In the case of Internet-based surveys the questionnaire versions have to be kept in separate directories. The site addresses need to be conveyed to respondents using separate letter or email versions. Apart from the above tasks, this research design has ramifications from a

statistical point. In most cases, the final treatment sizes produced, tended to be imbalanced. Such imbalanced treatments are vulnerable to violation of the equal variance assumption of ANOVA. Serious departure from that assumption would render ANOVA results being incorrectly interpreted.

In the current research, scripts were developed to perform the survey task with special emphasis on generation of equal sized treatment groups. This was done using an algorithm that generated and assigned numbers starting from one to 'Tn' (where 'Tn' was the maximum treatment number) to respondents in the order they accessed the site. After the "Tn^{th}" respondent, the series was repeated to produce sets of consecutive numbers in a cyclic order (e.g. 1,2,3,1,2,3,1,2,3...). Numbers assigned to respondents were recorded in a separate field (Treatment) in the database table. Based on this number, the Web page containing the corresponding treatment question was forwarded to the respondent's computer. An example of the scripts that performed the task is given below.

IF Rst("Treatment") = "1" and THEN RESPONSE REDIRECT(Point&Click. asp)
IF Rst("Treatment") = "2" and THEN RESPONSE REDIRECT(SearchEngine. asp)
IF Rst("Treatment") = "3" and THEN RESPONSE REDIRECT(Control. asp)

The scripts forwarded the treatment questions consecutively to respondents in the order they arrived to the survey site. As the order was cycled, there were equal numbers in the treatment groups. The above scripts eliminated the generation of treatment groups before hand. They also ensured equal numbers in the treatment groups. All respondents were asked to come to one Web site. This required the use of just one version of the cover and reminder letters.

6.5.2 Data Collation

When respondents entered their answer and clicked the "Next" button on the Web page, the action evoked scripts that opened a connection between the respondent's computer and the database on the server. Responses entered were transferred using forms into

corresponding fields in the database table. An example of a script that inserted response values into a field called "Age" in the database table is shown below.

IF Form ("Age") = Rst ("Age") = Form("Age")

Each respondent's answers were made ready for analysis almost immediately. The above line of script eliminated the task of checking questionnaires to see that they were answered logically. It also eliminated the manual entering of responses into a database. Not having to perform these regular survey tasks brought about considerable savings in cost and time.

6.5.3 Response-Based Question Skips

As the responses were entered into the database table almost immediately, the values were available to subsequent scripts to carry out tasks such as question skipping. Skips were achieved by the "IF...THEN and RESPONSE REDIRECT" conditional statement. An example of scripts that performed skips is give below.

IF Rst("Employment") = "1" and THEN RESPONSE REDIRECT(Salary.asp)
IF Rst("Employment") = "2" and THEN RESPONSE REDIRECT(Salary.asp)
IF Rst("Employment") = "3" and THEN RESPONSE REDIRECT(Age.asp)

After a connection has been established between the respondent's computer and the database on the server, the "IF" statement was executed to verify the value entered in the "Employment" field in the database table. Based on the value, the "RESPONSE REDIRECT" statement forwarded one of the Web pages (Salary.asp or Age.asp) to the respondent's computer.

The above scripts eliminated the inclusion of instructions about question skipping in the questionnaire. Hence, there was no requirement to check the questionnaires to see whether they were completed logically. The above scripts also ensured the production of a clean data set that was ready for analysis at the end of the survey.

6.5.4 Updating the Name Database

When the last question was submitted, a set of scripts was evoked that identified the respondent in the name table and deleted his or her information from the database. An example of the SQL statement used to carry out this task is given below:

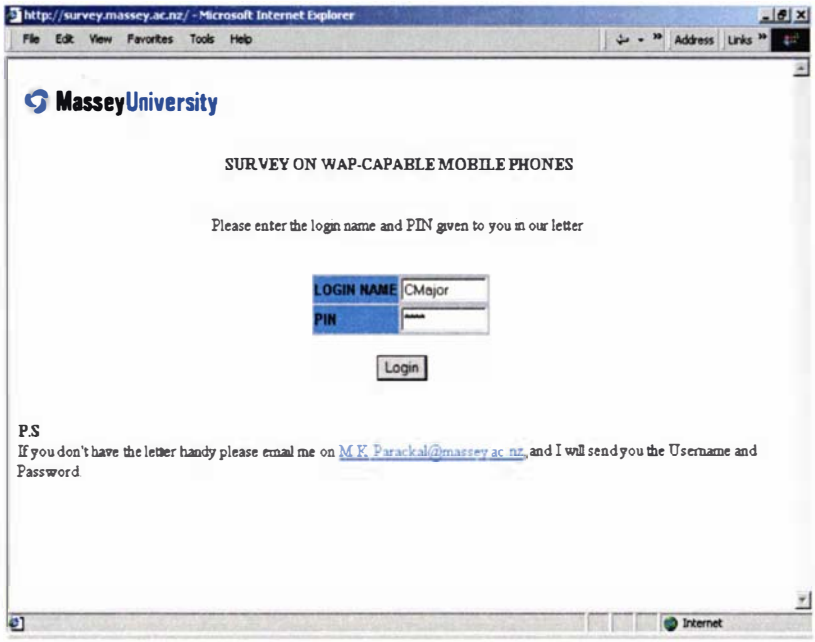
```
Dim Sql
Sql=Delete from NameTable Rst(Login Name)=Secession(Login Name) &
Rst(Access Code)=Session(Access Code)
```

The above lines of script kept the name table updated with only those who had not participated in the survey. The name table was ready at all times for preparing the reminder letters. These scripts also ensured that the research complies with the ethical requirement of having to delete respondents' information held in the database after the survey.

6.5.5 Survey Security

The survey site was secured by a database driven security system that required a login name and an access code to enter. On arrival at the site, respondents were presented with a login page that requested their login name and access code given to them in the cover and reminder letters. They were allowed into the site on typing in the correct login name and access code and hitting the "Login" button. If the entries were incorrect the login page was shown again with a request to type in the correct login name and access code. The security system secured the survey to be accessed only by those selected to participate in the survey. A sample of the Web page requesting the username and access code is shown in Figure 6.22 below.

Figure 6.22 Security Web Page used for Securing the Survey Site to Participants



6.6 Survey Results

In this section of the current chapter, the results of the two Internet-based surveys implemented in the research carried out for this thesis are reported. The section is divided into two parts; in the first part, results of the Vodafone survey are reported and in the second part, results of the New Zealand survey are reported. The two parts commence by providing an overview of the survey approach adopted.

6.6.1 Vodafone Survey

Response Rate

At the end of the survey period (30 days), 494 valid responses were received. Response rate calculated after removing the GNAs (Gone No Address) was 15% (see Table 6.3). The response rate was considered "low" when compared to surveys that have used just one contact letter (Brennan, 1994).

Table 6.3 GNAs, Refusals and Response Rate of the Vodafone Survey

Number of contacts made	3388
GNAs	84
Refusals	12
Survey participants	494
Response rate[1]	15%

As the survey was accessible only to Internet users, survey participants originated from the Internet population within the Vodafone clientele. To compare the response rate with other studies it was therefore necessary to weight the response rate by the incidence of the Internet in the Vodafone clientele; this, however, was not available. For the sake of making comparisons, the response rate was weighted by the incidence of the Internet in the New Zealand population at the time of the survey. The incidence of Internet in New Zealand households as reported by Nielsen//NetRatings (2002) and the New Zealand Ministry of Economic Development (2002) was 52%. The response rate calculated based on the above Internet incidence rate was 29%[2]. This is likely to be an inflated estimate as the incidence of Internet amongst the Vodafone clientele (mobile phone users) is bound to be higher than the national figure. The assumption is made based on the premise that individuals who are innovative in adopting mobile technology would exhibit similar innovativeness in adopting other communication systems.

An earlier New Zealand based study that used a similar approach of contacting respondents reported an estimated response rate of 22% (Brennan *et al.* 1999). Other Internet-based surveys that used census and probability samples have reported response rates ranging from 19% to 62% (Couper 2001; Jones & Pitt 1999; Dillman *et al.* 1998). The studies cited above, including the New Zealand one, surveyed populations in which everyone had Internet access (incidence rate of the Internet was 100%). The estimated response rate of 29% obtained in the current survey, though inflated, was within the range reported in the literature.

[1] Survey participants/(Number of contacts – GNAs – Refusals)*100
494/(3388-84) * 100 = 14.95%
[2] 14.95/0.52 = 28.8%

Data Quality

The survey approach employed in this research was one that is currently being developed. To assess whether the approach was successful in surveying a sample that matched the demographic make-up of the target population, comparisons were made between survey participants (referred to as "final sample" in this discussion) and the random sample selected from the Vodafone clientele (referred to as "original sample" in this discussion) on three demographic variables (age, employment and gender). Comparisons with the client list itself were not possible, as access to the list was not available. Information on the three demographic variables was obtained from the original client list.

Age

To compare the two samples on age, individuals were categorised as teenagers (19 yrs and below), young adults (20 - 24 yrs), adults (25 - 44 yrs), older adults (45 – 60 yrs), elderly (above 60 yrs) and age not disclosed (see Table 6.4). Comparisons were made between the two samples using descriptive statistics (frequencies) for these categories. Absolute differences in proportions ranged from 0.3 to 1.6. To make a relative assessment of these differences, absolute percentage differences were calculated (see Table 6.4); these ranged from 3% to 24%. For three categories (Adults, Older adults, Age not disclosed), the percentage difference was less than five percent and for two categories (Teenagers, Young adults), it was less than ten percent. Elderly participants exhibited the largest difference, being under represented by 24% in the final sample. This came as no surprise, as some people in this age category are not comfortable with using the Internet. Another assessment made was to compare the rank orders of the categories by proportions. The rank orders of categories were the same in the two samples (see Table 6.5).

Table 6.4 Proportions and Ranks of Age Categories in the Final and Original Samples

	Final Sample			Original Sample			Absolute
	N	%	Rank	N	%	Rank	Percentage Difference*
Teenagers (19 yrs & below)	20	4.0	5	146	4.3	5	7
Young adults (20 – 24 yrs)	99	20.0	2	628	18.5	2	8
Adults (25– 44 yrs)	209	42.3	1	1491	43.9	1	4
Older adults (45 – 59 yrs)	90	18.2	3	596	17.6	3	4
Elderly (above 60 yrs)	16	3.2	6	142	4.2	6	24
Age not disclosed	60	12.1	4	393	11.6	4	3
Total	**494**	**100**		**3396**	**100**		

* [(% Final Sample - % Original Sample)/Original Sample]*100

The above observations suggest that the final sample closely matched the original sample on all age categories except the elderly. The fact that all age categories were represented in the final sample and occurred in the same rank order as in the original sample suggests that the survey approach was successful in producing a sample that closely matched the sampling frame by age.

Employment

To compare the two samples on employment, individuals were categorised as employed, self-employed, student, retired, others, unemployed and employment not disclosed. The categorisation was based on employment details included in the original sample. Table 6.5 shows the proportions of these categories in the final and original samples. To make relative assessment of the differences, absolute percentage differences were calculated (see Table 6.5) and they ranged from 2% to 39%. For two categories (employment and others), differences were less than 5% and for two others (students and not disclosed), differences were less than 10%. These four categories made up 85% of the two samples. Absolute percentage differences were comparatively large for the remaining three categories (self-employed, retired, unemployed), ranging from 14% to 39%. Rank orders of the categories

were different in the two samples. Differences in rank orders were seen for three (self employed, others, not disclosed) of the seven categories. The composition of the final and original sample by employment appears to be somewhat different. All the same, the final sample had representation from all seven categories.

Table 6.5 Employment Status of Participants in the Final and Original Samples

	Final Sample			Original Sample			Absolute Percentage
	N	%	Rank	N	%	Rank	Difference*
Employed	298	60.3	1	2017	59.4	1	2
Self employed	38	7.7	4	345	10.2	3	25
Retired	9	1.8	7	71	2.1	7	14
Student	29	5.9	5	188	5.5	5	7
Others	49	9.9	3	351	10.3	2	4
Unemployed	21	4.3	6	105	3.1	6	39
Not disclosed	50	10.1	2	319	9.4	4	7
Total	**494**	**100**		**3396**	**100**		

* [(% Final Sample - % Original Sample)/Original Sample]*100

Gender

Proportions of male and female respondents in the final and original sample are shown in Table 6.6. Differences were marginal, with male respondents being less by 2% in the final sample, and female respondents and those who did not disclose their gender being more by 1% in the final sample. Difference in ratio between male and female respondents (excluding respondents who did not disclose their gender) was a meagre 0.1 (1.4 and 1.5 respectively). Rank orderings of the three groups were the same in the two samples. These observations suggest that the final sample closely matched the original sample on gender.

Table 6.6 Gender Split in the Final and Original Samples

	Final Sample			Original Sample		
	N	**%**	**Rank**	**N**	**%**	**Rank**
Male	268	54	1	1901	56	1
Female	194	39	2	1291	38	2
Not disclosed	32	7	3	204	6	3
Total	**494**	**100**		**3396**	**100**	

Treatment Generation

Researchers have been successful in eliminating the task of pre-assigning respondents to treatments in Internet-based surveys (Vehovar *et al.* 2000; Parackal & Brennan 1999). In the studies cited, respondents were allocated to treatments on arriving at the survey site (Parackal & Brennan 1999) or after answering some initial questions (Vehovar *et al.* 2000). These studies used non-probability sample techniques to select sample members. The approach, however, was advantageous, as it required just one version of the announcement that informed the Web site to all respondents. The task of reminding respondents was also simplified as only one version was required and there was no requirement of keeping track of respondents' treatment membership.

The studies cited above used scripts that allocated respondents randomly to treatments on their arrival at the survey site. As respondents access the survey site at times that suit them, randomness produced in the treatments was very much dependent on the randomness of arrivals. Whether this approach reproduced the variations of the population in the treatments requires verification. What was noticed in the studies cited above was that the treatments were not balanced (see Table 6.7). This restricted the number of satisfactory comparisons that were possible.

Table 6.7 Treatment Sizes Obtained in the Vodafone Survey, Parackal & Brennan (1999) and Vehovar *et al.* (2000)

	Vodafone Survey		Parackal & Brennan (1999)		Vehovar *et al.* 2000	
	N	%	N	%	N	%
Treatment One	170	34.4	117	70	**	60
Treatment Two	159	32.2	84	30	**	40
Treatment Three	165	33.4	*	*		
Total	**494**	**100**	**201**	**100**		**100**

* Not applicable ** Not reported

Having seen in the literature that online random allocation to treatment produced unequal treatment sizes (see Table 6.7) it was decided not to employ the approach in the current research. To produce equal treatment sizes, the current research developed and employed scripts that allocated respondents consecutively to treatments. These scripts, along with offering the benefits of online generation of treatments, were successful in producing reasonably balanced treatment sizes (34.4% 32.2% 33.4%). Table 6.7 shows the treatment sizes produced in the Vodafone survey and compares them with those produced in Vehovar *et al.* (2000) and Parackal & Brennan (1999).

Treatments produced were tested for homogeneity of variance using the Levene statistic. This test is included in SPSS to verify the equal variance assumption of ANOVA. The null hypothesis of Levene's test is that the difference in variances between treatments is zero. If the significance produced in the test is less than 5% then variances in the treatments are not equal. Levene's test was executed for the twelve and six months-probability data for WAP-capable mobile phones collected in the three treatments. In both instances, results confirm that the variances in the treatments were similar ($p > 0.05$) (see Table 6.8).

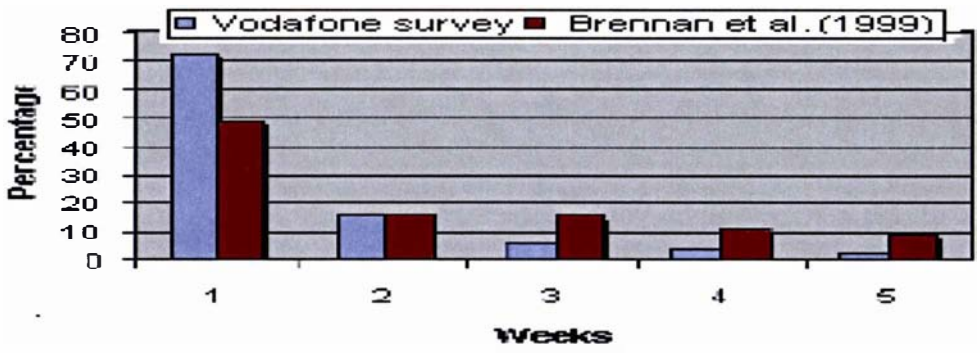
Table 6.8 Homogeneity of Treatments Produced in the Vodafone Survey

	Levene Statistic	df1	df2	Sig
Twelve-months probability data	.742	2	457	.477
Six-months probability data	.786	2	457	.456

Questionnaire Returns

One advantage of Internet-based surveys observed in the literature was the quick turn around of completed questionnaires (Brennan *et al.* 1999; Parackal & Brennan 1998; Rae & Brennan 1998; Tse 1998). This observation was re-confirmed in the Vodafone Survey with over 70% of respondents (see Figure 6.23) returning the questionnaire in the first week. This was a marked improvement over the previous New Zealand based study that reported 49% returning the questionnaire in the first week (Brennan *et al.* 1999). Figure 6.23 shows the weekly return of the Vodafone survey and that of Brennan *et al.* (1999).

Figure 6.23 Weekly Questionnaire Returns in the Vodafone Survey and Parackal & Brennan (1999)



Completion Rate

A major problem of Internet-based surveys raised in the literature was the high dropout rate (Brennan *et al.* 1999; Kottler 1997b). One reason for this was the incompatibility of respondents' computers to run the survey scripts (Aoki & Elasmar 2000; Kottler 1997b). In the current survey, the dropout rate was comparatively low with 93% completing all

questions pertaining to the research objectives ("completed the survey" + "completed most of the survey"). Table 6.9 shows the break down of completion rates in the Vodafone survey and in a previous New Zealand based study (Brennan *et al.* 1999).

Table 6.9 Completion Rates of the Vodafone Survey and Brennan *et al.* (1999)

	Vodafone Survey		Brennan *et al.* 1999	
	N	%	N	%
Completed the survey	443	90	154	70
Completed most of the survey	17	3	-	-
Abandoned the survey	34	7	66	30
Total	**494**	**100**	**220**	**100**

The satisfactory completion rate observed in the current survey could be because of more advanced models of computers used by respondents to access the survey site. This was evident from the entries of browser versions made in the log file of the server. Information from the log file revealed that almost everyone (99%) used either a version four or five of Internet Explorer or Netscape to access the survey site. The operating system required to run these versions is either Windows 3.1 or Windows NT. This suggests that almost everyone who participated in the survey used one of the recent models of computer to access the survey site. Table 6.10 summarises the browser versions of respondents obtained from the log file. In addition to the above observation, the fact that no respondents registered access or navigation problems suggests that the scripts employed ran successfully on their computers.

Table 6.10 Browser (Internet Explorer/Netscape) Versions of Respondents in the Vodafone Survey

Browser Versions	N	%
3	7	1
4	47	10
5	440	89
Total	**494**	**100**

6.6.2 New Zealand Survey

As the sample for this survey was drawn from the electoral roll, it included respondents who used mobile phones (mobile phone users) and those who did not use mobile phones (non-mobile phone users). Probability data for WAP-capable mobile phones were collected from the two groups separately. This survey also fielded the treatments required to collect the data required to achieve the second research objective. The questionnaire included probability questions on payment plans offered by the two major mobile service providers in New Zealand (Telecom and Vodafone). Mobile phone users were divided by the service provider and directed to corresponding sections in the questionnaire. As the sample included non-mobile phone users, the questionnaire also included a question directed at these individuals. They were asked to give their probability to sign up to use the services offered by the mobile phone service providers. Thus, it was possible in this survey to compare the treatments using different groups.

One clear draw back of the Vodafone survey was the low response rate, subjecting its results to non-response bias. To monitor the non-response bias caused by non-Internet users, Quigley *et al.*'s (2000) approach had to be employed in its entirety. The same was employed in the New Zealand survey. The alternative survey mode was implemented by including a reply paid post card in the letters sent out to respondents. The letters instructed non-Internet users to return the card by post to receive a paper version of the questionnaire. Data collected via the Internet was used to execute the comparisons required to achieve the

objectives. Data collected by the alternative survey mode was used to verify whether results of the Internet survey were suitable for this research.

While analysing the probability data collected in the Vodafone survey, it became evident that respondents in the treatments that provided contextual information (Point & Click and Search Engine) could wade through the interfaces without performing the task of viewing information. This would make the experimental treatments ineffective to such respondents, and including their responses would distort the comparisons. To prevent this from happening, it was decided to record the files that respondents viewed in the New Zealand survey. This allowed analyses to be performed using the responses of those who actually viewed the contextual information in the New Zealand survey.

Response Rate

The New Zealand survey produced 729 valid responses. The response rate calculated after removing refusals and GNAs (gone no address) was 30%. One reason for the low response rate could be the timing of the survey that coincided with the September 11th disaster in the United States of America. All the same, the response rate was comparable to similar approaches in the literature that ranged from 9% to 44% (Schonlau *et al.* 2001). It was also similar to the weighted response rate estimated in the Vodafone survey.

The number of respondents who completed the survey over the Internet was 403 (55%) and the number of respondents who completed the survey by filling in a hard copy of the questionnaire was 326 (45%). The proportion of Internet participants in the current study (55%) was comparable to the Internet population of New Zealand (52%) during the time of this study (Nielsen//NetRatings 2002; Ministry of Economic Development 2002). The response rate obtained for the Internet-based survey was 16%. Offering the alternative survey mode pushed the overall response rate to a modest 30%.

Representativeness of the Actual Sample

The crucial test for the survey approach employed as far as the current research was concerned was to find out whether the sample of respondents who participated via the Internet resembled the New Zealand population. To investigate this, comparisons were made on two demographic variables (age and gender) between Internet participants (referred to as "actual sample" for this discussion), Internet and mail participants (referred to as "final sample" for this discussion), the random sample selected from the electoral roll (referred to as "original sample" for this discussion), and the 2001 Census data (referred to as "2001 Census" for this discussion). Comparisons with the electoral roll were not possible as the list from which the sample was drawn was no longer in existence. Instead, the original sample was included in the comparison, as it theoretically resembled the sampling frame (electoral roll). The reason for making the comparison on age and gender was purely because of the availability of the information in all four datasets.

As a starting point, statistical differences in the actual and final sample were investigated on the two demographic variables (age and gender). For this purpose the two data sets were merged into a single data set. In the combined data set, the actual sample (Internet participants) was coded as "zero" and the final sample (Internet and mail participants) was coded as "one". Independent t-tests were executed to see if the mean age and proportion of one of the genders (male participants) were the same or different in the two samples. The analyses revealed that the actual and the final sample had similar gender distributions ($p > 0.05$). The samples, however, were significantly different on mean age ($p < 0.05$). Results of the analyses are summarised in Table 6.11.

Table 6.11 Mean Age and Proportion of Males in the Actual and Final Samples

	Actual sample	**Final sample**	**T**	**Sig**
Mean Age	45	49	4.232	0.000**
Proportion of Male participants	0.46	0.45	-0.312	0.755

** Significant at the 0.000 level

Age

To identify the age categories that caused the mean age to be different in the two samples, respondents' ages were categorised into intervals of 5, starting from 20 years of age as reported in the 2001 Census. Proportions of the categories were compared across the four samples (actual sample, final sample, original sample and 2001 census). Table 6.12 shows the proportions of these categories in the four samples and differences between the actual and other samples (A-B; A-C; A-D) for each category.

Table 6.12 Proportions of Age Categories in the Actual Sample, Final Sample, and 2001 Census

Age Categories	Actual sample (A)	Final sample (B)	Original sample (C)	2001 Census (D)	Absolute & Percentage difference					
					A-B	%	A-C	%	A-D	%
20-24	8	6	8	9	2	33	0	0	1	11
25-29	5	4	9	9	1	25	4	44	4	44
30-34	12	11	10	11	1	9	2	20	1	9
35-39	14	12	11	11	2	17	3	27	3	27
40-44	13	10	11	11	3	30	2	18	2	18
45-49	13	11	9	10	2	18	4	44	3	30
50-54	10	10	9	9	0	0	1	11	1	11
55-59	9	9	8	7	0	0	1	13	2	29
60-64	6	9	6	6	3	33	0	0	0	0
65 +	10	18	19	17	8	44	9	47	7	41
	100	**100**	**100**	**100**						

Examining the differences in proportions of categories revealed that "65 yrs & above" had the highest absolute difference (8%, 9% and 7%). The percentage difference for this category was also consistently high (44%, 47% and 41%). The fact that this category was under represented in the final sample came as no surprise. In general people of these age categories tend to use the Internet less. As for the current research, this outcome was

beneficial as individuals included from these categories were those who were familiar with using the Internet. Also, the test products could have more relevance to those who participated from these age categories than those who did not participate.

The absolute difference in proportion of the "60-64 yrs" category between the actual and final sample was 3% and the percentage difference was 33% (see Table 6.12). Proportions of this category in the actual sample, original sample and 2001 Census were the same (6%). Differences between the actual and final sample observed for these two categories (44% for the 65 & above; 33% for the 60-64 yrs) could be the cause for the significant results produced in the Independent t-test in Table 6.11 above.

Comparing the actual sample with the final sample showed that for two age categories (50-54 yrs and 55-59 yrs) proportions of participants were comparable (10% and 9% respectively)(see Table 6.12). For all other categories, differences seen were marginal with the actual sample producing slightly higher proportions. For five of these categories (20-24 yrs; 25-29 yrs; 30-34 yrs; 35-39 yrs; 45-49 yrs), differences were either 1% or 2% and, for one category (40-44 yrs), the difference was 3%. Ranges calculated for the two samples were comparable (71 and 73), suggesting that the two samples included individuals from the same age range.

Comparing the actual sample with the original sample and the 2001 Census showed that the proportions in the actual sample were less for two categories (25-29 yrs and the above 65 yrs). For one category (60-64 yrs), no difference was seen across the three data sets. For all the other categories (20-24, 30-34, 35-39, 40-44, 45-49, 50-54, and 55-59 years), the actual sample had marginally higher proportions, with differences ranging from 1% to 4%. Perhaps individuals in these categories tend to be more innovative towards technology, making them suitable for this survey.

Gender Split

Table 6.13 shows the comparison of the gender split between the actual sample, original sample and the 2001 Census data. The final sample was not included in this comparison as the actual and final sample had the same proportion of female and male participants (see Table 6.11). Differences across the three samples for male and female participants ranged from 2% to 3%.

The proportion of female participants in the actual sample was higher by 2% and that of male participants was lower by 2% when compared with the original sample. The differences were in the same direction when the actual sample was compared against the 2001 Census with the proportion of female participants being higher in the actual sample by 3% and that of male participates being lower by 3%. In all three data sets, the proportions of female participants were higher than those of male participants.

Table 6.13 Gender split in the Actual Sample, Original Sample and 2001 Census

	Actual sample	**Original sample**	**2001 Census**
Female	54	52	51
Male	46	48	49
Ratio	1.17	1.08	1.04
	100	100	100

Non-Response Bias

While the survey approach produced sufficient responses via the Internet to perform the necessary analyses, the response rate obtained was rather low. It was therefore essential to verify whether the data were of reasonable quality and not adversely affected by non-response bias. As the survey employed Quigley *et al.*'s (2000) approach in its entirety, data were collected from respondents who did not want to participate via the Internet (non-Internet users). These respondents completed a paper version of the questionnaire. The paper version did not field the treatments but followed the approach used in the Standard

treatment. Availability of this data permitted an investigation to find out whether the non-inclusion of non-Internet users could bias the results. For this purpose purchase probability data collected in the Mail survey were compared with those collected in the Internet treatment that implemented the Juster Scale on it own (Standard treatment). Comparisons were restricted to these two groups as they used the same approach of implementing the Juster Scale.

Mean probability scores obtained in the Internet-based survey were higher than those obtained in the Mail survey for Mobile phone users and Non-mobile phone users. This suggests that those who participated on the Internet were more innovative towards the test product than those who chose to participate by filling in a paper version of the questionnaire. In a sense, the Internet-based survey was successful in capturing individuals who were more innovative to the test product. This was beneficial for this research as purchase probability data were collected from individuals to whom the test product was relevant. Table 6.14 shows the mean probability scores obtained for the two groups in the two surveys.

Table 6.14 Mean Probability Scores of WAP-Capable Mobile Phones Obtained in the Internet-Based Survey and Mail Survey

		N	Mean	Std.Dev
Non mobile phone users				
12 months-purchase probability data	Internet	66	0.11	0.19
	Mail	127	0.08	0.18
6 months-purchase probability data	Internet	66	0.08	0.14
	Mail	126	0.07	0.19
Mobile phone users				
12 months-purchase probability data	Internet	206	0.33	0.34
	Mail	189	0.18	0.28
6 months-purchase probability data	Internet	206	0.22	0.30
	Mail	189	0.10	0.22

Independent t-tests were run to investigate whether there were statistical differences between the mean probability scores produced in the two surveys. In the case of non-mobile phone users, t value produced for the twelve months-probability data was –1.004 and the associated *p* was 0.317 and the t value for the six months-probability data was –0.219 and the associated *p* was 0.827. In both the twelve and six months-probability data, significance did not reach the rejection level ($\alpha = 0.05$). The results suggest that both Internet and Mail participants expressed similar purchase probability towards the product. As such, not including the data of those who participate in the Mail survey had no adverse effect on the aggregate purchase rate.

Table 6.15 Comparisons of Mean Probability Scores of WAP-Capable Phones between the Internet-Based Survey and Mail Survey for Non-Mobile Phone Users

		Levene's Test for Equality of Variances		T-test for Equality of Means		
		F	Sig	T	df	Sig
12 months purchase probability data	Equal variance assumed	2.061	0.153	-1.004	191.0	0.317
	Equal variance not assumed			-.989	126.4	0.325
6 months purchase probability data	Equal variance assumed	0.504	0.479	-0.219	190.0	0.827
	Equal variance not assumed			-0.243	173.1	0.809

In the case of mobile phone users, however, the results were different. T value produced for the twelve months-probability data was –4.652 and the associated *p* value was 0.000 and the t value for the six months-probability data was –4.482 and the associated *p* value was 0.000 (see Table 6.15). In both the twelve and six months-probability data, differences were statistically significant. The results suggest that the two groups were different, with Internet participants exhibiting higher aggregate adoption rates than mail participants (see Table 6.14). Estimates from the Internet-based survey alone may not be reflective of the general population. As the sample was a random selection from the electoral roll, those who participated via the Internet could be assumed to be part of the Internet population of New Zealand. Hence, estimates made in the Internet-based survey could be safely assumed to reflect the Internet population of New Zealand.

Table 6.16 Comparisons of Mean Probability Scores of WAP-Capable Phones between the Internet-Based Survey and Mail Survey for Mobile Phone Users

		Levene's Test for Equality of Variances		T-test for Equality of Means		
		F	Sig	T	df	Sig
12 month purchase probability data	Equal variance assumed	21.294	0.000	-4.613	393.0	0.000
	Equal variance not assumed			-4.652	388.3	0.000
6 months purchase probability data	Equal variance assumed	31.154	0.000	-4.422	393.0	0.000
	Equal variance not assumed			-4.482	373.2	0.000

Comparisons made earlier for age and gender (Tables 6.11, 6.12, 6.13) showed that Internet participants closely resembled the sampling frame and the national population on those variables. This could be because of the comparatively large Internet population that New Zealand has. The Internet population therefore could be viewed as being the optimal target population for forecasting studies (Kingsley & Anderson 1998; Weimann 1994) such as the one done in this research. In the articles cited, recommendation was made to survey innovators (Foxall & Goldsmith 1994; Roger 1983; Robertson 1971) to produce valid forecasts of future purchase behaviour. The survey approach used for this thesis was successful in achieving this for mobile phone users.

6.7 Chapter Summary

This research had two primary objectives. The first one was to investigate whether the Juster Scale required additional inputs to collect probability data in the context of making a purchase. The second one was to find out whether forecasts of mutually exclusive items based on probability scores that added up to one were any different to those based on

probability scores that did not add up to one. To achieve these objectives, the research designed and implemented two Internet-based surveys. The first survey was implemented on a random sample drawn from the client base of Vodafone New Zealand. This survey fielded the treatments required to collect data to achieve the first objective. The second survey was applied on a random sample drawn from the New Zealand electoral roll. This survey fielded the treatments required to collect data for the first and second objectives.

Respondents were contacted via their postal address with a request to participate in the respective Internet-based surveys. The contact letters communicated the usernames, access codes and the URL of the survey site to the respondents. In the Vodafone survey only one initial mail out was possible whereas in the New Zealand survey three mail outs were employed. In the latter survey the additional mail outs permitted the use of reminder letters to increase the survey response rate. In this survey, Quigley *et al.*'s (2000) approach was employed in its entirety, that is, Internet users were requested to complete the survey online and non-Internet users were offered the opportunity to participate in the survey by requesting for a paper version of the questionnaire. The approach allowed the monitoring of non-response bias caused by the exclusion of non-Internet users in the analyses.

To achieve the first objective, purchase probability data for WAP-capable mobile phones were collected on the Juster Scale in three separate treatments. In one, the scale was implemented on it own (Standard) and in the other two, the scale was implemented after respondents had the opportunity to view contextual information similar to when making a purchase (Point & Click and Search Engine). Mean probability scores obtained in the treatments were compared to find out whether the three approaches produced similar or different forecasts.

Respondents included in the Vodafone survey were all mobile phone users. In this survey, respondents were asked to indicate probability scores on the Juster Scale to replace their current mobile phone with WAP-capable ones. The New Zealand survey however was implemented on the general population of New Zealand, and the sample included respondents who owned and did not own mobile phones. This sample provided the

opportunity to compare the treatments on different sub groups. The questionnaire was designed to collect probability data from these sub groups separately. Those who owned mobile phones were asked to indicate probability scores on the Juster Scale to replace their current mobile phones with WAP-capable ones. Those who did not own mobile phones were asked to indicate probability scores on the Juster Scale to buy WAP-capable mobile phones. Probability data were collected for two time periods (six and twelve months) from both sub groups.

Treatments required to collect data for the second objective were built into the New Zealand survey. As the sample was drawn from the general population it included respondents who subscribed to the services of the two main mobile telephone service providers (Vodafone and Telecom) in New Zealand. The questionnaire included questions that asked respondents to indicate their probabilities to subscribe to the payment plans of their respective provider. These were collected in two separate treatments. In one, the electronic Constant Sum Scale was implemented. This scale forced respondents to give probability scores to the payment plans that added up to one. The other used a method that did not require respondents to give scores that added up to one. Mean probability scores obtained in the two treatments were compared to see whether they were statistically similar or different.

As the sample included non mobile phone users, the questionnaire also included a question directed at these individuals. They were asked to indicate the probabilities to use the services offered by mobile phone service providers. This provided an additional group in which the two treatments were compared.

The current research developed scripts for generation of equal sized treatment groups, collation of responses into a dataset, managing of skips, removal of personal details of respondents who completed the survey and a security system that secured the site to survey participants. Internet technology allowed the execution of these scripts to manage the survey process. It also ensured the production of clean data that were ready to be analysed at the closure of the survey.

This chapter also reported the results from the two Internet-based surveys (Vodafone and New Zealand surveys) implemented in the current research were reported. Results of the Vodafone survey reported here were based on an initial introductory letter mailed out to 3388 potential respondents, requesting their participation in an Internet-based survey. The survey received 494 valid responses at the end of the survey period (30 days) and the response rate calculated was 15%. The estimated response rate based on the incidence rate of the Internet in the national population (the overall incidence rate of Internet adoption was over 52%) was 29%. While this was most likely to be an inflated estimate it was within the range reported in the academic literature for Internet-based surveys that used probability samples (19% to 62%).

To assess the extent to which the survey participants resembled the target population, comparisons were made between survey participants and the original random sample selected from the client list on three demographic variables (age, employment, and gender). All age categories were represented in the final sample. The differences in proportions were quite marginal for all categories except the elderly. The latter category was underrepresented; this was not seen as a problem as the test products (WAP-capable mobile phones and payment plans) may not have been particularly relevant to members of this category.

In the case of employment there were some differences between the final and original sample. For three categories (self employed, retired, unemployed), the difference was quite large, with the absolute percentage differences ranging from 14% to 39%. Differences were seen in the rank orderings of the categories in the two samples. While there were differences between the two samples on this variable, all categories were represented in the final sample.

The differences in ratio of male and female respondents (excluding respondents who did not disclose their gender) in the two samples were marginal (1.4 and 1.5 respectively). The rank orderings of the three groups (males, females and respondents who did not disclose

their gender) were the same. By and large, the gender split reasonably matched in the actual and final samples.

The above observations suggest that the survey approach adopted was reasonably successful in producing a sample that closely matched the sampling frame (Vodafone clientele). A major drawback of the approach was the low response rate and the possible effect of non-response bias on the results.

Other investigations included comparing completion rate, questionnaire returns and treatment generation with similar studies in the literature. Results showed that there were overall improvements in completion rate, time taken for receiving completed questionnaires, and response rates. The approach employed was also successful in producing equal treatment size that had similar variance. The survey approach was seen as being efficient in collecting probability data required to achieve the objectives of the current research.

The New Zealand survey was implemented on a random sample of 3000 respondents selected from the 2001 New Zealand electoral roll. Two reminder letters were used to encourage non-respondents to participate in the survey. As with the previous survey, the letter provided each respondent an unique username and an access code to access the survey site. This survey employed Quigley *et al.*'s (2000) approach in its entirety and data were collected via the Internet and by a mail survey. The number of valid responses after two reminder letters was 729. The number of participants via the Internet was 403 (55%) and those who chose to fill in a paper version of the questionnaire were 326 (45%). Analyses required for the research objective were performed using responses received via the Internet.

To assess the extent the participants of this survey resembled the population, comparisons were made between Internet participants (actual sample), Internet and mail participants (final sample), the random sample selected from the electoral roll (original sample), and the 2001 Census data on two demographic variables (age and gender). Independent t-tests

showed that the actual and final sample were the same on the gender split but different on the mean age of respondents.

Comparing the actual sample with the original sample and the 2001 Census data showed that the "65 years and above" and the "20-24" years categories were under represented in the actual sample. For all other age categories (30-34, 35-39, 40-44, 45-49, 50-54, and 55-59), the actual sample had slightly higher proportions of respondents than the other two data sets. Individuals in these age categories apparently are those who would use mobile phones (test product) and also be familiar with using the Internet (survey mode), making the sample optimal for collecting probability data for this research.

As data were also collected from those who did not want to participate via the Internet, an investigation was carried out to see whether their exclusion in the analyses gave rise to non-response bias. Results of this investigation showed that for non-mobile phone users the samples obtained via the Internet and by the Mail survey expressed similar adoption rates. The Internet sample could be regarded as being reflective of the target population. Similar conclusions could not be made for the mobile phone users. However, it must be stated that data collected for the latter group came from respondents who were seen as being innovative. The higher expressed adoption rate of this group suggests that the test product (WAP-capable mobile phones) had more relevance to these individuals. In this sense, the survey approach was successful in obtaining a sample that was optimal for providing probability data for the test products.

7 CONTEXT OF THE JUSTER SCALE: RESULTS AND DISCUSSION

7.1 Overview of the Research

Contextual literature reviewed in Chapter Three suggested that the context of questions influenced their response distributions (Schuman *et al.* 1983; Sudman & Bradburn 1982; Schuman *et al.* 1981; Schuman & Presser 1981; Duncan & Schuman 1980). The literature cited emphasised the importance of presenting survey questions in the right context to elicit accurate responses. The Juster Scale studies were reviewed in the light of the knowledge gained from the contextual literature. This review revealed that in many Juster Scale studies (McDonald & Alpert 2001; Brennan 1995; Brennan *et al.* 1995a; 1995b; Brennan & Esslemont 1994a; 1994b; Gan *et al.* 1986; Juster 1966; Byrnes 1964), the context of the scale was not given due consideration. The review in Chapter Three also identified three factors (Question order, Respondents' interpretations and Survey practices) that exhibited tendencies towards altering the context of the scale. Juster Scale investigators failed to recognise these factors as being detrimental to the context of the Juster Scale; hence, no effort was made in any of the Juster Scale studies reviewed to control them. It is possible for these factors to individually or collectively influence the context of the Juster Scale, resulting in the scale being implemented in contexts different to what was originally intended. Results of such Juster Scale studies may not be comparable as was seen in the contextual literature. Therefore the Juster Scale requires fresh testing in a controlled context to generate comparable results to evaluate its true reliability.

Concern raised above regarding the comparability of Juster Scale studies, warrants investigations of the factors mentioned. Before undertaking these investigations, it was imperative that the contextual requirement of the Juster Scale be standardised. With this in view, an attempt was made in this current research to standardise the context of the Juster Scale through an empirical investigation. The research objective set for this purpose is stated below:

> To investigate whether the Juster Scale required additional contextual information to collect purchase probability data in a purchasing context.

To achieve the objective stated above by using statistical techniques, the following null and alternative hypotheses were formulated:

> **Null hypothesis (H0)**: Mean probability scores obtained in the treatment that implemented the Juster Scale on its own without providing contextual information will be similar to those obtained in treatments that implemented the scale after contextual information was provided (ˆx1 =ˆx2=ˆx3).
>
> **Alternative hypothesis (H1)**: Mean probability scores obtained in the treatment that implemented the Juster Scale on its own without providing contextual information will be lower than those obtained in treatments that implemented the scale after contextual information was provided (i.e. at least one of the means was different).

The above hypotheses were tested using probability data collected on the Juster Scale through quantitative survey techniques implemented over the Internet. Probability data were collected in three separate treatments to make the necessary comparisons. In the first treatment, the Juster Scale was implemented on its own (Standard treatment). Respondents allocated to this treatment were not given any additional contextual information apart from a brief explanation of the test product (WAP-capable mobile phone). In the second and third treatments, the Juster Scale was implemented after respondents had the opportunity to search and view contextual information (Point & Click treatment and Search Engine treatment). In the Point & Click treatment, contextual information was provided via an interface that listed items as hyperlinks to external Web sites (Urban *et al.* 1966). In the Search Engine treatment, contextual information was provided via a search engine (Brucke 1985). Respondents allocated to these treatments searched and viewed information about the test product similar to when making an online purchase. On completion of searching and viewing information, respondents were asked to indicate their probability scores on the Juster Scale. Mean probability scores obtained in the treatments were compared for statistical differences using Analysis of variance (ANOVA).

The three treatments were built into the two Internet-based surveys (Vodafone and New Zealand survey) implemented in the current research. The Vodafone survey was executed on a random sample selected from the clientele of Vodafone New Zealand. All

respondents in this sample were mobile phone users and were asked the following questions:

- *"Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next **TWELVE MONTHS**, that is up to the end of < >?"*
- *"Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next **SIX MONTHS**, that is up to the end of < >?"*

The New Zealand survey was implemented on a random sample selected from the New Zealand electoral roll. The sample included both mobile phone users and non-mobile phone users. Mobile phone users were asked the same questions shown above. Non-mobile phone users were asked the following questions:

- *"Taking everything into account, what are the chances that you would purchase a WAP-capable mobile phone within the next **TWELVE MONTHS**, that is up to the end of < >?"*
- *"Taking everything into account, what are the chances that you would purchase a WAP-capable mobile phone within the next **SIX MONTHS**, that is up to the end of < >?"*

Respondents answered the above question by indicating their probabilities on the Juster Scale. In the following sections, results of analyses carried out on the data from the two surveys are reported separately. The chapter concludes with discussion of the results.

7.2 Vodafone Survey

The Vodafone survey produced 460 usable responses for the analyses carried out. ANOVA tests were executed to see if the mean probability scores obtained in the treatments were statistically similar or different. Tables 7.1 and 7.2 summarise the results of the ANOVA tests executed on the twelve and six months-probability data.

Table 7.1 Comparisons of Mean Probability Scores for the Twelve Months-Probability Data

	N	Mean	F	Sig
Standard	167	0.407		
Point and Click	139	0.390	0.17	0.84
Search Engine	154	0.412		

Table 7.2 Comparisons of Mean Probability Scores for the Six Months-Probability Data

	N	Mean	F	Sig
Standard	167	0.259		
Point and Click	139	0.243	0.22	0.81
Search Engine	154	0.237		

Analysis carried out on the twelve months-probability data produced an F statistic of 0.17 with a p value of 0.84. Analysis carried out on the six months-probability data produced an F statistic of 0.22 with a p value of 0.81. In both analyses significances of the p values were greater than the alpha level set ($\alpha = 0.05$) for this research. The Scheffe test for joint pair wise comparisons included with the ANOVA in SPSS, grouped the three means into one subset for the twelve and six months-probability data, suggesting that they were similar. Evidence provided by the twelve and six months-probability data was not sufficient to reject the null hypothesis. From this, it can be concluded that mean probability scores obtained in the three treatments were similar with differences observed being caused by sampling error.

One drawback of the Vodafone survey was that no procedure was in place to learn whether respondents actually engaged in the antecedent activities. Interfaces facilitating antecedent activities in the Point & Click and Search Engine treatments were designed to allow respondents to perform these activities as they would in the real world. For this, respondents were required to perform the antecedent activities on their own. Hence, there was the possibility of some not performing the required antecedent activities. While this always happens in the real world, in this research, the treatment effect was made void for such respondents. Including their responses in the analyses would make

the comparisons within the context of the current research inappropriate. The non-significant results in the above analyses might have been caused by this drawback. As this was identified after the completion of the survey, it was not possible to rectify the effect of the drawback in the Vodafone survey. This was rectified in the New Zealand survey by recording the files that respondents viewed, allowing analyses to be done using respondents who performed the required antecedent task.

7.3 New Zealand Survey

The New Zealand survey provided comparisons between the three treatments on a different sample. This survey was implemented on a random sample of 3000 respondents selected from the New Zealand electoral roll. The sample included both mobile phone users and non-mobile phone users, providing comparisons between the treatments on two separate groups. An initial question (see Figure 7.1) enquired whether respondents owned a mobile phone or not. Based on the response given to this question, respondents were directed to corresponding sections in the questionnaire.

Figure 7.1 Web Page with the Question that Enquired about Mobile Phone Ownership

Frequency of responses to the above question revealed that 308 (77%) respondents indicated that they owned a mobile phone and 92 (23%) respondents indicated that they

did not own one. There were 25 respondents who abandoned the survey prior to answering the questions on WAP-capable mobile phones. Hence, the actual numbers of usable responses in the two groups were 294 (mobile phone users) and 81 (non-mobile phone users) respectively. The latter group (non-mobile phone users) was comparatively small; all the same it provided comparisons between the treatments using a different set of respondents.

To rectify the drawback of the Vodafone survey in the New Zealand survey, compliance status of respondents to perform the required antecedent activities were recorded in the two treatments (Point & Click and Search Engine). This allowed analyses to be executed using the responses of respondents who viewed information in the two treatments (Point and Click and the Search Engine). In the following sections, results of these analyses performed using data from mobile phone users and non-mobile phone users are reported separately

7.3.1 Mobile Phone Users

ANOVA tests were executed to investigate whether mean probability scores obtained in the treatments were different because of sampling error or because of the treatment effect (that is, providing contextual information prior the Juster Scale). Analysis done on the twelve month-probability data produced a F statistic of 0.96 with a p value of 0.385 (see Table 7.3) and that done on the six months-probability data produced a F statistic of 1.151 with a p value of 0.22 (see Table 7.4). In both analyses, significance was higher than the alpha level set ($\alpha = 0.05$) for this research. The Scheffe test (pair wise comparisons of means) executed on the twelve and six months-probability data confirmed that the means obtained in the three treatments were similar. Thus, the evidence provided by the twelve and six months-probability data was not sufficient to reject the null hypothesis. These results concurred with the observations made in the Vodafone survey.

Table 7.3 Comparisons of Mean Probability Scores for the Twelve Months-Probability Data

	N	Mean	F	Sig
Standard	103	0.33		
Point and Click	85	0.27	0.96	0.385
Search Engine	106	0.27		

Table 7.4 Comparisons of Mean Probability Scores for the Six Months-Probability Data

	N	Mean	F	Sig
Standard	103	0.22		
Point and Click	85	0.16	1.515	0.222
Search Engine	106	0.17		

Compliance status recorded revealed that 62% (n = 53) in the Point & Click treatment and 37% (n = 39) in the Search Engine treatment complied with the requirement of searching and viewing contextual information about WAP-capable mobile phones. The number of information items viewed by these respondents ranged from one to fourteen. The rest (38% and 63% in the respective treatments) indicated their probability scores without viewing any information. Treatment effect for the latter respondents was made void and having them in the analyses could distort the results. The non-significant results obtained in the previous analyses could be because the treatment effect was nullified for a good number of respondents in the two treatments that provided contextual information (Point & Click and Search Engine). Similar behaviour could have occurred in the Vodafone survey, in which case the comparisons discussed above may not be appropriate as far as the current research objective was concerned.

As compliance status was secured in the New Zealand survey, it was possible to repeat the above analyses after removing the responses of respondents who failed to view information in the two treatments that provided information (Point & Click and Search Engine). Tables 7.5 and 7.6 summarise the results of the ANOVA executed on the twelve and six months-probability data.

Table 7.5 Comparisons of Mean Probability Scores after Applying the Information Viewing Criterion for the Twelve Months-Probability Data

	N	Mean	F	Sig
Standard	103	0.325		
Point and Click	53	0.277	0.449	0.639
Search Engine	39	0.289		

Table 7.6 Comparisons of Mean Probability Scores after Applying the Information Viewing Criterion for the Six Months-Probability Data

	N	Mean	F	Sig
Standard	103	2.20		
Point and Click	53	1.68	0.623	0.537
Search Engine	39	1.95		

The F value obtained for the twelve months-probability data was 0.449 and the associated *p* value was 0.639 and that obtained for the six months-probability data was 0.623 and the associated *p* value was 0.537. In both the above analyses, significance of the F values were over the alpha level set ($\alpha = 0.05$) for this research. These results confirmed those of the previous analyses and those of the Vodafone survey (reported in Section 7.3).

In the above analyses, sizes of treatments that provided information (Point & Click, n = 53; Search Engine, n = 39) were reduced because of the information-viewing criterion. Consequently, treatments were not well balanced (see 'n' in Table 7.6), violating the equal treatment size assumption of the ANOVA. If there was a serious departure from this underlying assumption, then results would be invalid. Hence, the above comparison was repeated using Kruskal-Wallis H, an alternative non-parametric test (Gibbons 1985; Hollander & Wolfe 1973). The H statistic for this test is derived from the mean ranks of probability scores in the treatments and is compared against a Chi-square distribution. The associated *p* value calculated for the Chi-square value enables conclusions to be made. A definite advantage of this test for the current data was that it does not require treatments to be balanced. Tables 7.7 and 7.8 summarise the results of the Kruskal-

Wallis H test executed on the twelve and six months-probability data collected in the three treatments.

Table 7.7 Comparisons of Mean Ranks after Applying the Information Viewing Criterion for the Twelve Months-Probability Data

	N	Mean Ranking	Chi Square	df	Sig
Standard	103	100.6			
Point and Click	53	92.6	0.724	2	0.696
Search Engine	39	98.4			

Table 7.8 Comparisons of Mean Ranks after Applying the Information Viewing Criterion for the Six Months-Probability Data

	N	Mean Ranking	Chi Square	df	Sig
Standard	103	99.4			
Point and Click	53	93.1	0.623	2	0.732
Search Engine	39	101.0			

The Chi-square value obtained for the twelve months-probability data was 0.724 and the associated *p* value was 0.696 and that for the six months-probability data was 0.623 and the associated *p* value was 0.732. In both analyses, Chi-square values were significant at levels higher than the alpha level set ($\alpha = 0.05$) set for this research, thereby not being sufficient to reject the null hypothesis. Results of these analyses concur with the earlier observations.

7.3.2 Non-Mobile Phone Users

The total number of respondents in this group was 81. Consequently, the number in each treatment was rather small (Standard: n = 33; Point & Click: n = 24; Search Engine: n = 24). Again, this raised the concern of the equal variance assumption of ANOVA being violated. To verify whether violation occurred, Levene's homogeneity

test was executed on the three treatments for the twelve and six months-probability data. Results of the Levene's test are summarised in Table 7.9 below.

Table 7.9 Homogeneity Tests of the Twelve and Six Months-Probability Data Collected From Non-Mobile Phone Users

	Levene Statistic	df1	df2	Sig
Twelve months-probability data	7.002	2	78	0.002**
Six months-probability data	9.126	2	78	0.000*

** Significant at 0.01
*** Significant at 0.000

The Levene's statistic produced for the twelve months-probability data was 7.002 and the associated *p* value was 0.002 and that produced for the six months-probability data was 9.126 and the associated *p* value was 0.000, that is, within the rejection level (α = 0.05). Consequently, the null hypothesis (difference in variance between treatment was zero) was rejected, confirming that variances in the treatments were not the same. This conclusion ruled out the use of ANOVA to analyse the data. Hence, Krushal-Wallis H was used to analyse the data collected from this group.

Krushal-Wallis H was executed separately on the twelve and six months-probability data collected in the three treatments for non-mobile phone users (see Tables 7.10 and 7.11). Chi-square produced for the twelve months-probability data was 9.146 and the associated significance was 0.009. Chi-square produced for the six months-probability data was 6.636 and the associated significance was 0.034. In both instances, the null hypothesis was rejected ($p > 0.05$), supporting the alternative hypothesis. Results of these analyses were different to those produced for mobile phone users.

Table 7.10 Comparisons of Mean Ranks for the Twelve Months-Probability Data

	n	Mean Rank	Chi square	df	Sig
Standard	33	40.30			
Point and Click	24	32.56	9.146	2	0.009**
Search Engine	24	50.40			

**Significant at the .01 level.

Table 7.11 Comparisons of Mean Ranks for the Six Months-Probability Data

	n	Mean Rank	Chi square	df	Sig
Standard	33	45.06			
Point and Click	24	32.96	6.636	2	0.034*
Search Engine	24	43.46			

*Significant at the 0.05 level.

There was no technique available in conjunction with Kruskal-Wallis H to identify treatments that were different. A technique called Games-Howell (Toothacker 1993), recommended when variances are not equal or variances and group sizes are not equal was found to be appropriate for this data. This technique essentially tolerates small groups or treatment sizes, provided they are greater than 5. It provides pair wise comparisons of treatment means based on the q-statistic distribution. Results of the Games-Howell's test executed on the twelve and six months-probability data are summarised in Table 7.12 and Table 7.13 respectively.

In the case of the twelve months-probability data (Table 7.12), mean difference for treatments that provided contextual information (Point & Click and Search Engine) was 0.138 and the associated p value was 0.028. The significance of the mean difference was within the rejection level set ($\alpha = 0.05$) in this research, leading to the conclusion that treatments that provided contextual information were different. To further investigate the reasons for difference between these treatments, the compliance status of respondents in performing antecedent activities was examined. This revealed that none viewed information in the Point & Click treatment, whereas nine (38%) viewed information in the Search Engine treatment. The number of items viewed by the latter

group ranged from one to six. Differences in the compliance status in the two treatments could have caused this result. As the numbers in the treatments were very small, further analysis was not attempted.

Mean difference between the Standard treatment and Point & Click treatment was 0.084 with an associated *p* value of 0.054 and that between the Standard treatment and Search Engine treatment was -0.053 with an associated *p* value of 0.641. In both these latter comparisons the *p* value did not reach the rejection level ($\alpha = 0.05$) set for research (see Table 7.12). These latter results aligned with those obtained for mobile phone users in the Vodafone and New Zealand survey.

Table 7.12 Pair wise Comparisons of Mean Differences for Twelve Months-Probability Data

Treatment (I)	Treatment (J)	Mean Diff (I-J)	Std Error	Sig
Standard	Point & Click	0.084	0.034	0.054
	Search Engine	-0.053	0.059	0.641
Point & Click	Standard	-0.084	0.035	0.054
	Search Engine	-0.138	0.050	0.028 *
Search Engine	Standard	0.053	0.059	0.641
	Point & Click	0.138	0.050	0.028*

*Significant at the .05 level.

Table 7.13 Pair wise Comparisons of Mean Differences for Six Months-Probability Probability Data

Treatment (I)	Treatment (J)	Mean Diff (I-J)	Std Error	Sig
Standard	Point & Click	0.084	0.024	0.018*
	Search Engine	-0.053	0.037	0.974
Point & Click	Standard	-0.084	0.024	0.018*
	Search Engine	-0.138	0.285	0.092
Search Engine	Standard	0.053	0.366	0.974
	Point & Click	0.138	0.028	0.092

*Significant at the .05 level.

Games-Howell's test executed on the six months-probability data produced different results (see Table 7.13). Mean difference between the Standard treatment and the Point & Click treatment was 0.084 and the associated p value was 0.018. The p value was within the rejection level ($\alpha = 0.05$), thereby supporting the alternative hypothesis. Compliance status in the Point & Click treatment revealed that no one viewed information. Consequently, the treatment effect was totally redundant and the two treatments (Standard and Point & Click) were, in essence, the same. Hence, the difference may not have been caused by the treatment effect. There was, however, a difference in the formatting of the questionnaires between the two treatments. In the Standard treatment respondents were presented with the Juster Scale directly whereas in the Point & Click treatment they were routed via the interface that facilitated information search. Maybe, the samples did not have sufficient variances to nullify this difference. This was evident in the response distributions obtained in the three treatments. The response distributions revealed that 67% in the Standard treatment, 92% in the Point & Click treatment and 67% in the Search Engine treatment gave zero probability scores. Proportions of zero probability score were sizeable in the three treatments (see Table 7.14). This suggested that this group of respondents were clearly not interested in the test product used.

Table 7.14 Distributions of Probability Scores obtained in the Treatments

Probability scores	Standard				Point & Click				Search Engine			
	12 months		6 months		12 months		6 months		12 months		6 months	
	N	**%**	**N**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
0	22	67	21	64	19	79	22	92	9	38	16	67
1	3	9	5	15	4	17	2	8	7	29	5	21
2	1	3	3	9	1	4	-	-	3	13	-	-
3	2	6	3	9	-	-	-	-	3	12	1	4
4	2	6	-	-	-	-	-	-	-	-	1	4
5	2	6	-	-	-	-	-	-	-	-	1	4
6	-	-	1	3	-	-	-	-	-	-	-	-
7	1	3	-	-	-	-	-	-	1	4	-	-
8	-	-	-	-	-	-	-	-	1	4	-	-
9	-	-	-	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-	-	-
Mean scores	**0.11**		**0.08**		**0.03**		**0.01**		**0.16**		**0.07**	

7.4 Discussion

The two Internet-based surveys together collected probability data for WAP-capable mobile phones from three separate groups. Twelve and six months-probability data were collected from each of these groups, permitting a total of six separate comparisons between the treatments. Four of the six comparisons were performed using data collected from mobile phone users. The remaining two comparisons were performed using data collected from non-mobile phone users. ANOVA tests were executed to find out whether the mean probability scores obtained in the treatments were different because of sampling error or because of the treatment effect.

In all four comparisons made using data collected from mobile phone users, the ANOVA tests returned non-significant results. Compliance status of respondents to perform the antecedent activities collected in the New Zealand survey allowed additional analyses to be performed. These additional analyses were executed after removing the responses of respondents who failed to perform the required antecedent

activities in the two treatments that provided contextual information (Point & Click and Search Engine). ANOVAs returned non-significant results for the twelve and six month probability data ($p > 0.05$).

Removing the respondents who failed to view information in the Point & Click and Search Engine treatments reduced their sizes. Consequently, treatments became imbalanced, raising the possibility of violating the equal variance assumption of ANOVA. If there was serious violation then the result would be invalid. Hence, the comparisons were repeated using the Kruskal-Wallis H test, a non-parametric alternative that did not require treatments to be of equal size. The test was applied on the twelve and six months-probability data collected from mobile phone users. In both cases the results were non-significant ($p > 0.05$), confirming the results of the previous analyses.

In all the analyses discussed, so far, there was no evidence to reject the null hypothesis. Hence, the alternative hypothesis (H1) could not be supported. It can be concluded that the Juster Scale implemented on its own produced mean probability scores (forecasts) similar to those produced when the scale was implemented after respondents had viewed contextual information. This observation seems to imply that the Juster Scale does not require any additional contextual input.

The above observation was based on tests carried out on one product category (WAP-capable mobile phones). It is therefore essential to see if the results could be re-produced on other product categories. As for this product category (WAP-capable mobile phones), investigation can progress to the next phase outlined in the review of literature in Chapter Three. This will be to investigate whether the factors (question order, the practice of concurrent testing of the Juster Scale and respondents' interpretation) identified in the review of literature altered the context of the Juster Scale. If they were found to alter the context then methods to control these factors must be investigated.

Results based on the non-mobile phone users were not conclusive. The data failed to meet the assumptions of ANOVA, hence, analyses were performed using Kruskal-Wallis H. The test was executed separately on twelve and six months-probability data.

In both instances, the null hypothesis was rejected, supporting the alternative hypothesis. Multiple comparisons of means using Games-Howell's method were employed to identify the treatment that caused the null hypothesis to be rejected. Results produced for the twelve months-probability data revealed that difference occurred between treatments that provided contextual information (Point & Click and Search Engine) ($p < 0.05$). Differences between the treatment that did not provide contextual information (Standard) and the ones that provided contextual information (Point & Click and Search Engine), however, were not significant at 0.05. This latter result supported the observation made by the analyses performed on data collected from mobile phone users.

Results of Games-Howell's test for the six months-probability data revealed a significant difference between the Standard treatment and the Point & Click treatment ($p < 0.05$). This was the only result that was different out of all the analyses performed. Analysing the compliance status in the Point & Click treatment revealed that no respondents performed the antecedent task required. Hence, this treatment was not different to the Standard treatment. The only known difference between these treatments (Standard and Point & Click) was the format of the questionnaire. It was not possible to confirm whether this (format) was responsible for the significant difference between these two treatments, as there was no way to investigate this issue further.

The adoption rates (mean probability scores) obtained for the twelve and six months-probability data exhibited logical increment. However, in both instances, there was a large majority who gave zero probability scores (see Table 7.14). Also, those who gave non-zero probability scores were clustered at the lower end of the scale. The low mean probability scores obtained in the treatments for the twelve and six months-probability data suggest that respondents belonging to this group (non mobile phone users) had very little interest in the test product (WAP-capable mobile phone). Hence, the task of getting respondents to participate in antecedent purchase activities was not successful. Information on compliance status in the Point & Click treatment revealed that no one viewed information; useable responses for any further analysis were zero. In the Search Engine treatment, nine respondents viewed information; this sub-sample was too small to carry out further meaningful analyses. Overall, results based on the six months-

probability data collected from non-mobile phone users were not conclusive, hence, must be interpreted with caution.

7.5 Chapter Summary

The Vodafone survey collected twelve and six months-probability data for WAP-capable mobile phones in the three treatments. ANOVA tests showed that the differences in the mean probability scores between the treatments were chiefly due to sampling error. This was the case for both the twelve and six months-probability data.

The New Zealand survey produced twelve and six months-probability data that originated from mobile phone users and non mobile phone users. ANOVA performed on data from mobile phone users for both the time periods produced F values that were not significant at the alpha value of 0.05 ($p > 0.05$). These results confirmed the observation made in the Vodafone survey. In this survey, compliance status of respondents to perform the antecedent activities was recorded in the treatments that provided contextual information (Point & Click and Search Engine treatments). This allowed further analyses to be executed using responses of respondents who actually performed the antecedent activities in the Point & Click and Search Engine treatments. ANOVA and Kruskal-Wallis H tests confirmed that differences observed were due to sampling error and not because of treatment effect.

In all the above analyses, the p values did not reach the rejection level ($\alpha = 0.05$). Hence, the evidence gathered was insufficient to reject the null hypothesis. From this, it was concluded that the Juster Scale on its own, produced forecasts that were similar to when the scale was implemented with additional contextual information. The Juster Scale, as far as this product is concerned (WAP-capable phone), can be applied on its own to collect probability data in a purchasing context.

Results based on data collected from non-mobile phone users were not very conclusive. Results of the multiple comparisons (Games-Howell's test) performed on the twelve months-probability data supported the conclusion made by the analyses performed on data collected from mobile phone users. The six month-probability data, however, showed otherwise. Not much confidence could be placed on the results of the six

months-probability data, as a sizable proportion of the sample failed to view information in the treatments that provided contextual information (Point & Click and Search Engine). The analysis could not be taken further as the treatments became very small. Hence, results for the six months-probability data collected from non-mobile phone users, must be interpreted with caution.

8 MUTUALLY EXCLUSIVE BEHAVIOURS: RESULTS AND DISCUSSION

8.1 Overview of the Research

Researchers have employed the Juster Scale to collect probability data for mutually exclusive behaviours. Studies examined on this topic revealed a major difficulty with this Juster Scale application (Flannelly *et al.* 2000a; 2000b; 1999; Parackal & Brennan 1999; Hoek & Gendall 1996; 1993). That is, most respondents failed to understand how probability operated when making a choice from a set of mutually exclusive alternatives. This observation was evident from the way respondents assigned probability scores to each of the mutually exclusive alternatives in the studies mentioned above. In these studies, the total probability score assigned to the alternatives did not tally up to ten. Consequently, probability scores failed to convey the purchase behaviour of respondents towards each alternative. The mean probability scores also failed to reflect the purchase behaviour of the sample towards each alternative.

Two approaches were used in the literature to overcome this problem. One was a weighting process[1] applied to the raw probability scores collected for a set of mutually exclusive alternatives (Hoek & Gendall 1993). The mean weighted scores obtained for the alternatives added up to one and they logically reflected the individuals' and the sample's behaviour towards the alternatives. The second approach comprised of employing the Constant Sum Scale to collect probability scores for a set of mutually exclusive alternatives (Hamilton - Gibbs *et al.* 1994). This scale was successful in collecting probability scores that tallied up to ten. The two approaches mentioned above have produced satisfactory results in separate studies (Flannelly *et al.* 2000a; 2000b; 1999; 1998; Hoek & Gendall 1997). However, there has been no attempt in the literature to establish which one produced better forecasts. Comparing the two approaches would also elucidate whether forecasts for mutually exclusive alternatives based on data that needed weighting were any different to those that did not require weighting. In the current research, an attempt was made to establish the method that was

[1] Weighting was done by dividing probability scores of each item by the total probability score of all mutually exclusive items {P (x1)/P (x1+x2+x3+x4)}*100

best suited to collect probability data for mutually exclusive alternatives. The research objective for this purpose is stated below:

> To investigate whether forecasts of mutually exclusive behaviours based on probability scores that did not add up to ten (weighted probability scores) were more accurate than those based on scores that added up to ten (Constant Sum Scale).

To achieve the objective using statistical techniques, the following null and alternative hypotheses were formulated:

> **Null hypothesis (H_0)**: Forecasting accuracy of mutually exclusive items based on the weighted probability scores will be similar to those made on the Constant Sum Scale ($\hat{x}_1 = \hat{x}_2$).
>
> **Alternative hypothesis (H_2):** Forecasting accuracy of mutually exclusive items based on the weighted probability scores will be less accurate than those made on the Constant Sum Scale ($\hat{x}_1 \neq \hat{x}_2$).

Data required to test the above hypotheses were secured through the New Zealand survey. The two approaches introduced above were implemented in separate treatments built into this survey. For the sake of simplicity, the treatments will be called "Constant Sum Scale" and "Weighted-scores" in the following discussions. As the sample for this survey was drawn from the electoral roll, it included respondents who owned mobile phones (mobile phone users) and those who did not own mobile phones (non-mobile phone users). There were 291 respondents who answered the question related to mobile phone users. These respondents were grouped according to their mobile service provider as Telecom subscribers (n = 180) and Vodafone subscribers (n = 111). Respondents were asked to indicate their probability scores to subscribe to six payment plans offered by their respective provider. Payment plans were made mutually exclusive by including a seventh option called "Others" to cover all other plans. There were 75 respondents who answered the question related to non-mobile phone users. These respondents were asked to indicate their probabilities for signing up to use the services offered by mobile phone companies. The three groups (Telecom, Vodafone and non-mobile phone users) together produced 18 separate comparisons between the two approaches (Weighted-scores and Constant Sum Scale).

Raw scores collected in the Weighted-scores treatment were prepared for analyses by weighting them to one. Weighting was applied by dividing the probability score of each payment plan by the total probability score obtained by adding the probability scores across the payment plans. This was done for each respondent separately before calculating item wise means for the weighted probability scores. Raw scores obtained on the Constant Sum Scale did not require any other preparation. In the following section, results of the analyses carried out are reported. Following that is a discussion on the results.

8.2 Telecom Subscribers

Respondents subscribing to Telecom New Zealand were asked to indicate their chances to either of remaining on or changing to another payment plan. Payment plans included with this question were Mytime 50, Mytime 200, Anytime 40, Anytime 80, Anytime 200, Prepaid, and Others. Of the 180 Telecom subscribers who completed this question, 87 (48%) gave probability scores on the Constant Sum Scale and 94 (52%) gave probability scores for each payment plan separately (Weighted-scores).

In the Constant Sum Scale treatment, three respondents did not answer this question. They were recognised by the zeros entered for all the payment plans by the corresponding survey scripts. These respondents were excluded from the analyses, leaving the final number in this treatment at 84. In the Weighted-scores treatment there were nine respondents who gave zeros for one, some or all the payment plans. Two respondents entered numbers outside the range of zero and ten (86 and 34). They were all excluded from the analyses, leaving the final number in this treatment at 83.

8.2.1 Mean Probability Scores

Examining the raw probability scores that respondents registered for the payment plans in the two treatments revealed that the Constant Sum Scale was successful in collecting probability data that added up to one. The mean probability score for each payment plan was interpreted as the proportion of the sample that would subscribe to that payment plan (see Table 8.1).

Table 8.1 Mean Probability Scores Based on the Raw Scores for Telecom Subscribers

	Constant Sum Scale (n = 84)	Weighted-scores (n = 83)
	Mean	Mean
Mytime 50	0.04	0.11
Mytime 200	0.07	0.11
Anytime 40	0.03	0.08
Anytime 80	0.02	0.06
Anytime 200	0.13	0.10
Prepaid	0.37	0.54
Others	0.34	0.30
Total	**1.00**	**1.30**

In the Weighted-scores treatment, 15 respondents (18%) registered probability scores for all the seven payment plans. The remaining 68 respondents (82%) registered probability scores for one to six of the seven payment plans, with the majority doing so for one (n = 60; 88%). For the data to have the mutually exclusive nature, a transformation, converting the missing entries to zero was performed. For five (Mytime 50; Mytime 200; Anytime 40; Anytime 80; Prepaid) of the seven payment plans, mean probability scores obtained in this treatment were higher than in the Constant Sum Scale treatment. The sum of the mean probability scores in the Weighted-scores treatment was greater than one (1.30 in Table 8.1). Consequently, mean probability scores failed to reflect the subscription behaviour of the sample towards each payment plan. The raw scores had to be weighted to one to meaningful interpret the subscription behaviours of the sample.

8.2.2 Rank Orders

In the Weighted-scores treatment, individual probability scores were weighted to one prior to calculating the means. The weighted probability scores across the payment plans summed up to "one" for each respondent and the mean probability scores reflected the subscription behaviour of the sample (see Table 8.2). Rank orders of the mean probability scores in the treatments were the same for the most preferred (Prepaid, Others, and Anytime 200) and least preferred (Anytime 40 and Anytime 80) payment

plans. For two payment plans (Mytime 50 and Mytime 200) the order was reversed in the two treatments (see Table 8.2).

Table 8.2 Mean Probability Scores and Ranks of Payment Plans Obtained in the Treatments for Telecom Subscribers

	Constant Sum Scale (n = 84)		Weighted-scores (n = 83)	
	Mean	Rank	Mean	Rank
Mytime 50	0.04	5	0.05	4
Mytime 200	0.07	4	0.04	5
Anytime 40	0.03	6	0.03	6
Anytime 80	0.02	7	0.01	7
Anytime 200	0.13	3	0.06	3
Prepaid	0.37	1	0.51	1
Others	0.34	2	0.30	2
Total	**1.00**		**1.00**	

8.2.3 Comparisons of Treatments

Student's t-tests were executed to find out whether the mean probability scores produced in the treatments were statistically similar or different. Results of the Student's t-tests are summarised in Table 8.3. The Levene's test of homogeneity built in with the Student's t-test in SPSS showed that treatment variances were similar for five pairs (My time 50, Anytime 40, Anytime 80, Prepaid and Others) ($p > 0.05$) and for two pairs (Mytime 200 and Anytime 200), treatment variances were not similar ($p < 0.05$) (see Table 8.3). For these latter two pairs, Student's t-tests with equal variance not assumed were used to interpret the results.

In all seven comparisons, t-statistics were significant at levels higher than the alpha value ($\alpha = 0.05$) set for this research (see Table 8.3). Thus, the evidence collected in the treatments was not sufficient to reject the null hypothesis. This led to the conclusion that mean probability scores obtained in the treatments were similar and differences seen were due to sampling errors.

Table 8.3 Comparisons of Mean Probability Scores of Payment Plans for Telecom Subscribers

		Levene's Test for Equality of Variances		T-test for Equality of Means		
		F	Sig	T	df	Sig
Mytime 50	Equal variance assumed	0.327	0.568	-0.337	165.0	0.736
	Equal variance not assumed			-0.337	153.8	0.737
Mytime 200	Equal variance assumed	4.525	0.035	1.164	165.0	0.246
	Equal variance not assumed			1.166	150.6	0.246
Anytime 40	Equal variance assumed	0.093	0.761	0.147	165.0	0.884
	Equal variance not assumed			0.146	161.5	0.884
Anytime 80	Equal variance assumed	3.156	0.077	0.816	165.0	0.416
	Equal variance not assumed			0.818	144.9	0.415
Anytime 200	Equal variance assumed	7.882	0.006	1.612	165.0	0.109
	Equal variance not assumed			1.615	154.4	0.108
Prepaid	Equal variance assumed	3.805	0.053	-1.898	165.0	0.059
	Equal variance not assumed			-1.897	164.3	0.060
Others	Equal variance assumed	0.004	0.948	0.579	165.0	0.563
	Equal variance not assumed			0.579	164.9	0.563

A close examination of the raw probability scores revealed a major problem in the above comparisons (see Table 8.4). In the Constant Sum Scale treatment, everyone gave scores that added up to ten; this was very much what was anticipated in this treatment. But in the Weighted-scores treatment, contrary to what was anticipated, a large majority (75%; n = 62) gave scores that added up to ten. About 97% (n = 60) of these respondents gave ten as the probability to the payment plan of their choice and zero to the rest. Hence, the probability scores collected in the treatments were similar in nature, which could be the reason for the non-significant results.

Table 8.4 Proportion of Respondents whose Probability Scores Added up to Ten

	Constant Sum Scale		Weighted-scores	
	N	%	N	%
Scores that added up to ten	84	100	62	75
Scores that did not add up to ten	0	0	21	25
Total	**84**	**100**	**83**	**100**

The above analyses were repeated using the responses from the 21 respondents (25%) whose probability scores did not add up to ten in the Weighted-scores treatment (see Table 8.4). This way, probability scores that actually failed to add up to ten were

compared with those that added up to ten. As there were only 21 usable responses in the Weighted-scores treatment, treatment sizes became imbalanced (Constant Sum Scale: n = 84; Weighted-scores: n = 21). This violated the equal sample size assumption of the Student's t-test. Hence, the Mann-Whitney U test, a non-parametric equivalent of the Student's t-test was used. This test does not require the sample to be normally distributed and can be used on small samples (n < 30). It compared the mean ranks of probability scores between the two treatments. Results for the Mann-Whitney U tests are summarised in Table 8.5 below.

Table 8.5 Comparisons of Mean Ranks of Payment Plans for Telecom Subscribers

		N	Mean Rank	Mann-Whitney U	Sig
Mytime 50	Constant Sum Scale	84	49.54	591.0	0.001**
	Weighted-scores	21	66.86		
Mytime 200	Constant Sum Scale	84	50.57	678.0	0.017*
	Weighted-scores	21	62.71		
Anytime 40	Constant Sum Scale	84	50.68	687.0	0.011*
	Weighted-scores	21	62.52		
Anytime 80	Constant Sum Scale	84	50.62	682.0	0.007**
	Weighted-scores	21	62.52		
Anytime 200	Constant Sum Scale	84	52.80	865.5	0.405
	Weighted-scores	21	53.79		
Prepaid	Constant Sum Scale	84	52.49	839.0	0.353
	Weighted-scores	21	55.05		
Others	Constant Sum Scale	84	53.67	826.0	0.313
	Weighted-scores	21	50.33		

* Significant at the 5% level
** Significant at the 1% level

Results of the Mann-Whitney U tests showed that for four (Mytime 50: U = 591, p = 0.001; Mytime 200: U = 678, p = 0.017; Anytime 40: U = 687, p = 0.011; Anytime 80: 682, p = 0.007) of the seven payment plans, mean ranks were significantly different (p < 0.05). Results obtained in these four comparisons supported the alternative hypothesis (H2). The Weighted-scores treatment produced higher mean scores for these four payment plans.

8.3 Vodafone Subscribers

Respondents who subscribed to Vodafone New Zealand were asked to indicate their chances to either remain or change to another one of the payment plans. Payment plans included with the question were Get 70, Get 200, Daytime 400, Daytime 800, Daytime 20, Prepay, and Other plans. Of the 112 Vodafone subscribers who participated, 61 (54%) were allocated to the Constant Sum Scale treatment and 51 (46%) to the Weighted-scores treatment. There were three and five respondents in the Constant Sum Scale treatment and the Weighted-scores treatment respectively who gave zeros for all payment plans. They were removed from the analyses and the final numbers in the Constant Sum Scale treatment and the Weighted-scores treatment were 58 and 46 respectively.

8.3.1 Mean Probability Scores

The Constant Sum Scale was successful in collecting probability data that added up to one for the payment plans. The mean probability scores logically reflected the subscription behaviour of the sample towards each payment plan (see Table 8.6).

Table 8.6 Mean Probability Scores Based on the Raw Scores for Vodafone Subscribers

	Constant Sum Scale (n = 58)	Weighted-scores (n = 46)
	Mean	Mean
Get 70	0.066	0.074
Get 200	0.124	0.172
Daytime 400	0.069	0.057
Daytime 800	0.028	0.076
Daytime 20	0.074	0.080
Prepay	0.524	0.587
Other plans	0.116	0.130
Total	**1.00**	**1.18**

In the Weighted-scores treatment, 10 respondents (20%) registered probability scores for all payment plans. The remaining 41 respondents (80%) registered probability scores

for some of the seven payment plans. The missing entries were transformed to zeros so that the probability data exhibited the mutually exclusive nature. Mean probability scores obtained for six of the seven payment plans (Get 70, Get 200, Daytime 800, Daytime 20, Prepay and Other plans) were higher in this treatment (see Table 8.6). The total of the mean probability scores across the payment plans failed to add up to "one" (see Table 8.6). The mean probability scores had to be weighted to "one" to interpret the subscription behaviour logically.

8.3.2 Rank Orders

In the Weighted-scores treatment, individual scores were weighted before calculating the means. The mean probability scores calculated for the weighted scores added up to one. The weighted mean probability scores were able to explain the subscription behaviour of the sample towards each payment plan (see Table 8.6). In both treatments, rank orders of the three most preferred payment plans (Prepay, Get 200 and Other plans) were the same. The rank orders of the remaining payment plans, however, were different in the two treatments (see Table 8.7).

Table 8.7 Mean Probability Scores of Payment Plans obtained for Vodafone Subscribers

	Constant Sum Scale (n = 58)		**Weighted-scores (n = 46)**	
	Mean	Rank	Mean	Rank
Get 70	0.066	6	0.047	6
Get 200	0.124	2	0.142	2
Daytime 400	0.069	5	0.038	7
Daytime 800	0.028	7	0.059	4
Daytime 20	0.074	4	0.053	5
Prepay	0.524	1	0.529	1
Other plans	0.116	3	0.132	3
Total	**1.00**		**1.00**	

8.3.3 Comparisons of Treatments

Student's t-tests were executed to find out whether the differences in the mean probability scores between the two treatments were statistically significant. The Levene's tests included with the Student's t-test showed that the treatment's variances were similar for all payment plans ($p > 0.05$). The t-statistics produced in all seven comparisons were significant at levels higher than the alpha value ($\alpha = 0.05$) set for this research, that is, not sufficient to reject the null hypothesis. Thus, the means produced by the two approaches were similar across the seven comparisons. These results were in concurrence with those obtained for Telecom subscribers. Results of the Student's t-tests are summarised in Table 8.8.

Table 8.8 Comparisons of Mean Probability Scores of Payment Plans for Vodafone Subscribers

		Levene's Test for Equality of Variances		T-test for Equality of Means		
		F	Sig	T	df	Sig
Get 70	Equal variance assumed	0.730	0.395	0.509	102.0	0.612
	Equal variance not assumed			0.521	101.9	0.603
Get 200	Equal variance assumed	0.666	0.417	-0.323	102.0	0.747
	Equal variance not assumed			-0.316	87.3	0.753
Daytime 400	Equal variance assumed	2.377	0.126	0.752	102.0	0.454
	Equal variance not assumed			0.788	98.6	0.433
Daytime 800	Equal variance assumed	2.851	0.094	-0.933	102.0	0.353
	Equal variance not assumed			-0.892	74.8	0.375
Day time 200	Equal variance assumed	1.105	0.296	0.524	102.0	0.602
	Equal variance not assumed			0.535	101.8	0.594
Prepay	Equal variance assumed	0.582	0.477	-0.054	102.0	0.957
	Equal variance not assumed			-0.053	95.2	0.958
Others	Equal variance assumed	0.653	0.421	-0.260	102.0	0.795
	Equal variance not assumed			-0.256	90.1	0.799

As was observed for the Telecom subscribers, many respondents (65%; n = 30) in the Weighted-scores treatment gave probability scores that added up to ten. Table 8.9 shows the break down of respondents who gave scores that added up to ten and those that did

not add up to ten in the two treatments. This was again similar to what was observed for the Telecom subscriber and could be the reason for the non-significant results obtained.

Table 8.9 Proportion of Respondents whose Probability Scores Added up to Ten

	Constant Sum Scale		Weighted-scores	
	N	**%**	**N**	**%**
Scores that added up to ten	58	100	30	65
Scores that did not add up to ten	0	0	16	35
Total	**58**	**100**	**46**	**100**

The analyses were repeated using the responses of respondents whose probability scores failed to add up to ten (n = 16) in the Weighted-score treatment. The treatment sizes became imbalanced (Weighted-scores n = 16; Constant Sum Scale n = 58), hence, Mann-Whitney U test was employed to compare mean ranks of probability scores in the two treatments. Probability scores in the Weighted-scores treatment were weighted before executing the Mann-Whitney U tests. Results showed that the two treatments were different for two (Daytime 40: U = 352.5, $p = 0.024$; Daytime 80: U = 332.5, $p = 0.007$) out of the seven payment plans ($p < 0.05$) (see Table 8.10). For these two payment plans, the null hypothesis was rejected. Mean ranks obtained in the Weighted-scores treatment were higher than those obtained in the Constant Sum Scale treatment. This pattern was consistent with that seen for Telecom payment plans that exhibited differences (see Table 8.5, p. 185).

Table 8.10 Comparisons of Mean Ranks of Payment Plans for Vodafone Subscribers

		N	Mean Rank	Mann-Whitney U	Exact Sig
Get 70	Constant Sum Scale	58	36.12	384.0	0.091
	Weighted-scores	16	42.50		
Get 200	Constant Sum Scale	58	35.83	367.0	0.068
	Weighted-scores	16	43.56		
Daytime 40	Constant Sum Scale	58	35.58	352.5	0.024*
	Weighted-scores	16	44.47		
Daytime 80	Constant Sum Scale	58	35.23	332.5	0.007**
	Weighted-scores	16	45.72		
Daytime 200	Constant Sum Scale	58	36.42	401.5	0.137
	Weighted-scores	16	41.41		
Prepay	Constant Sum Scale	58	39.28	361.0	0.081
	Weighted-scores	16	31.06		
Others	Constant Sum Scale	58	38.06	431.5	0.254
	Weighted-scores	16	35.47		

*Significant at 0.05
** Significant at 0.01

8.4 Non-Mobile Phone Users

Respondents who stated that they do not use mobile phones were allocated to one of the two treatments (Constant Sum Scale n = 38; Weighted-scores n = 37). They were asked to give their probability for signing up with the mobile service providers (Vodafone or Telecom). To make the options mutually exclusive, two additional alternatives were included ("Other providers" and "Will not need the service of a mobile service provider"). There were three respondents in the Constant Sum Scale treatment and 12 respondents in the Weighted-scores treatments who gave zeros to all the alternatives. They were excluded, leaving 35 respondents in the Constant Sum Scale treatment and 25 respondents in the Weighted-scores treatment whose responses were used in the analyses.

8.4.1 Mean Probability Scores and Rank Orders

As seen in the previous two sets of analyses, the Constant Sum Scale was successful in collecting probability data that logically reflected the behaviour of the sample towards

each option. In the Weighted-scores treatment, the mean probability scores for the four options added to one (on rounding) (see Table 8.11); this was different from the previous two analyses. The two treatments had the same rank orders for the options, suggesting that the data collected in the two treatments were similar in nature.

Table 8.11 Mean Probability Scores Based on the Raw Scores for Non-Mobile Phone Users

	Constant Sum Scale (n = 35)		Weighted-scores (n = 25)	
	Mean probability scores	Rank	Mean probability scores	Rank
Telecom	0.183	3	0.240	3
Vodafone	0.126	2	0.148	2
Other providers	0.037	1	0.048	1
Will not need the service of a mobile service provider	0.654	4	0.568	4
Total	**1.00**		**1.004**	

Examining the response distributions obtained in the Weighted-scores treatment revealed that a majority of respondents (68%) gave probability scores that logically reflected their behaviour towards the options (see Table 8.12). Over half of these respondents (56%) assigned all 10 chances to the option "Will not need the service of a mobile service provider". This conveyed that most respondents were confident of their choice and had no dilemma in assigning all ten probabilities to that option.

8.12 Proportion of Respondents whose Probability Scores Added up to Ten

	Constant Sum Scale		Weighted-scores	
	N	**%**	**N**	**%**
Scores that added up to ten	35	100	17	68
Scores that did not add up to ten	0	0	8	32
Total	**35**	**100**	**25**	**100**

8.4.2 Comparisons of Treatments

Observations made on the mean probability scores, rank orders and the pattern of assigning probability scores to options suggest that the probability scores collected in the two treatments were similar. To confirm this, Student's t-tests were executed to compare the mean probability scores between the treatments. F statistics produced in the Levene's tests confirmed that treatment variances were similar for all four options. T-statistics produced for the four options were significant at levels greater than 0.05 and was not sufficient to reject the null hypothesis. From this it can be concluded that means obtained in the treatments were similar and differences seen were due to sampling error (see Table 8.11).

Table 8.13 Comparisons of Mean Probability Scores for Non-Mobile Phone Users

		Levene's Test for Equality of Variances		T-test for Equality of Means		
		F	**Sig**	**T**	**df**	**Sig**
Vodafone	Equal variance assumed	0.000	0.985	0.056	58.0	0.956
	Equal variance not assumed			0.055	51.5	0.956
Telecom	Equal variance assumed	3.331	0.073	-0.904	58.0	0.370
	Equal variance not assumed				42.6	0.393
Other providers	Equal variance assumed	0.160	0.691	0.119	58.0	0.906
	Equal variance not assumed			0.127	57.8	0.899
Not required	Equal variance assumed	2.288	0.136	0.583	58.0	0.562
	Equal variance not assumed			0.575	49.3	0.568

8.5 Discussion

Purchase probability data for mutually exclusive alternatives were collected from three separate groups (Telecom subscribers, Vodafone subscribers and Non-mobile phone users). Telecom and Vodafone subscribers were required to give their probability scores for subscribing to the payment plans offered by their providers. Non-mobile phone users were required to give their probability scores for signing up with the mobile telephone service providers. In total probability data were secured for 18 individual items,

permitting 18 separate comparisons between the two treatments (Constant Sum Scale and Weighted-scores). The data collected enabled the assessment of the quality and differences of the two methods.

For all three groups, the Constant Sum Scale produced purchase probability data that logically conveyed the purchase behaviour of individuals and samples towards the mutually exclusive alternatives. This, however, was not the case for the probability data collected in the Weighted-scores treatment. For two groups (Telecom and Vodafone subscribers), raw probability scores had to be weighted to convey the purchase behaviour of the individuals and the sample towards the mutually exclusive items. In one group (non-mobile phone users), the mean probability score of each option added up to one.

The Constant Sum Scale was successful in producing data ready for analysis. It also got respondents to convey their purchase probability scores for all the mutually exclusive items. Thus, the data collected did not require further processing to convey the behaviour of the individual or the sample. In the Weighted-scores treatment, however, there was a high number of missing data. Missing entries were transformed to zeros to give the data the mutually exclusive nature. There were more respondents choosing not to answer the question in the Weighted-scores treatment, resulting in the size (n) being smaller for this treatment. Perhaps, the task required of respondents in the Weighted-scores treatment was more difficult. The Constant Sum Scale therefore is recommended for its ability to produce clean data that are ready for analysis and to secure probability data from more number of respondents.

Probability data collected for the individual items in the two treatments provide separate comparisons. In total 18 separate comparisons were made using the Student's t-test to investigate whether the mean probability scores obtained in the treatments were similar or different. In all 18 tests, results failed to reject the null hypothesis. The results concluded that the weighting process and the Constant Sum Scale produced similar mean probability scores (forecasts).

Examining the raw probability scores revealed that the response patterns were similar in the Weighted-scores and the Constant Sum Scale treatments. The Constant Sum Scale

always produced scores that added up to ten. In the Weighted-scores treatment, however, contrary to what was anticipated a considerable number of respondents gave scores that added up to ten. Therefore, probability data obtained in the two treatments were similar in nature and this was confirmed by the Student's t-tests.

In both treatments, a large majority of respondents assigned all 10 chances to a single item. This was the case for all the three groups (Telecom subscribers, Vodafone subscribers and Non-mobile phone users), which suggests that most respondents were absolutely sure about their choices. From this observation it could be concluded that if respondents were certain of their choice then it did not matter which approach was used for collecting probability data. Similar observation was made in the polling studies done on two-party/candidate elections in the literature (Flannelly *et al.* 2000b; Flannelly *et al.* 1999; Flannelly *et al.* 1998). In the studies cited, probability data were collected using a similar approach employed in the Weighted-scores treatment. These studies did not require the weighting process and the forecasting accuracy reported was reasonable. However, if respondents were uncertain of their choice then scores would be more widely distributed across the items. In such instances, a greater degree of irrational assigning of probability scores can arise. This, however, was not seen in the data collected in this research. Most respondents in this research exhibited absolute certainty in their choices (10 out of 10 chances). The overall amount of irrational assigning of scores was considerably small in all three groups. Probably this was the case in the current research because of the test product used. Services rendered by payment plans are consumed continuously and replacement purchases in most cases tend to be habitual renewal of the existing plan. As such, most respondents were sure of what payment plan they would be on in the next six months.

Analyses were attempted to see whether forecasts made for mutually exclusive items based on probability data that added up to ten were statistically different or same as those based on probability data that did not add up to ten. These analyses were performed by retaining the responses of respondents whose scores did not tally up to ten in the Weighted-scores treatment. A large proportion of the treatment became ineligible for these analyses, reducing the treatment size considerably. Consequently, the treatments became imbalanced and comparisons were made using the Mann-Whitney U

test, a non-parametric equivalent of the Student's t-test that compares mean ranks in treatments.

As the data collected from non mobile phone users were essentially similar in the two treatments, they were not appropriate for these latter analyses. Mann-Whitney U tests were executed only on data collected from mobile phone users, providing 14 separate comparisons between treatments. In six (43%) comparisons the mean ranks were significantly different ($p < 0.05$). For all these items, mean ranks produced in the Constant Sum Scale treatment were lower than those produced in the Weighted-scores treatment. This suggests that the latter approach produced higher estimates than the former. In the remaining eight comparisons (57%), mean ranks were not statistically different ($p > 0.05$). No obvious pattern could be recognised in these two sets of items (items that showed difference and those that did not show difference). It was not possible to explore the data further as the numbers were very small in the Weighted-scores treatments.

The six comparisons that exhibited differences (43%) warrant the next stage of the research, that is, the validation of the forecasts. As there were more comparisons (57%) that exhibited no difference in this research, drawing conclusions based on such a validation study may not be appropriate. Hence a conclusive recommendation could not be made as far as saying which of the two approaches was better, perhaps because of the test products used. Respondents in general were able to express absolute confidence in their future plans for these products. This resulted in comparatively fewer irrational assigning of probability scores in the Weighted-scores treatment. Hence, data collected in the treatment became inappropriate as far as the required comparisons were concerned.

8.6 Chapter Summary

In this chapter, results of analyses carried out on probability data collected for mutually exclusive items in the Constant Sum Scale treatment and the Weighted-score treatment were reported. Probability data were collected from three separate groups (Telecom subscribers, Vodafone subscribers and Non-mobile phone users). In all three groups the Constant Sum Scale produced purchase probability data that logically conveyed the

purchase behaviour of the individuals and the sample towards the mutually exclusive items. Probability data collected in the Weighted-scores treatment had to be weighted particularly for the Telecom and Vodafone subscribers to convey the purchase behaviour of the individuals and the sample. For non-mobile phone users, however, the mean scores did convey behaviour logically; hence, the weighting process was not required. Clearly the Constant Sum Scale produced clean data that was ready for analysis. For this reason, the Constant Sum Scale is recommended over the Weighting process.

Conclusions regarding the difference between scores that added up to ten and those that did not add up to ten were not conclusive. In this research, respondents seem to have had greater confidence in their purchase plan. This was evident from the large number of respondents assigning all ten chances to their preferred payment plan. Hence, the discrepancy anticipated in the Weighted-scores treatment was not severe. Analyses were attempted by retaining only those responses that exhibited discrepancy in the Weighted-scores treatment. Results were mixed; in six out of the fourteen comparisons differences occurred, whereas in the remaining eight there was no difference. Further analyses were not attempted because of the small treatment sizes; hence, a conclusive outcome was not secured.

9. CONCLUSIONS

9.1 Introduction

The Juster Scale was the most preferred probability scale in the literature reviewed in Chapter Two. All the same, the review of literature in Chapters Two and Three revealed that the scale's accuracy was not consistent across product categories and studies. Whether this inconsistency was due to methodological issues of the studies reviewed needed investigation. In the research undertaken for this thesis, attempts were made to address two methodological issues. This chapter provides the conclusions derived from the investigations carried out.

9.2 Context of the Juster Scale

In the review of literature in Chapter Three, the context of the Juster Scale was raised as causing some of the variations in its forecasting accuracy. Juster Scale investigators failed to recognise this. Consequently, no effort was made to ensure proper context was present while implementing the scale. The review carried out identified three factors that exhibited strong tendencies to alter the context of the Juster Scale. These factors were question order, the practice of testing the Juster Scale concurrently on different product categories and respondents' interpretation of the question accompanying the Juster Scale. These factors were pertinent to most Juster Scale studies examined. Their influence could have occurred in varying degrees, resulting in the scale being implemented in contexts different to what was originally intended. Results of such studies may not be comparable as was seen in the contextual literature reviewed in Chapter Three. If this was the case, then the Juster Scale would require fresh testing in a controlled context to know its reliability. To systematically investigate the factors mentioned above, it was necessary to first standardise the context of the Juster Scale. In the research carried out for this thesis an attempt was made to standardise the context of the Juster Scale. The objective set for this part of the research was:

> To investigate whether the Juster Scale requires additional contextual information to collect purchase probability data in a purchasing context.

To achieve the above objective, probability data for WAP-capable mobile phones were collected in three separate treatments. In the first treatment, the Juster Scale was implemented on its own without providing additional information (Standard). In the second and third treatments, the Juster Scale was implemented after respondents had the opportunity to search and view information (Point & Click and Search Engine). Probability data were collected for twelve and six months-probability periods. Mean probability scores obtained in the treatments were compared for statistical differences. The three treatments were built into the two Internet-based surveys implemented in this research (Vodafone survey and New Zealand survey). The Vodafone survey was applied on the clientele of Vodafone New Zealand, securing data from mobile phone users. The New Zealand survey was applied on the national population, securing data from mobile phone users and non-mobile phone users. In the following sections, conclusions, limitations and future directions of investigation carried out are discussed.

9.2.1 Conclusions

Mobile Phone Users

Eight separate comparisons were made using data collected from mobile phone users. Two of these used the twelve and six months-probability data collected in the Vodafone survey. The remaining six comparisons were made using the twelve and six months-probability data collected in the New Zealand survey. The latter six comparisons included those made after the information-viewing criterion was applied in the treatments that provided contextual information. ANOVA and Kruskal-Wallis H were employed to investigate whether mean probability scores obtained in the treatments were different because of sampling error or treatment effect.

In all eight analyses, results showed no significant difference between the treatments. The Juster Scale implemented on its own without additional contextual information produced mean probability scores (forecasts) similar to those produced when the scale was implemented after contextual information was provided. Thus it can be concluded, at least for the product (WAP-capable mobile phones) used in this research, that the Juster Scale is a robust forecasting instrument that collected probability data in a typical purchasing environment.

Non-Mobile Phone Users

Two separate comparisons were made using the twelve and six months-probability data collected from non-mobile phone users. Levene's test of homogeneity revealed that treatment variances were not the same, hence, comparisons were made using Kruskal-Wallis H. Results of the two Kruskal-Wallis H tests showed that the mean probability score was different for at least one treatment. Games-Howell's method was used to identify the treatment that was different. In the case of the twelve months-probability data, difference was seen between treatments that provided contextual information. However, there was no difference between the Standard treatment and the Point & Click treatment and between the Standard treatment and the Search Engine treatment. Results of the latter two comparisons were in line with those made for mobile phone users.

In the case of the six months-probability data difference occurred between the Standard treatment and the Point & Click treatment. This result was different from all the others. Investigations into the compliance status of respondents to undertake the required antecedent activities revealed that a large majority failed to view contextual information in the Point & Click treatment. Hence, the comparison was not in line with the objective of the research. Because only a few respondents actual performed the required task, further investigation into possible reasons for the different result could not be attempted.

In Chapter Three, the review of literature discussed three studies (Day *et al.* 1991; Gan *et al.* 1986; Clawson 1971) that tested the Juster Scale concurrently on different product categories. In those studies, forecasting accuracy was inconsistent across the product categories included. Observations on the probability distributions discussed in that section (see Section 3.3) suggest that the samples used were not suitable to collect probability data for product categories for which forecasting accuracy was poor. In those studies, the inconsistent results across product categories appear to be caused by the sample make-up rather than the Juster Scale. A similar situation seems to have occurred for non-mobile phone users in the current research. The test product (WAP-capable mobile) seems to be clearly not suitable for this group. This was evident from the large majority who gave zero probability scores. Also, the fact that many respondents refused to view information was suggestive of the group's lack of interest in the product. Therefore, results of analyses based on data collected from non-mobile

phone users must be interpreted with much caution and, hence, were not given extra emphasis.

9.2.2 Limitations

The survey approach employed was successful in generating equal sized treatments. However, the compliance status criterion implemented reduced the numbers in the treatments, making them small and imbalanced. This was primarily due to low response rate, a consequence of not using standard survey techniques proven to increase response rate. In the Vodafone survey, due to restrictions imposed by the research partner, only one mail out was executed and no incentive was offered. In the New Zealand survey, two reminder letters were used to increase response rate, but an incentive was not offered due to limited funds. Hence, due to small numbers in the treatments, Kruskal-Wallis H, a non-parametric equivalent of ANOVA, was employed to compare treatments. Results, however, would have been more robust if they were based on standard parameters for which variances could be known. This was not possible because of the limitations of the data.

The test product (WAP-capable mobile phones) used in the current research was relevant to mobile phone users. The survey approach was successful in collecting probability data for WAP-capable mobile phones from this group. The product was not so relevant to non-mobile phone users. The poor interest in the product resulted in fewer respondents performing the antecedent activities. While the Vodafone survey brought to light the problem of respondents not performing the antecedent behaviour required, this could have been identified prior to the survey by conducting a pilot study. Piloting would also have revealed whether the test products were appropriate to produce the required comparisons. Consequently, treatments failed to generate the desired comparisons as far as the research was concerned. Hence, conclusions here are restricted to mobile phone users only. Future studies should ensure that the selection of test products is appropriate to all sub-groups.

The Internet technology was successfully employed to field the treatments in this research. Treatments facilitating information search were successful in achieving their purposes. The problem of respondents not performing the required antecedent task,

however, had serious ramifications. While the problem was addressed in the New Zealand survey, it went unnoticed in the Vodafone survey. The extent to which this oversight impacted on the results of the Vodafone survey remains unknown. As results of analyses executed after removing the responses of respondents who failed to view information were in agreement with the other analyses, the suspected ramification may not be severe in this study. Future studies that require respondents to perform certain online tasks must employ appropriate methods to monitor whether the task is carried out.

9.2.3 Future Research Direction

Observations and conclusions made in this research were based on tests carried out for one test product (WAP-capable mobile phone). As the primary objective in this research was to verify the contextual requirement of the Juster Scale, it was necessary to keep all possible biases in check. To eliminate the bias caused by order effect discussed in Chapter Three (in Section 3.3.2), the two Internet-based surveys implemented in the current research collected probability data for just one test product. While the conclusion made in this research might be applicable to similar communication-based products, it is necessary to find out whether it also holds for other products categories such as fast moving consumer items, durables and services. Such replications are essential to corroborate the conclusion made in this research and also to generalise the recommendation.

In this research the contextual requirement of the Juster Scale was standardised for WAP-capable mobile phones. For this product, investigation can move onto the next phase of the current research stream. In the review of literature in Chapter Three question order effect, the practice of testing product concurrently and respondents' interpretation were identified as potential deflectors of the Juster Scale context during implementation (see Section 3.3.2). Investigations can now be undertaken to find out whether these factors truly deflected the context standardised in this research. If they are found to do so, then appropriate methods to keep them in check can be developed and tested. Hence, the next stage of this research theme will be to:

- Investigate whether the question asked before the Juster Scale deflected its context.

- Investigate whether the practice of collecting purchase probability data for different products concurrently was appropriate from the contextual point of view.
- Investigate whether the question form accompanying the Juster Scale fostered the purchasing context.

9.2.4 Managerial Implications

This research showed that the Juster Scale does not require additional contextual background to collect purchase probability data. The scale can be confidently used in mail, telephone and Internet-based surveys without researchers having to be worried about contextual modification. While there are other factors identified in the literature review such as context modifiers, the results of the current research add more strength to the recommendation that the Juster Scale is a satisfactory scale for collecting purchase probability data in a purchasing context.

9.3 Mutually Exclusive Behaviour

The second objective set in this thesis was to address a problem faced while collecting probability data for mutually exclusive behaviours. Researchers in the past have used the Juster Scale to forecast the adoption of mutually exclusive behaviours. All studies examined on this topic reported that respondents in general failed to understand how probability operated when making a choice from a list of mutually exclusive alternatives (Parackal & Brennan, 1999; Hoek & Gendall, 1997; 1996; and 1993). Respondents tended to give probability scores treating each item independent of the others. When probability scores given to items were reviewed together, they failed to explain the purchase behaviour towards the item at the individual and aggregate levels.

Investigators have used two independent approaches to overcome this problem when making forecasts for mutually exclusive behaviours. One was a weighting process implemented on the raw probability scores collected for the mutually exclusive items (Hoek & Gendall 1993). The weighted scores added up to ten or one and also logically reflected the behaviour of individuals and the sample towards each item. The second

approach used the Constant Sum Scale to collect probability data for mutually exclusive items (Hoek & Gendall 1997). Probability scores collected on this scale always added up to ten or one, providing logical interpretations of the behaviour at the individual and aggregate levels. The two approaches were found to produce satisfactory results in separate studies (Flannelly *et al.* 2000; 1999; 1998; Hoek & Gendall 1997; 1996; 1993). There has been no previous attempt directed at finding out which of the two approaches produced better forecasts for mutually exclusive behaviours. Such a comparison would also elucidate whether forecasts for mutually exclusive behaviours based on probability scores that added up to ten were any different to those based on probability scores that did not add up to ten. Hence, the current research attempted a systematic comparison of the above two approaches.

Data required to achieve the above objective were collected by the New Zealand survey. The questionnaire implemented in this survey included questions on subscription plans offered by mobile phone companies. The New Zealand survey implemented the Constant Sum Scale and the weighting process approach in separate treatments. The sample provided comparisons between treatments using three separate groups (Telecom subscribers, Vodafone subscribers, and Non-mobile phone users). Telecom and Vodafone subscribers were asked to give their probability scores to subscribe to seven payment plans offered by their respective provider. Non-mobile phone users were asked to give their probability to subscribe to the services of mobile phone service providers. In the following sections, conclusions, limitations and future directions of the investigation carried out are discussed.

9.3.1 Conclusions

In total 18 separate comparisons were made using data collected from the three groups (Vodafone subscribers, Telecom subscribers and non-mobile phone users). In all comparisons, the Constant Sum Scale was successful in collecting probability data for the mutually exclusive items that added up to ten. The mean probability scores were successful in logically conveying the purchase behaviour of the sample towards each item. This was not the case for probability scores obtained in the Weighted-scores treatment. The raw probability scores collected from two groups (Telecom and Vodafone subscribers) had to be weighted to one to reflect the purchase behaviour of

the sample. In the case of non-mobile phone users, though discrepancy was present, it was not serious. The Constant Sum Scale was able to produce clean probability data that was ready for analysis. For this reason the Constant Sum Scale is recommended over the approach that employs the weighting process.

Student's t-tests were used to investigate whether mean probability scores obtained using the Constant Sum Scale were any different to those obtained by the weighting process. In all 18 comparisons, mean probability scores were statistically similar; suggesting that the two approaches produced similar forecasts. Examination of the probability distribution obtained in the Weighted-scores treatment revealed that a large majority of respondents gave scores to items that added up to ten. The scoring pattern of respondents suggests that they were confident with their purchase plans and hence had no dilemma in assigning all ten chances to their preferred option. Services rendered through payment plans are consumed continuously and those using the services for a while would habitually renew their payment plans with ones they have become comfortable with. Only a small proportion of respondents gave scores that failed to refect their purchase behaviour. The bias caused by this latter group may be random and hence its effect may be also negligible on aggregation. From the above observations it can be concluded that either of the two approaches might be used when a majority of respondents are confident of their behaviour. Similar observations were made in polling data collected for two-candidate elections (Flannelly *et al.* 2000b; Flannelly *et al.* 1999). Discrepancy in these studies was negligible; consequently, results were interpreted without applying the weighting process.

In the Weighted-scores treatment, contrary to what was anticipated, a large majority of respondents gave scores that added up to ten. The two treatments essentially were the same for non-mobile phone users. However, in the Telecom and Vodafone groups there were a few respondents whose scores failed to add up to ten (Telecom n = 21; Vodafone n = 16). Mann-Whitney tests were executed to see if the mean ranks of the probability scores collected for payment plans were different in the two treatments. Results of these comparisons were mixed with six showing significant differences and the remaining eight showing no significant differences. Hence, it was not possible to extend the investigation to conclusively establish whether scores that added up to ten were more

accurate than scores that did not add up to ten. All the same, from this research the Constant Sum Scale could be recommended because of the clean data that it generated.

9.3.2 Limitations

As mentioned earlier, services provided by Vodafone and Telecom payment plans are consumed regularly and individuals tend to habitually renew their existing payment plans. This appears to be the case for most individuals sampled in this research. Consequently, the number of respondents in the Weighted-scores treatment whose probability scores failed to convey their purchase behaviour was comparatively small. This was contrary to what was anticipated and, hence, comparisons made were not appropriate to achieve the research objective. Analyses were attempted by retaining the irrational responses in the Weighted-scores treatment. Treatment sizes were reduced and imbalanced hence, a non-parametric test was used. Results were mixed and therefore a clear conclusion could not be arrived at in this research. This problem may have been averted if the study was piloted to find out whether the test products were appropriate for achieving the objectives set.

9.3.3 Future Research Direction

The Constant Sum Scale was seen to produce clean probability data that did not require further processing. The results also showed that when respondents were reasonably confident of their behaviour it did not matter which method was used. Results pertinent to the current research objective, however, were not totally conclusive. Hence, the topic remains open for further investigation into the appropriateness of probability data that fails to sum up to one for forecasting purchase behaviour of mutually exclusive items.
For any such investigation in the future, attention ought to be given to the selection of test items. It is imperative that the Weighted-scores treatment has sufficient probability scores that need weighting. It would be best not to use items that are habitually purchased to avoid what happened in the current research. Instead, competing brands of durables might be a better set of test products for this purpose. Most individuals tend to be unsure of such purchases and hence would have to distribute the scores across items to convey their purchase preferences. This could prompt more respondents to distribute scores irrationally, producing sufficient responses for making the desired comparisons.

For most comparisons made in this research, the Constant Sum Scale and the Weighted-scores method produced similar mean probability scores. From this, it could be concluded that both methods produced similar results. All the same, the Constant Sum Scale is recommended because of the clean data produced and it may be beneficial to develop this scale further in the future. Currently, the Constant Sum Scale has been successfully implemented in a face-to-face setting and in Internet based surveys. It is suggested that the scale be developed for implementation in mail and telephone surveys.

In this research, the test products used were frequently used items. The discrepancy caused by irrational assigning of probability scores in the Weighted-scores method was minimal. For this product category, respondents seemed to be confident of their purchase plans, hence, it did not matter which method was used. The research needs to be repeated using other products and behaviour categories (e.g. competing brands, voting behaviour), for which there would be a better spread of probability scores across alternatives. This would result in more discrepancies and hence achieve a robust experimental comparison.

9.3.4 Managerial Implications

In forecasting mutually exclusive behaviours, attempts are made to capture the relative position of these behaviours. Data collected is of ratio level; hence, the behaviours in question can be rank ordered and the distance between them measured. Such forecasting has immense value to managers, as it helps them to identify the relative position of their product in the market. The results of this research suggest that the Constant Sum Scale is a suitable tool for collecting probability data for frequently occurring mutually exclusive behaviours.

10. REFERENCES

1. Aaker DA & Day GS. Marketing Research (4th ed), 1990 Singapore: John Wiley & Sons.
2. AcessNZ. Statistics of visitors to AcessNZ. 1999. [Online][cited on October 1999] Available from Internet: URL: http://www.accessnz.co.nz/statistics/os/
3. Alexrod JN. Attitude measures that predict purchase. *Journal of Advertising,* 1964; 8 (1): 3-17.
4. Aoki K & Elasmar M. Opportunities and challenges of a Web survey: A field experiment. *Proceedings of the 55th Annual Conference of the American Association for Public Opinion Research*, Portland, Oregon, May 18-21, 2000.
5. Armstrong SJ. Research on forecasting: A quarter-century review, 1960-1984. *Interfaces*, 1986; 16 (1): 89-103.
6. Armstrong, JS, Morwitz VG, & Kumar V. Sales forecasts for existing consumer products and services: Do purchase intentions contribute to accuracy? *International Journal of Forecasting,* 2000; 16, 383-397.
7. Askew R, Craighill PM, & Zukin, C. Internet surveys: Fast, easy, cheap, and representative of whom? *Proceedings of the 55th Annual Conference of American Association for Public Opinion Research*, Portland, Oregon, May 18-21, 2000.
8. Bankston K. Caught up in the Web. *Credit Union Management*, 1996; 19 (9): 14-19.
9. Batagelj Z. & Vehovar V. Technical and methodological issues in WWW surveys In: Software and methods for conducting Internet surveys. *Proceedings of the 52nd Annual Conference of the American Association for Public Opinion Research*, St. Louis, 1998.
10. Belson W.A. Validity in Survey Research. *London*: Gower Publishing Co, 1986.
11. Belson WA. The Design and Understanding of Survey Questions. Aldershot, England: Gower Publishing Co, 1981.
12. Berners-Lee T & Connolly D. Hypertext mark-up language - 2.0, 1995. [Online][Cited October 2001] Available from the Internet:URL: http://www.w3.org/MarkUp/html-spec/html-spec_toc.html
13. Berrens RP, Bohara AK, Jenkins-Smith H, Sivia C, & Weimer DL. The advent of Internet survey for political research: A comparison of telephone and Internet

samples. [Online][Cited on March 2002] Available from the Internet: URL: <www.lafollette.wisc.edu/FacultyStaff/Faculty/Weimer/Resources/tellnet.pdf>

14. Bowditch A, Fitall S & Wilde RD. Through the looking glass - Primary research in multi - country forecasting. *Marketing and Research Today,* 1995; Nov: 275-283.
15. Bradburn N & Sudman S. Polls and surveys: Understanding what they tell us. San Francisco: Jossey-Bass, 1998.
16. Bradley N. Sampling for Internet surveys. An examination of respondent selection for Internet research. *International Journal of Market Research*, 1999; 41(4): 387-395.
17. Brennan M & Esslemont D. The accuracy of the Juster Scale for predicting purchase rates of branded, fast-moving consumer goods. *Marketing Bulletin,* 1994a; 5: 47-52.
18. Brennan M & Esslemont D. Estimation of the price-demand relationship: A marketing perspective. *Marketing Bulletin* 1994b; (2): 18-30.
19. Brennan M, Esslemont D & Hini D. Obtaining purchase predictions via telephone interviews. *Journal of the Market Research Society*, 1995a; 37 (3): 241-250.
20. Brennan M, Esslemont D & U C. Using the Juster Scale to estimate the demand - price relationship. *Asia-Australia Marketing Journal,* 1995b; 3(1): 27-37.
21. Brennan M, Rae N & Parackal M. Survey-based experimental research via the Web: Some observations. *Marketing Bulletin*, 1999; 10: 83 – 92.
22. Brennan M. Constructing demand curves from purchase probability data: An application of the Juster scale. *Marketing Bulletin*, 1995; 6: 51-58.
23. Brennan, M. Techniques for improving mail survey response rates. *Marketing Bulletin*, 1992; 3: 24-37.
24. Brucks M. The effect of product class knowledge on information search behaviour. *Journal of Consumer Research*, 1985; 12 (June): 1-15.
25. Brucks M. Search Monitor: An approach for computer-controlled experiments involving consumer information search. *Journal of Consumer Research*, 1988; 15(June): 117-121
26. Bruner GC. The effect of problem recognition style on information search. *Journal of the Academy of Marketing Science;* 1985; 15 (4): 33-41.
27. Burke R. Do you see what I see? The future of virtual shopping. *Academy of Marketing Science,* 1997; 25 (4): 352-360.

28. Burr MA, Levin KY & Becher A. Examining Web vs. Paper mode effects in a federal government customer satisfaction study. *Proceedings of 2001 AAPOR Annual Conference*, Montreal, May 17-20, 2001.
29. Byrnes JC. Consumer intentions to buy. *Journal of Advertising Research,* 1964; 4 (September): 49-51.
30. Carroll NM. Stimulated purchase 'chip' testing. *Marketing And Research Today,* 1989; November: 240-244.
31. Chatman S. (2002). Going beyond the conversion of paper survey forms to web surveys. Student Affairs Online, 2002; 3 (Feb15). [Online][Cited Feb 15 2002] Available from Internet: URL: http://www.studentaffairs.com/ejournal/Winter_2002/surveys.htmls
32. Clancy KJ & Garsen R. Why some scales predict better. *Journal of Advertising Research,* 1970; 10 (5): 33-38.
33. Clawson CJ. How useful are 90-day purchase probabilities? *Journal of Marketing,* 1971; 35(October): 43-47.
34. Claxon JD, Fry JN & Portis B. A taxonomy of pre-purchase information gathering patterms. *Journal of Consumer Research*, 1974; 1(December): 35-42.
35. Converse JM & Presser S. Survey questions: Handcrafting the standardised questionnaire. New Delhi, India: Sage Publications, 1986.
36. Cook WA. Telescoping and memory's other tricks. *Journal of Advertising Research*, 1987; 27: RC5-RC8.
37. Corkindale D & List D. The adoption of Internet market research in Australia. *Proceedings of the Market Research Society of Australia conference*, Adelaide, October 1999.
38. Couper MP, Traugott M & Lamias M. Web survey design and administration. *Public Opinion Quarterly,* 2001; 65 (2): 230–253.
39. Crawford SD, Couper MP & Lamias MJ. Web surveys: Perception of burden. *Social Science Computer Review*, 2001; 19: 146-162.
40. Ely D & Srinivasa V. The predictive power of Internet-based product concept testing using visual depiction and animation. *Journal of Product Innovation Management,* 2000; 17(2): 99- 110.
41. Daly B, Thomson G & Cross, J. Web vs. paper surveys: Lessons from a direct large-scale comparison. Paper presented at the Annual Meeting of California AIR, 2000.

42. Danenberg N & Sharp B. Measuring loyalty in subscription markets using probabilistic estimates of switching behaviour. *Proceedings of the Australia New Zealand Marketing Educators Conference,* 1996a, Auckland.
43. Danenberg N. & Sharp B. Testing Probabilistic Measures of Behaviour as Measures of Customer Loyalty. *Proceedings of the Australian Marketing Educators' Conference*, 1996b Adelaide, Marketing Science Centre, University of South Australia.
44. Danenberg N. & Sharp, B. Examining the predictive ability of two-loyalty segmentation approaches. *Proceedings of the 28th European Marketing Academy Conference*, Berlin, Germany, 1999.
45. Darcy R & Schramm SS. Comment on Kernell. *American Political Science Review*, 1979; 75: 543-545
46. Dawes J. Further evident on the predictive accuracy of the verbal probability scale – The case of household bill payments. *Proceedings of the Australia New Zealand Marketing Educators Conference,* Auckland, 2000.
47. Day D. An examination of the accuracy of two versions of the Juster scale for predicting consumer purchase behaviour using self-completion questionnaires. Unpublished student research report, Department of Marketing, Massey University, 1987.
48. Day D, Gan B, Gendall P & Esslemont D. Predicting purchase behaviour, *Marketing Bulletin*, 1991; 2: 18-30.
49. Deutskens E, Ruyter K, Wetzels M & Oosterveld P. Response rate and response quality of Internet-based surveys: An experimental study, 2003. [Online][Cited on Dec 2003] Available from Internet: URL: http://www.personeel.unimaas.nl/ec.deutskens/content/Response%20Rate%20and%20Response%20Quality%20of%20Online%20Surveys,%20An%20Experimental%20Study.pdf
50. Dillman DA, Phelps G, Tortora R, Swift K, Kohrell J & Berck J. Response rate and measurement differences in mixed mode surveys using mail, telephone, interactive voice response and the Internet. *Proceedings of the 56th Annual Conference of the American Association for Public Opinion Research*, Montreal, Quebec, May 17-20, 2001.

51. Dillman DA, Tortora R, Conradt J & Bowerk D. Influence of plain versus fancy design on response rates for Web surveys. *Proceedings of the Annual Meeting of the American Statistical Association,* Dallas, Texas, 1998.
52. Driver J & Foxall, G How scientific is advertising research? *International Journal of Advertising*, 1986; *5*, 147-160.
53. Duncan OD & Schuman H. Effects of question wording and context: An experiment with religious indicators. *Journal of the American Statistical Association* 1980, 75 (350): 269-275.
54. Duncan OD, Schuman H & Duncan B. Social changes in a metropolitan community, New York: Russell Sage Foundation 1973.
55. Ehrenberg ASC. Repeat Buying Facts: Theory and Applications. New York: Oxford University Press, 1998.
56. Ehrenberg ASC & Uncles MD. Dirichlet-Type Markets, South Bank University and University of New South Wales, London, 1998.
57. Ferber R & Piskie RA. Subjective probabilities and buying intentions. *The Review of Economics and Statistics,* 1965; 47(August): 322-325.
58. Flannelly KJ, Flannelly LT & McLeod MS Jr. Comparison of election predictions, voter preference and candidate choice on political polls. *Journal of the Market Research Society*, 1998; 40 (4): 337-346.
59. Flannelly KJ, Flannelly LT & McLeod MS Jr. Judgement certainty using forced-choice and subjective probability scales. *Proceedings of the Psychonomic Society,* Los Angeles, CA, 20 November 1999.
60. Flannelly KJ, Flannelly LT & McLeod MS Jr. Reducing undecided voters and other sources of error in election survey. *International Journal of Marketing Research*, 2000a; 42 (2): 231-237.
61. Flannelly KJ, Flannelly LT & McLeod, MS Jr. Comparison of forced-choice and subjective probability scales measuring behavioural intentions. *Psychological Reports,* 2000b; 86: 321-332.
62. Gabor A & Granger CWJ. Ownership and acquisition of consumer durables: Report on the Nottingham consumer durable project. *European Journal of Marketing,* 1972; 6(4): 234-248.
63. Gan BC, Esslemont D & Gendall P. A test on the accuracy of the Juster Scale as a predictor of purchase behaviour. *Market Research Centre Report No. 45,* Massey University, ISSN 0110-5426 MRC 600, 1986.

64. Garland R. Estimating customer defection in personal retail banking International *Journal of Bank Marketing*, 2002; 20(7): 317-324.

65. Garner WR. Rating scales, discriminability, and information transmission. *The Psychological Review*, 1960; 67 (6): 343-352.

66. Gatt A & O'Dwyer L. E-commerce technology for business: XML - A panacea for e-business? School of Business Information Technology working paper series, 2/2002 ISBN: 0 86459 197 7, RMIT [Online][Cited April 2002] Available frorn Internet:URL: http://www.businessit.bf.rmit.edu.au/eportal/resources/Gatt_Research.pdf

67. Gendall P. A framework for questionnaire design: Labaw revisited. *Marketing Bulletin*, 1998; 9: 28-39.

68. Gendall P & Hoek J. Question of wording. *Marketing Bulletin*, 1990; 1: 25-36.

69. Gendall P, Esslemont D. & Day D. A comparison of two versions of the Juster Scale using self-completion questionnaires. *Journal of the Marketing Research Society*, 1991; 33(3): 257-263.

70. Golden J, Milewicz J & Herbig P. Forecasting: Trials and tribulations. *Management Decision*, 1994; 32(1): 33-36.

71. Golicic Sl, Davis DF & McCarthy TM, Mentzer JT. Bring order out of chaos: Forecasting e-commerce. *The Journal of Business Forecasting Methods & Systems*, Spring 2001; 20(1): 11-17.

72. Goodman LA. Some multiplicative models for the analysis of cross-classification data. In: L.Le Cam et al. Berkeley: University of California Press editors. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. 1972; 649-696.

73. Gruber A. Purchase intent and purchase probability. *Journal of Advertising Research,* 1970; 10(1): 23-27.

74. Hair JF, Bush RP & Ortinau DJ. Sampling: Theory, designs and issues in marketing research. In: *Marketing Researching: A practical approach for the new millennium.* Published by McGraw –Hill Book co, Singapore. 2000.

75. Hamilton - Gibbs D, Esslemont D & McGuinness D. Predicting the demand for frequently purchased items. *Marketing Bulletin,* 1992; 3: 18-23.

76. Hardie BGS, Fader PS & Wisniewski M. An empirical comparison of new product trial forecasting models. *Journal of Forecasting*, 1998; 17: 209-229

77. Hauser JR, Urban GL & Weinberg, BD. How consumers allocate their time when searching for information. *Journal of Marketing Research*, 1993; 30: 452-466.
78. Heald GI. The relation of intentions to buy consumer durables with levels of purchase. *British Journal of Marketing*, 1970; Summer: 87-97.
79. Hoek J & Gendall P. A new method of predicting voting behaviour. *Journal of the Market Research Society*, 1993; 35: 361-371.
80. Hoek J & Gendall P. A Constant-sum method for determining voting probabilities. *Paper presented at the 52nd Annual Conference of the American Association for Public Opinion Research*, Norfolk, Virginia, May 16-18, 1997a.
81. Hoek J & Gendall P. Factors affecting political poll accuracy: An analysis of undecided respondents. *Marketing Bulletin,* 1997b; 8: 1-11.
82. Huggins V & Eyerman J. Probability based Internet surveys: A synopsis of early methods and survey research results. Federal Committee of Statistical Methods conference, 2001. [Online][Cited Feb 2002] Available from the Internet: URL: www.fcsm.gov/01papers/huggins.pdf.
83. Hyman HH & Sheatsley PB. The current status of American Public opinion In: Payne JC (ed.) The teaching of contemporary affairs. Twenty-first year book of the National Council of Social Studies, 1950; 11-34.
84. Jones R & Pitt N. Health surveys in the workplace: Comparison of postal, email and World Wide Web methods. *Occupational Medicine*, 1999; 49: 556-558.
85. Juster FT. Prediction and consumer buying intentions In: Income, Consumption and Savings, *American Economic Association*, 1964; 604-622.
86. Juster FT. Consumer buying intention and purchase probability. National Bureau of Economic Research, Columbia University Press, 1966.
87. Kalton G & Schuman H. The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society*, 1982; 145 (1), 42-57.
88. Katona G. & Muller E. A study of purchase decision. In: *Consumer Behaviour: The Dynamics of Consumer Reaction*, ed. Clark L.H. New York: New York University, 30-87 1955.
89. Kaye BK & Johnson TJ. Research methodology: Taming the cyber frontier. *Social Science Computer Review*, 1999; 17: 323-337.
90. Keil G.C. & Layton R.A. Dimensions of consumer information seeking behaviour. *Journal of Market Research,* 1981, 13(May): 233-239

91. Killen M. Internet: Global penetration 1996 and forecast for the year 2000. [Online] [Cited on June 1997] Available from the Internet: URL: http://www.killen.com
92. Kingsley P & Anderson T. Market research in an electronic community: Generalizing results from web-based surveys. In: Pelton L & Schnedlitz P editors. *Proceedings of the AMA Marketing Exchange Colloquium*, Vienna, Austria, 1998; 322-331.
93. Klein L.R. & Lansing J.B. Decisions to purchase consumer durable goods. *Journal of Marketing*, 1955, 20 (2), 109, 132.
94. Knoth J. Web survey prompts worldwide response. *Computer Aided Engineering*, 1997; 16 (10): 18-23.
95. Kottler RE. Exploiting the research potential of the World Wide Web. *Paper presented at Research '97*, London, October, 1997a.
96. Kottler RE. Web surveys, the professional way. *Paper presented at ARF Conference*, New York, April, 1997b.
97. Kreuels B. Caveats and options of Internet surveys. *Journal of E-Business,* 2001; 1(1): 24-33.
98. Krotki, K & Dennis JM. Probability-based survey research on the Internet. *Proceedings of the 53rd Conference of the International Statistical Institute*, Seoul Korea, Aug 22-29, 2001.
99. Labaw P. Advanced questionnaire design. Abt. Books: Cambridge MA, 1980.
100. Laswad F. Interpretations of probability expressions by New Zealand standard setters. Accounting Horizons, 1997; 11(4): 16-23
101. Luce RD. Individual Choice Behaviour. New York: John Wiley and Sons, 1959.
102. McDonald & Alpert. Using the Juster Scale to predict adoption of an innovative product. *Proceedings of the Australian and New Zealand Marketing Academy Conference*, Massey University, Auckland, New Zealand, 2001.
103. McFarland S. Effects of question order on survey responses. *Public Opinion Quarterly*, 1981; 16: 381-398.
104. Mehta R & Sivadas E. Comparing response rates and response content in mail versus electronic mail surveys. *Journal of the Market Research Society,* 1995; 37: 429 – 439.
105. Metfessel M. A proposal for quantitative reporting of comparative judgements. *Journal of Psychology*, 1947; 24: 229-235.

106. Microsoft. Active Server Pages. 2002[online] [cited July 2002] Available from Internet: URL:http://msdn.microsoft.com/library/default.asp?url=/library/enus/iisref/html/psdk/asp/iiwawelc.asp

107. Microsoft. About Microsoft COM. 1999[online] [cited July 1999] Available from Internet: URL: http://www.microsoft.com/com/about.asp

108. Ministry of Economic Development. Statistics on Information Technology in New Zealand, 2002; Updated to 2002: Part 2, ISSN 1175-8368.[Online][Cited on August 2002] Available from Internet: URL: http://www.med.govt.nz/pbt/infotech/it-stats/it-stats-2002/part02/index.html

109. Ministry of Economic Development. Statistics on Information Technology in New Zealand, 2003. [Online][Cited on Nov 2003] Available from the Internet: URL: http://www.med.govt.nz/pbt/infotech/it-stats/it-stats-2003/index.html

110. Neisen / Netratings. Reports that nearly 15 Million people worldwide gained Internet access in Q3, 2001. [Online][Cited on March 2002] Available from Internet: URL: http://www.eratings.com/news/2001/20011206.htm

111. Newman JW & Staelin R. Prepurchase information seeking for new cars and major household appliances. *Journal of Marketing Research*, 1972; 9(August): 249-257.

112. Nichols E & Sedivi B. Economic data collection via the Web: A Census Bureau case study. *Proceedings of the section on Survey Research Methods, American Statistical Association,* Alexandria, 1998; 366–371.

113. Nysveen H & Pedersen PE. Search mode and purchase intention in online shopping behavior. *International Journal of Internet Marketing and Advertising*, 2004; 1 (3): (In press).

114. Parackal M & Brennan M. Obtaining purchase probabilities via a Web based survey: Some corrections. *Marketing Bulletin*, 1999; 10: 93 – 101.

115. Peter JP. Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, 1979; 16 (Feb): 6-17.

116. Peterson, RA., & Wilson, WR. Measurring customer satisfaction: Fact & artefact. Journal of the Academy of Marketing Science, 1992; 20(1): 61-71.

117. Pickering, J.F., & Greatorex, M. Evaluations of individual consumer durables: differences between owners and non-owners and buyers and non-buyers. *Journal of the Market Research Society*, 1980, *22* (2), 97-141.

118. Pickering JF & Isherwood BC. Purchase probabilities and consumer durable buying behaviour. *Journal of Market Research Society*, 1974; 16 (3): 203-226.

119. Pitkow JE & Recker MM. Using the web as a survey tool: Results from the second WWW user survey, 1994a. [Online][Cited on October 1999] Available from Internet: URL: http://www.cc.gatech.edu/gvu/user_surveys/User_Survey_Home.html

120. Pitkow JE & Recker MM. Bottom-up-and top-down methodologies for intelligent hypertext. *Proceeding of the third International Conference on Information and Knowledge Management*. Maryland: NIST, 1994b.

121. Punj GN & Staelin R. A model of consumer information search for new automobiles. *Journal of Consumer Research*, 1983; 9: 366-380.

122. Quigley BR, Riemer RA, Cruzen DE & Rosen S. Internet versus paper survey administration: Preliminary finding on response rates. *Proceedings of the 42nd Annual Conference of the International Military Testing Association,* Edinburgh, Scotland, 2000.

123. Rademacher EW & Smith AE. Poll Call. *Public Perspective*, 2001; March/April: 36-37.

124. Rae N & Brennan M. The Relative Effectiveness of Sound and Animation in Web Banner Advertisements. *Marketing Bulletin*, 1998, 9, 76-82.

125. Raggett D, Le Hors A & Ian J. HTML 4.0 Specification. 1997 [Online] [Cited on 17th October 2001] Available from the Internet: URL: http://www.w3.org/TR/WD-html40-970917/cover.html#toc

126. Raggett D. HTML 3.2 Reference Specification. 1997 [Online] {Cited on 17th October 2001] Available from the Internet: URL: http://www.w3.org/TR/REC-html32.html

127. Raggett D, Le Hors A & Ian J. HTML 4.01 Specification. 1999 [Online] [Cited on 17th October 2001] Available from the Internet: URL: http://www.w3.org/TR/html401/

128. Reibstein D.J. Prediction of individual probabilities of brand choice. *Journal of Consumer Research*, 1978, 5(30): 163-168.

129. Reid M & Wood A (2002). Motivating non-donors to donate blood. *Proceeding of the Australia New Zealand Marketing Academic Conference*, Auckland, New Zealand, 2002; 3411-3417.

130. RFL Communications. Harris interactive uses election 2000 to prove its online MR efficacy and accuracy. *Research Business Report*, 2000; Nov: 1-2. [Online] [Cited on Nov 2000] Available from Internet: URL: http://www.imaa.org/events/archive/2001/eoty/Gordon_Black_ResearchBusinessReport1.pdf

131. Riebe E, Lowndes M, Kennedy R & Romaniuk J. Verbal versus ordinal scales in the measurement of service quality and switching estimates. *Proceedings of the Australian and New Zealand Marketing Academy Conference,* University of Otago, New Zealand, 1998.

132. Riebe E, Danenberg N, Sharp B & Rungie, C. Verifying the distribution of probabilistic scales. *Proceeding of the Australian New Zealand Marketing Association Conference*, University of New South Wales, Sydney, Australia Dec 1999.

133. Riebe E. Identifying variations in the accuracy of probabilistic predictions. *Proceedings of the ANZMAC 2000 Visionary Marketing for the 21st Century: Facing the Challenge,* 2000; 1063 –1067.

134. Riquier C, Luxton S & Sharp, B. Probabilistic Segmentation Modelling. *Journal of the Market Research Society*, 1997; 39 (October), 571-587.

135. Rungie C & Danenberg N. Modelling the Juster Scale. *Proceedings of the Australian and New Zealand Marketing Academy Conference,* University of Otago, New Zealand, 1998.

136. Rutkowski T. Internet survey reaches 16.1 million Internet host level: FE growth continues. Biannual Strategic Note, General Magic, Inc., 13102 Weathervane Way, Sunnyvale, CA 94086, 1997.

137. Ryan C & Huyton J. Who is interested in Aboriginal tourism in the northern territory of Australia? A cluster analysis. *Journal of Sustainable Tourism*, 2000; 8(1): 53-88.

138. Ryan C & Huyton J. Dispositions to buy postcards with Aboriginal designs at Uluru-Kata Tjuta National Park. *Journal of Sustainable Tourism*, 1998; 6(3): 254-259.

139. Schillewaert N, Langerak F & Duhamel T. Non-probability sampling for WWW surveys: A comparison of methods. *Journal of the Market Research Society*, 1998; 40: 307-323.

140. Schonlau M, Fricker RD.Jr., Elliott MN. Conducting research surveys via email and the Web, 2002 MR-1480-RC. RAND, Santa Monica [Online] [Cited on Oct 2002] Available from the Internet: URL: http://www.rand.org/publications/MR/MR1480/

141. Fricker RD Jr. & Schonlau M. Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods.* 2002; 14: 347-367.

142. Schuman H & Presser S. Questions and answers in attitude surveys. New York: Academic Press, 1981.

143. Schuman H, Presser S & Ludwig J. Context effects on survey responses to questions about abortion. *Public Opinion Quarterly*, 1981; 45: 216-223.

144. Schuman H, Kalton G, & Ludwig J. Context and contiguity in survey questionnaires. *Public Opinion Quarterly*, 1983; 47: 112-115.

145. Seymour P, Brennan M & Esslemont D. Predicting purchase quantities: Further investigation of the Juster Scale. *Marketing Bulletin,* 1994; 5: 203-226.

146. Sigleman L. Question-order effects on presidential popularity. *Public Opinion Quarterly*, 1981, 45: 199-207

147. Smith TW. Happiness: time trends, seasonal variations, inner survey differences and other mysteries. *Social Psychology Quarterly*, 1979; 42: 18-30.

148. Solomon, DJ. Conducting web-based surveys. *Practical Assessment Research and Evaluation*, 2001; 7. [Online] [Cited Feb 2002] Available from the Internet: URL: http: //paraeonline.net/getvn.asp?v=7&n=19

149. Stapel J (1968). Predictive attitudes In: Adler, L & Crespi, I Attitude research on the rocks. *American Marketing Association,* 1968; 96-115.

150. Statistic New Zealand. Census 2001 [Online] [Cited on Nov 2000] Available from the Internet: URL: http: //www.stats.govt. nz/census.htm

151. Sudman S & Bradburn N. Asking questions: A practical guide to questionnaire design. San Francisco: Jossey-Bass, 1982.

152. Taylor H, Bremer J, Overmeyer C, Siegel JW & Terhanian G. Touchdown! Online polling scores big in November 2000. *Public Perspective*, 2001; (March/April): 38-39.

153. Theil H & Kosobud RF. How informative are consumer buying intentions surveys? *The Review of Economics and Statistics,* 1968; 50-57.

154. Thiel H. On the estimation of relationships involving quantitative variables. *American Journal of Sociology*, 1970; 76: 103-154.

155. Toothacker LE. Multiple comparisons procedures *in Quantitative Applications in the Social Sciences* series #89, Thousand Oaks, CA: Sage Publications, 1993.

156. Tse ACB. (1998) Comparing the response rate, response speed and response quality of two methods of sending questionnaires: Email versus mail. *Journal of the Market Research Society,* 1998; 40: 353–361.

157. Turner CF & Krauss E. Fallible indicators of the subjective state of the nation. *American Psychologist,* 1978; 33: 456-470.

158. Urban GL, Weinberg BD & Hauser JR. Pre-market forecasting of really new products. *Journal of Marketing,* 1996; 60 (January): 47-60.

159. Urban GL, Hauser JR, Qualls WJ & Weinberg BD. Information acceleration: Validity and lessons from the field. *Journal of Marketing Research,* 1997; 34 (1): 143-153.

160. Vehovar V. Batagelj Z, & Lozar K. Web surveys: Can the weighting solve the problem? *Proceedings of the Section on Survey Research Methods. American Statistical Association*, Alexandria, 1999.

161. Vehovar V, Katja L & Batagelj Z. Design issues In WWW Surveys. *Proceedings of the Section on Survey Methodology, 55th Annual Conference of American Association for Public Opinion Research*, Portland, Oregon, USA, May 18-21, 2000.

162. Waddel D & Sohal AS. Forecasting: The key to managerial decision-making. *Management Decision,* 1994; 32 (1); 41-49.

163. Weimann G. The Influential. Albany: State University of New York, 1994.

164. Worcester RM & Burns TR A statistical examination of the relative precision of verbal scales. *Journal of the Market Research Society,* 1975; 17(3): 181-197.

165. Wright M, Lees G, & Garland R. Venture Taranaki Trust: An aerial cableway for Taranaki? A report prepared for Venture Taranaki Trust, 2002.

166. Wright M, Sharp A & Sharp B. Market statistics for the Dirichlet model: Using the Juster scale to replace panel data. *International Journal of Research in Marketing,* 2002; 19(1): 81-90.

Appendix 11.1

Mail outs

Appendix 11.1.1 Vodafone Survey: Cover Letter

Massey University
COLLEGE OF BUSINESS

Department of Marketing
Private Bag 11 222,
Palmerston North,
New Zealand
Telephone: 64 6 350 5593
Facsimile: 64 6 350 2260

Miss Stacey Davenport
56a Connell Street
Blockhouse Bay
Auckland 1007

17th July 2001

Dear Stacey

SURVEY ON WAP-CAPABLE MOBILE PHONES

The Department of Marketing, Massey University is conducting a survey in conjunction with Vodafone New Zealand. Your name has been selected at random from Vodafone's client list.

We wish to ask you, a few questions about your views on WAP-capable mobile phones and about some payment options available from Vodafone. We would be grateful if you could help us by answering a short questionnaire at this web site:

Http://survey.massey.ac.nz

You will need to log in as **Sdavenport** with the access code **C1566**. It will take about 9 minutes to answer all the questions; however, this time could be more or less depending on your Internet connection. Full information about the project is available on the Web site.

The survey is completely confidential, and is being carried out under the Code of Practice of the Market Research Society of New Zealand. It has been reviewed and approved by the Massey University Human Ethics Committee, PN Protocol 01/1. If you have any questions, please contact me or Andrew Stevenson of Vodafone New Zealand. My phone number is 06 353 5580 and my email address is M.K.Parackal@massey.ac.nz and Andrew's phone number is 09 357 5100 and his email address is Andrew.Stevenson@vodafone.co.nz.

Thank you in advance for your co-operation. We look forward to receiving your completed questionnaire.

Yours sincerely,

Mathew Parackal
Department of Marketing
Massey University

Appendix 11.1.2 New Zealand Survey: Cover Letter

Massey University
COLLEGE OF BUSINESS

Department of Marketing
Private Bag 11 222
Palmerston North
New Zealand
Telephone: 64 6 350 5593
Facsimile: 64 6 350 2260

3948
Miss Raylene Gaye Laverty
11 Ethel Street
Invercargill 9501

12 September 2001

Dear Raylene

SURVEY ON WAP-CAPABLE MOBILE PHONES

The Marketing Department at Massey University has, over a number of years, been developing a method for forecasting sales. The method involves asking people to assess the probability that they will buy a product within some specified time period. So far the research has found this to be a reasonably accurate way of forecasting, and we are continuing to develop the method.

As part of the requirements for my Ph.D. thesis, I am applying this method to forecast the rate of adoption of WAP-capable mobile phones. I am also trying to use the method to forecast the switching between some payment options available from mobile phone service providers. Your name has been selected at random from the electoral roll and I am writing to request your help with this research.

I would be grateful if you could help me by answering a short questionnaire at this Web site:

Http://survey.massey.ac.nz

You will need to log in as RLaverty with the personal identity number 3948. It will take about 14 minutes to answer all the questions; however, this time could be more or less depending on your Internet connection. Full information about the project is available on the web site.

If you do not have access to the Web, please use the reply paid post card enclosed with this letter to request a hard copy of the questionnaire.

The survey is completely confidential, and will only be reported in an aggregate form (e.g. 30% of the sample use mobile phones). It is being carried out under the Code of Practice of the Market Research Society of New Zealand, and has been reviewed and approved by the Massey University Human Ethics Committee, PN Protocol 01/1. My supervisors are Associate Professor Tony Lewis (A.Lewis@massey.ac.nz, or 06 350 5588) and Dr Ron Garland (B.R.Garland@massey.ac.nz or 06 350 5581); both my supervisors and I will be glad to answer any questions you may have. You can reach me at 06 350 5580, or email me at M.K.Parackal@massey.ac.nz.

I hope that you will be able to help me by taking part in the survey. Thank you in advance for your help. I look forward to receiving your completed questionnaire.

Yours sincerely,

Mathew Parackal

Te Kunenga ki Pūrehuroa

Appendix 11.1.3 New Zealand Survey: First Reminder Letter

Massey University

Department of Marketing
Private Bag 11 222,
Palmerston North,
New Zealand
Telephone: 64 6 350 5583
Facsimile: 64 6 350 2260

2872
Mrs Juanita Zillah Geange
30 A Argyle Avenue
Levin 5500

31 October 2001

Dear Juanita

SURVEY ON WAP-CAPABLE MOBILE PHONES

Six weeks ago I contacted you requesting your help with my survey research. As I have not received the completed questionnaire from you I'm writing again to remind you of the survey and request your help.

The Marketing Department, Massey University has over a number of years been developing a method of forecasting sales. The method involves asking people to assess the probability that they will buy a product within some specified time period. So far the research has found this to be a reasonably accurate way of forecasting, and we are continuing to develop the method.

As part of the requirements for my Ph.D. thesis, I am applying this method to forecast the adoption of the WAP-capable mobile phone. I am also trying to use the method to forecast the switching between some payment options available from mobile phone service providers. Your name has been selected at random from the electoral roll and I am writing to request your help with this research.

I would be grateful if you could help me by answering a short questionnaire at this Web site:

Http://survey.massey.ac.nz

You will need to log in as **JGeange** with the personal identity number **2872**. It will take about 14 minutes to answer all the questions. Full information about the project is available on the web site.

If you do not have access to the Web, please use the reply paid post card enclosed with this letter to request for a paper copy of the questionnaire.

The survey is completely confidential, and will only be reported in an aggregate form (e.g. 30% of the sample use mobile phones). It is being carried out under the Code of Practice of the Market Research Society of New Zealand, and has been reviewed and approved by the Massey University Human Ethics Committee, PN Protocol 01/1. My supervisors are Associate Professor Tony Lewis (A.Lewis@massey.ac.nz, or 06 350 5588) and Dr Ron Garland (B.R.Garland@massey.ac.nz or 06 350 5581); both my supervisors and I will be glad to answer any questions you may have. You can reach me at 06 350 5580, or email me at M.K.Parackal@massey.ac.nz.

I hope that you will be able to help me by taking part in the survey. Thank you in advance for your help. I look forward to receiving your completed questionnaire.

Yours sincerely,

Mathew Parackal

Te Kunenga ki Purehuroa
Inception to Infinity: Massey University's commitment to learning as a life-long journey

Appendix 11.1.4 New Zealand Survey: Second Reminder Letter

Massey University

Department of Marketing
Private Bag 11 222,
Palmerston North,
New Zealand
Telephone: 64 6 350 5580
Facsimile: 646 350 2260

2872
Mrs Juanita Zillah Geange
30 A Argyle Avenue
Levin 5500

31 October 2001

Dear Juanita

SURVEY ON WAP-CAPABLE MOBILE PHONES

Six weeks ago I contacted you requesting your help with my survey research. As I have not received the completed questionnaire from you I'm writing again to remind you of the survey and request your help.

The Marketing Department, Massey University has over a number of years been developing a method of forecasting sales. The method involves asking people to assess the probability that they will buy a product within some specified time period. So far the research has found this to be a reasonably accurate way of forecasting, and we are continuing to develop the method.

As part of the requirements for my Ph.D. thesis, I am applying this method to forecast the adoption of the WAP-capable mobile phone. I am also trying to use the method to forecast the switching between some payment options available from mobile phone service providers. Your name has been selected at random from the electoral roll and I am writing to request your help with this research.

I would be grateful if you could help me by answering a short questionnaire at this Web site:

Http://survey.massey.ac.nz

You will need to log in as JGeange with the personal identity number 2872. It will take about 14 minutes to answer all the questions. Full information about the project is available on the web site.

If you do not have access to the Web, please use the reply paid post card enclosed with this letter to request for a paper copy of the questionnaire.

The survey is completely confidential, and will only be reported in an aggregate form (e.g. 30% of the sample use mobile phones). It is being carried out under the Code of Practice of the Market Research Society of New Zealand, and has been reviewed and approved by the Massey University Human Ethics Committee, PN Protocol 01/1. My supervisors are Associate Professor Tony Lewis (A.Lewis@massey.ac.nz, or 06 350 5588) and Dr Ron Garland (R.R.Garland@massey.ac.nz or 06 350 5581): both my supervisors and I will be glad to answer any questions you may have. You can reach me at 06 350 5580, or email me at M.K.Parackal@massey.ac.nz.

I hope that you will be able to help me by taking part in the survey. Thank you in advance for your help. I look forward to receiving your completed questionnaire.

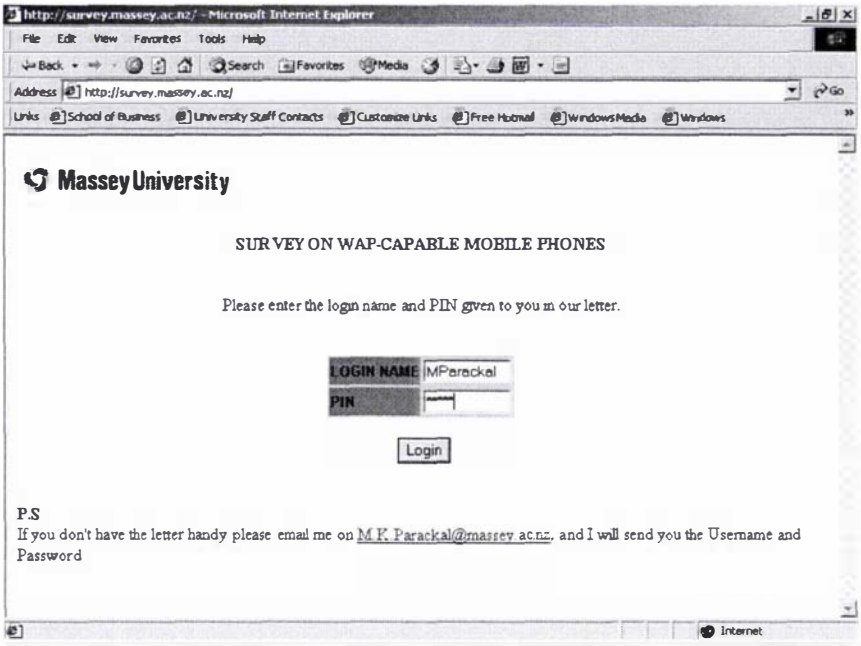
Yours sincerely,

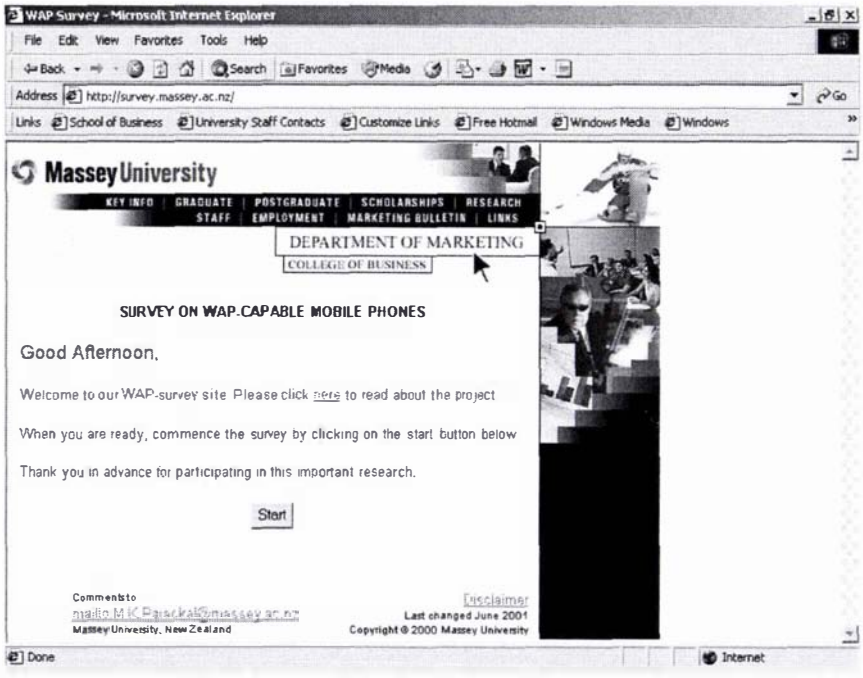
Mathew Parackal

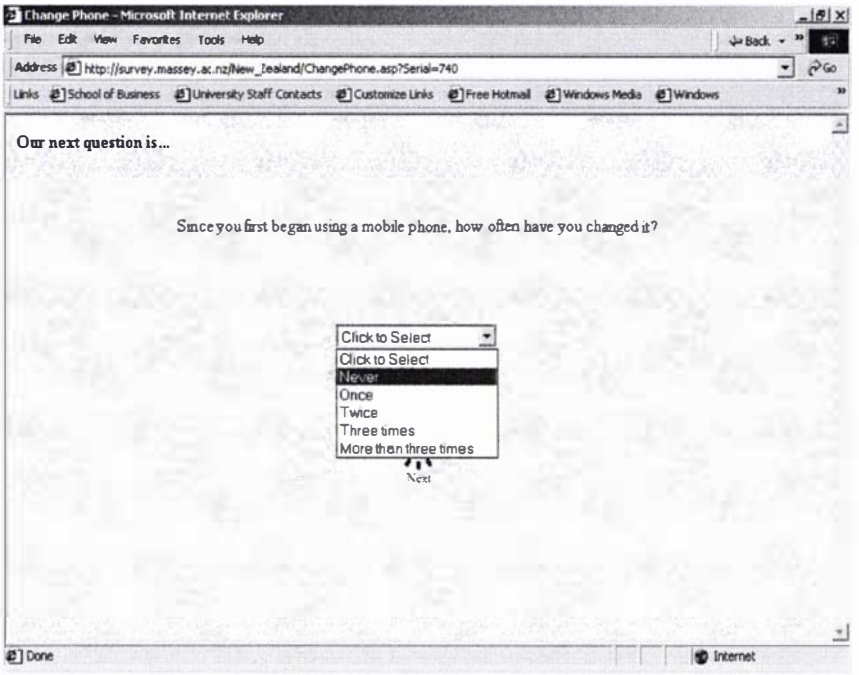
Te Kunenga ki Pureluiroa

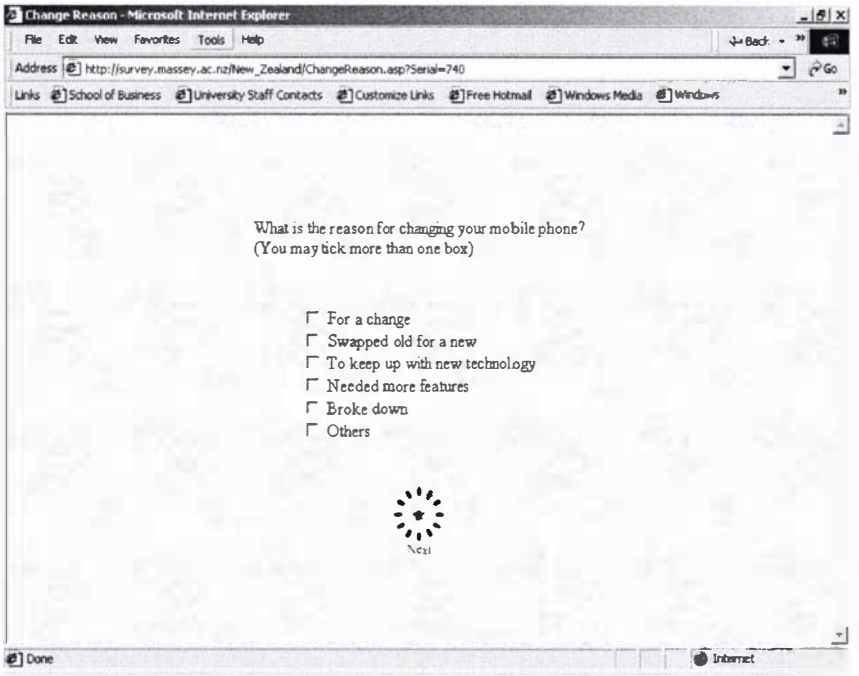
Appendix 11.2

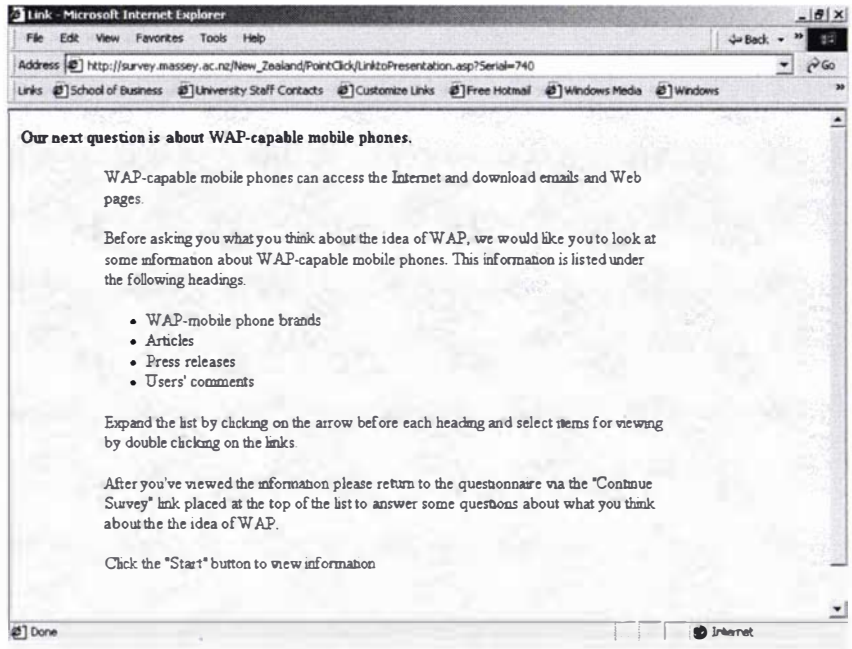
Questionnaire

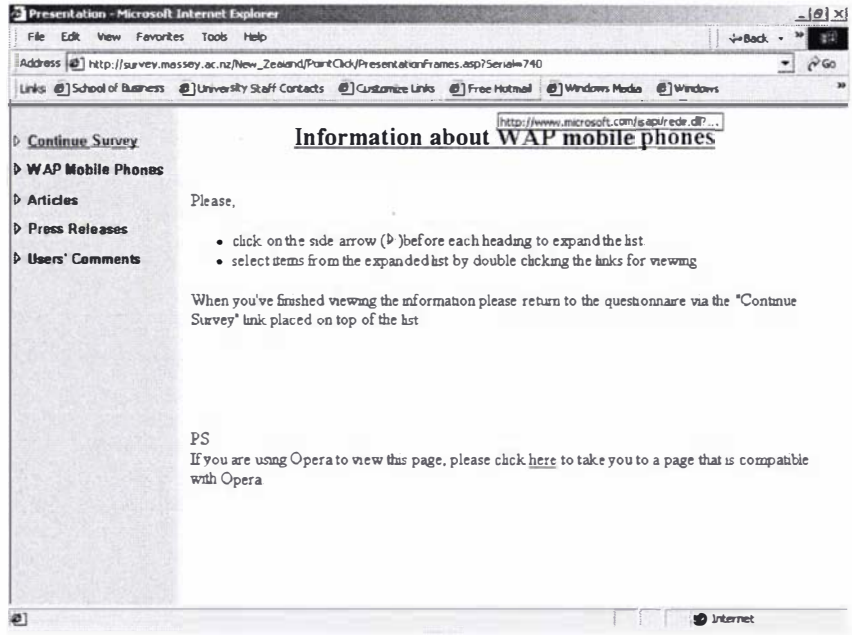

http://survey.massey.ac.nz/ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://survey.massey.ac.nz/
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Massey University
SURVEY ON WAP-CAPABLE MOBILE PHONES
Please enter the login name and PIN given to you in our letter.
LOGIN NAME MParackal
PIN
Login
P.S
If you don't have the letter handy please email me on M.K.Parackal@massey.ac.nz, and I will send you the Username and Password
Internet

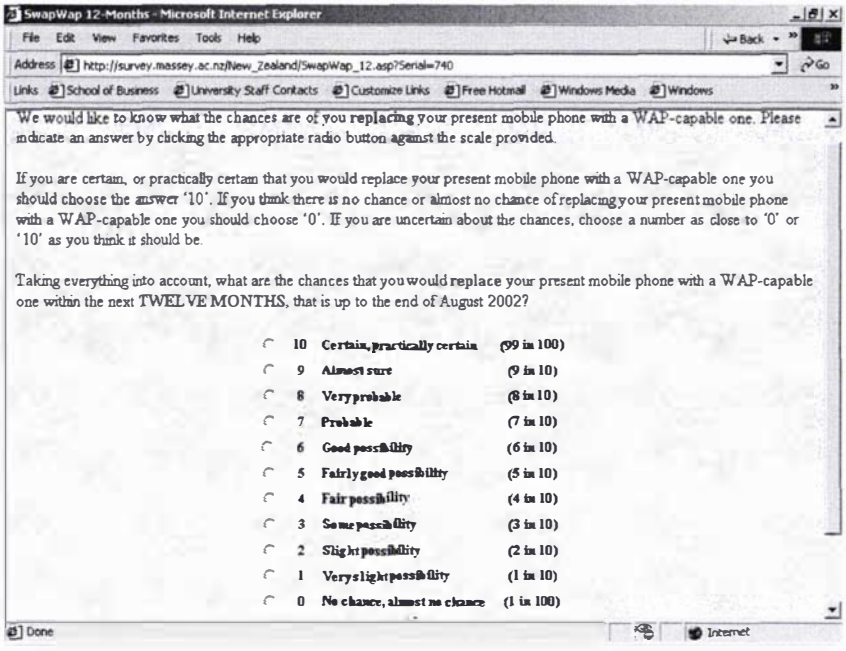

WAP Survey - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://survey.massey.ac.nz/
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Massey University
KEY INFO | GRADUATE | POSTGRADUATE | SCHOLARSHIPS | RESEARCH
STAFF | EMPLOYMENT | MARKETING BULLETIN | LINKS
DEPARTMENT OF MARKETING
COLLEGE OF BUSINESS
SURVEY ON WAP-CAPABLE MOBILE PHONES
Good Afternoon,
Welcome to our WAP-survey site. Please click here to read about the project
When you are ready, commence the survey by clicking on the start button below
Thank you in advance for participating in this important research.
Start
Comments to
mailto:M.K.Parackal@massey.ac.nz
Massey University, New Zealand
Disclaimer
Last changed June 2001
Copyright © 2000 Massey University
Done
Internet

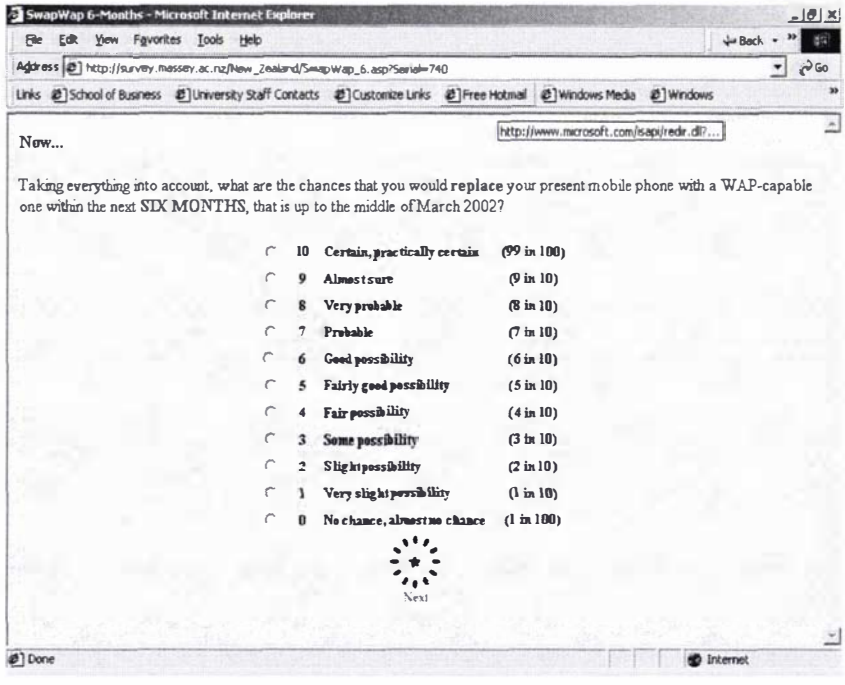

Change Phone - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/ChangePhone.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Our next question is...
Since you first began using a mobile phone, how often have you changed it?
Click to Select
Click to Select
Never
Once
Twice
Three times
More than three times
Next
Done
Internet

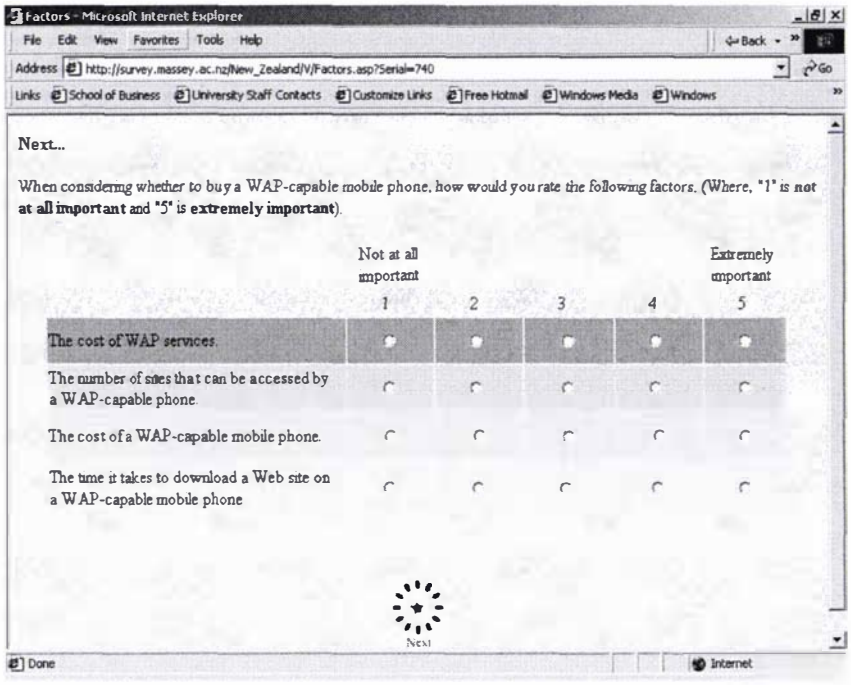

Change Reason - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/ChangeReason.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
What is the reason for changing your mobile phone?
(You may tick more than one box)
For a change
Swapped old for a new
To keep up with new technology
Needed more features
Broke down
Others
Next
Done
Internet

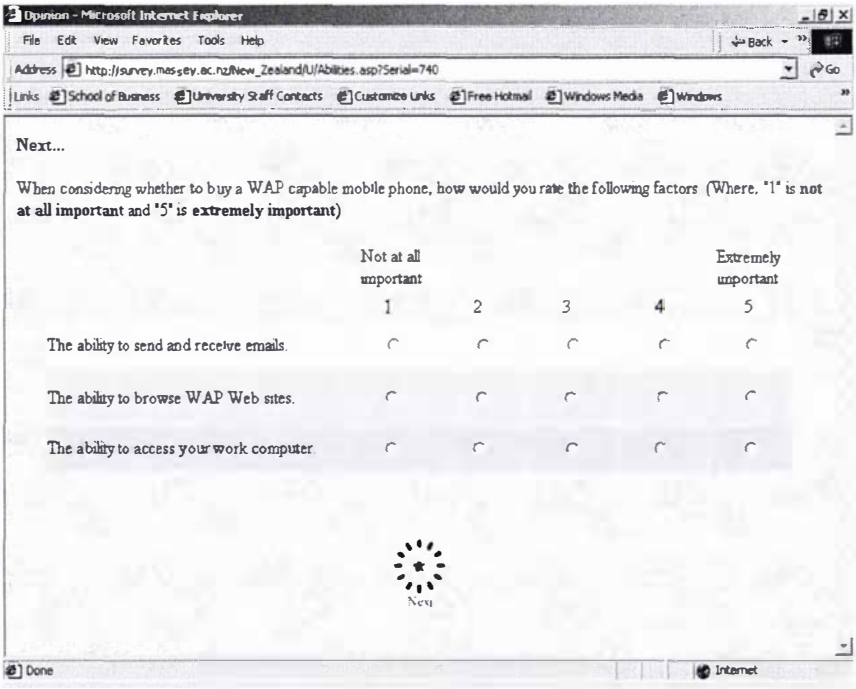

Link - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/PointClick/LinktoPresentation.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Our next question is about WAP-capable mobile phones.
WAP-capable mobile phones can access the Internet and download emails and Web pages.
Before asking you what you think about the idea of WAP, we would like you to look at some information about WAP-capable mobile phones. This information is listed under the following headings.
WAP-mobile phone brands
Articles
Press releases
Users' comments
Expand the list by clicking on the arrow before each heading and select items for viewing by double clicking on the links.
After you've viewed the information please return to the questionnaire via the "Continue Survey" link placed at the top of the list to answer some questions about what you think about the the idea of WAP.
Click the "Start" button to view information
Done
Internet

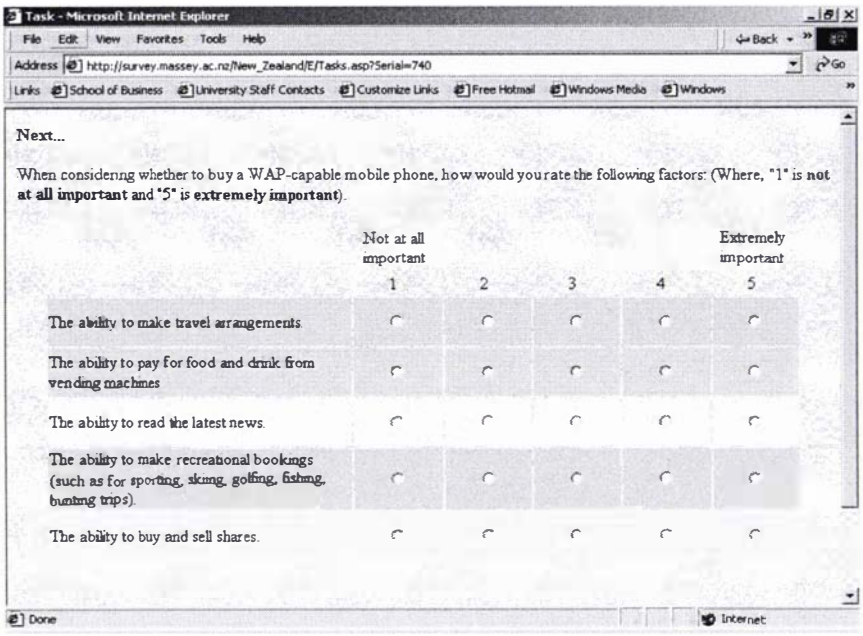

Presentation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/PointClick/PresentationFrames.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Continue Survey
WAP Mobile Phones
Articles
Press Releases
Users' Comments
Information about WAP mobile phones
Please,
click on the side arrow (▷)before each heading to expand the list
select items from the expanded list by double clicking the links for viewing
When you've finished viewing the information please return to the questionnaire via the "Continue Survey" link placed on top of the list
PS
If you are using Opera to view this page, please click here to take you to a page that is compatible with Opera
Internet

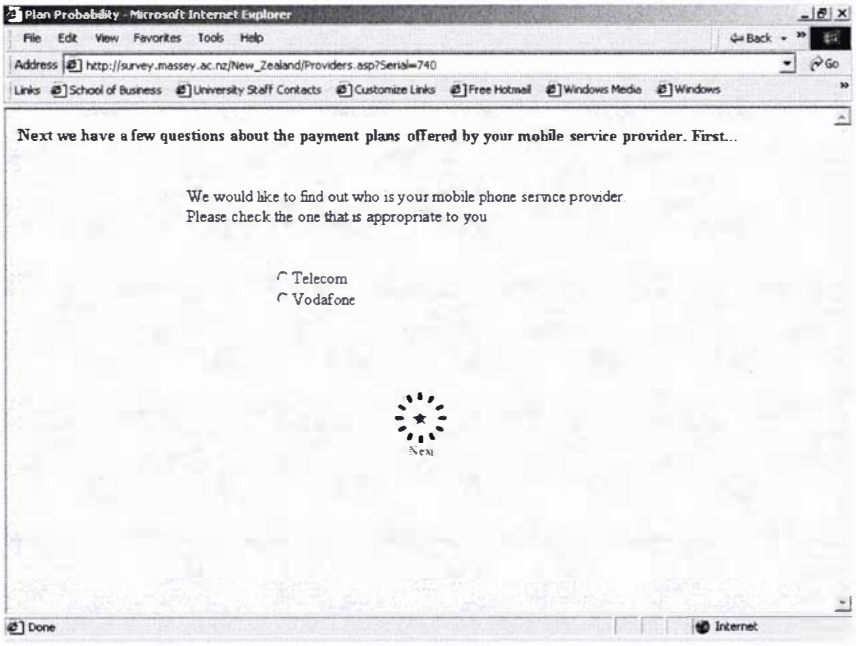

SwapWap 12-Months - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/SwapWap_12.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
We would like to know what the chances are of you replacing your present mobile phone with a WAP-capable one. Please indicate an answer by clicking the appropriate radio button against the scale provided.
If you are certain, or practically certain that you would replace your present mobile phone with a WAP-capable one you should choose the answer '10'. If you think there is no chance or almost no chance of replacing your present mobile phone with a WAP-capable one you should choose '0'. If you are uncertain about the chances, choose a number as close to '0' or '10' as you think it should be.
Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next TWELVE MONTHS, that is up to the end of August 2002?
10 Certain, practically certain (99 in 100)
9 Almost sure (9 in 10)
8 Very probable (8 in 10)
7 Probable (7 in 10)
6 Good possibility (6 in 10)
5 Fairly good possibility (5 in 10)
4 Fair possibility (4 in 10)
3 Some possibility (3 in 10)
2 Slight possibility (2 in 10)
1 Very slight possibility (1 in 10)
0 No chance, almost no chance (1 in 100)
Done
Internet

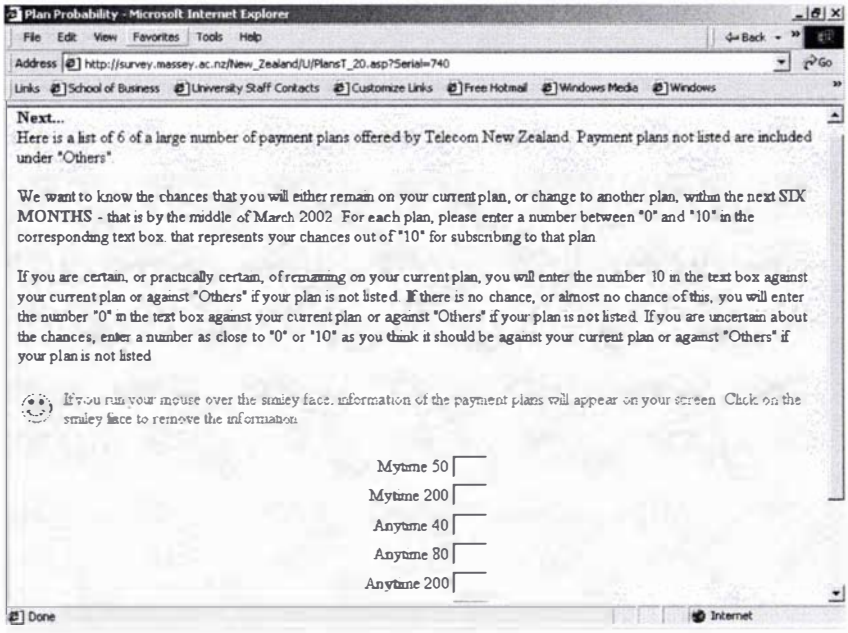

SwapWap 6-Months - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/SwapWap_6.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
http://www.microsoft.com/isapi/redir.dll?...
Now...
Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next SIX MONTHS, that is up to the middle of March 2002?
10 Certain, practically certain (99 in 100)
9 Almost sure (9 in 10)
8 Very probable (8 in 10)
7 Probable (7 in 10)
6 Good possibility (6 in 10)
5 Fairly good possibility (5 in 10)
4 Fair possibility (4 in 10)
3 Some possibility (3 in 10)
2 Slight possibility (2 in 10)
1 Very slight possibility (1 in 10)
0 No chance, almost no chance (1 in 100)
Next
Done
Internet

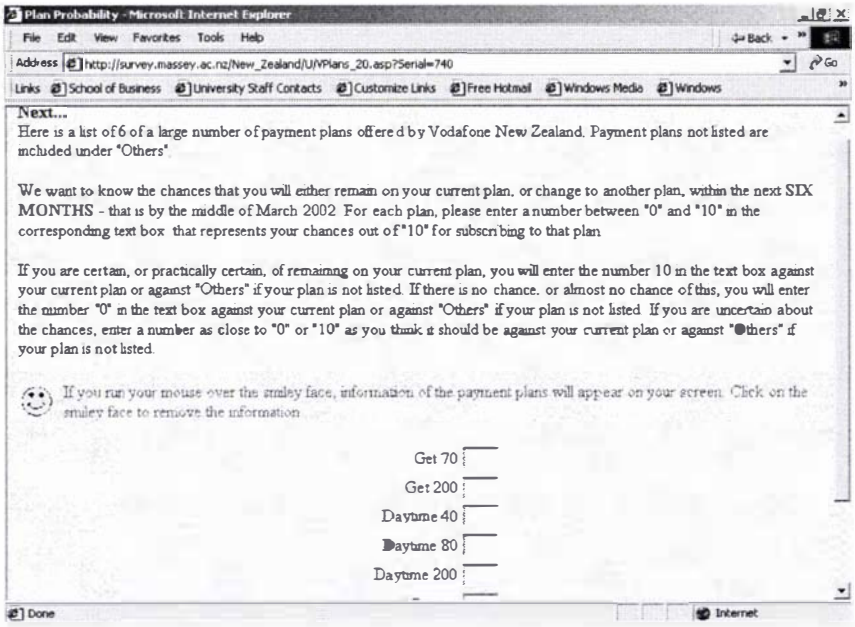

Factors - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/V/Factors.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Next...
When considering whether to buy a WAP-capable mobile phone, how would you rate the following factors. (Where, "1" is not at all important and "5" is extremely important).
Not at all important
Extremely important
1 2 3 4 5
The cost of WAP services.
The number of sites that can be accessed by a WAP-capable phone.
The cost of a WAP-capable mobile phone.
The time it takes to download a Web site on a WAP-capable mobile phone.
Next
Done
Internet

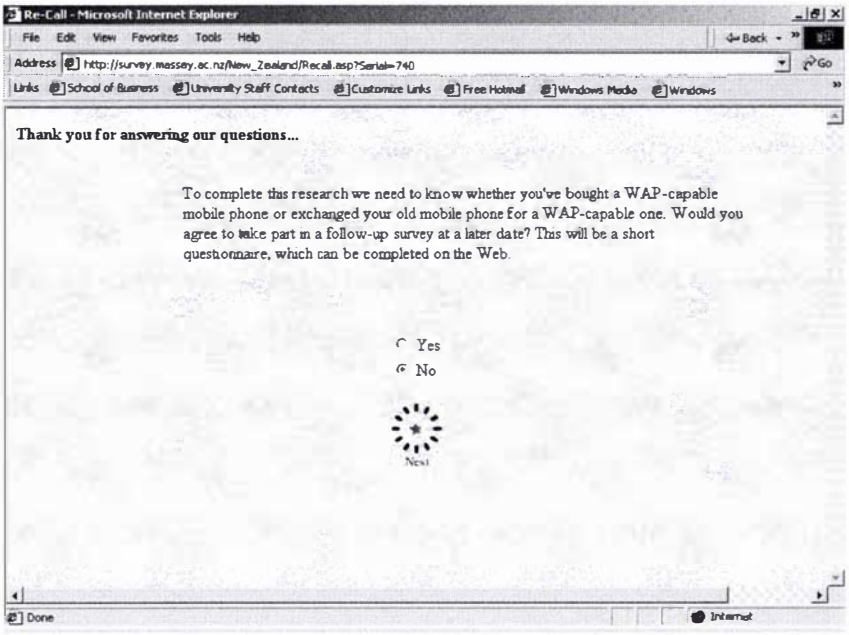

Opinion - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/U/Abilities.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Next...
When considering whether to buy a WAP capable mobile phone, how would you rate the following factors (Where, "1" is not at all important and "5" is extremely important)
Not at all important
Extremely important
1 2 3 4 5
The ability to send and receive emails.
The ability to browse WAP Web sites.
The ability to access your work computer.
Next
Done
Internet

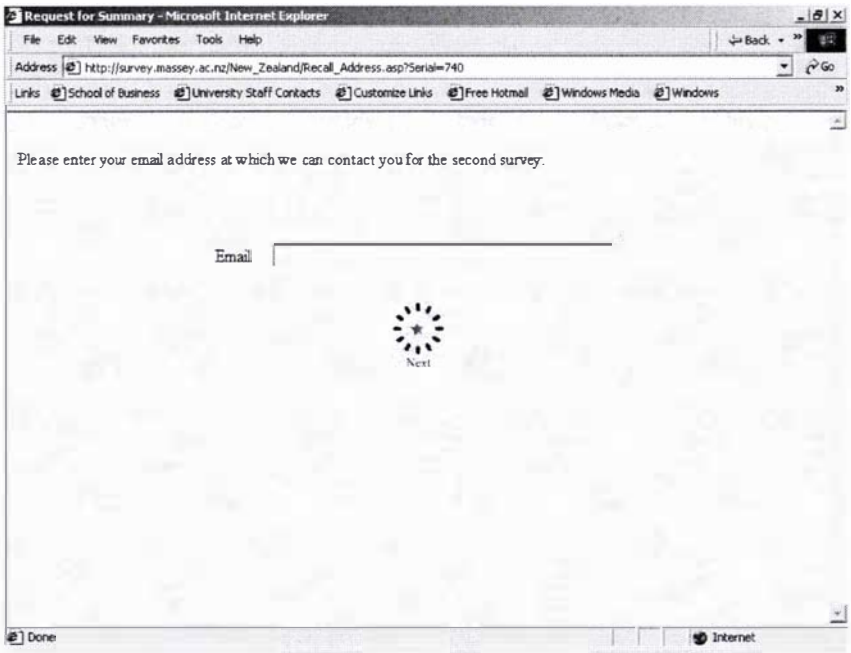

Task - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/E/Tasks.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Next...
When considering whether to buy a WAP-capable mobile phone, how would you rate the following factors: (Where, "1" is not at all important and "5" is extremely important).
Not at all important
Extremely important
1 2 3 4 5
The ability to make travel arrangements.
The ability to pay for food and drink from vending machines
The ability to read the latest news.
The ability to make recreational bookings (such as for sporting, skiing, golfing, fishing, hunting trips).
The ability to buy and sell shares.
Done
Internet

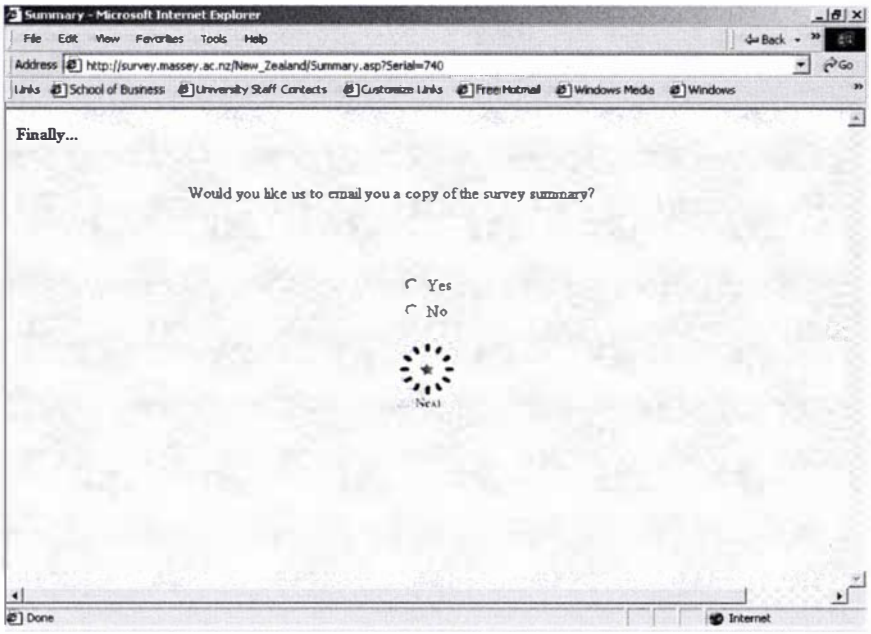

Plan Probability - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/Providers.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Next we have a few questions about the payment plans offered by your mobile service provider. First...
We would like to find out who is your mobile phone service provider.
Please check the one that is appropriate to you
Telecom
Vodafone
Next
Done
Internet

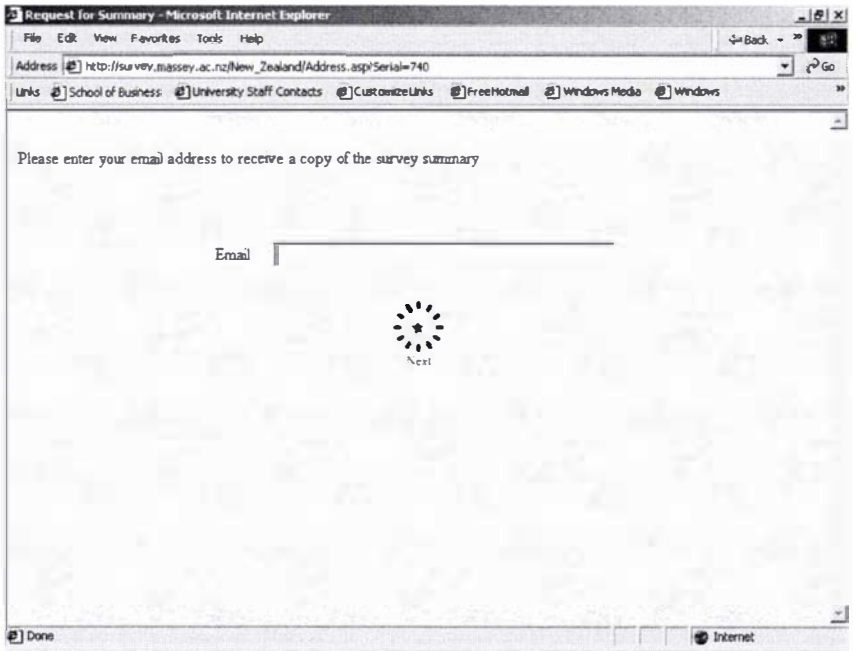

Plan Probability - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/U/PlansT_20.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Next...
Here is a list of 6 of a large number of payment plans offered by Telecom New Zealand. Payment plans not listed are included under "Others".
We want to know the chances that you will either remain on your current plan, or change to another plan, within the next SIX MONTHS - that is by the middle of March 2002. For each plan, please enter a number between "0" and "10" in the corresponding text box that represents your chances out of "10" for subscribing to that plan.
If you are certain, or practically certain, of remaining on your current plan, you will enter the number 10 in the text box against your current plan or against "Others" if your plan is not listed. If there is no chance, or almost no chance of this, you will enter the number "0" in the text box against your current plan or against "Others" if your plan is not listed. If you are uncertain about the chances, enter a number as close to "0" or "10" as you think it should be against your current plan or against "Others" if your plan is not listed.
If you run your mouse over the smiley face, information of the payment plans will appear on your screen. Click on the smiley face to remove the information.
Mytime 50
Mytime 200
Anytime 40
Anytime 80
Anytime 200
Done
Internet


Plan Probability - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/U/Plans_20.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Next...
Here is a list of 6 of a large number of payment plans offered by Vodafone New Zealand. Payment plans not listed are included under "Others".
We want to know the chances that you will either remain on your current plan, or change to another plan, within the next SIX MONTHS - that is by the middle of March 2002. For each plan, please enter a number between "0" and "10" in the corresponding text box that represents your chances out of "10" for subscribing to that plan.
If you are certain, or practically certain, of remaining on your current plan, you will enter the number 10 in the text box against your current plan or against "Others" if your plan is not listed. If there is no chance, or almost no chance of this, you will enter the number "0" in the text box against your current plan or against "Others" if your plan is not listed. If you are uncertain about the chances, enter a number as close to "0" or "10" as you think it should be against your current plan or against "Others" if your plan is not listed.
If you run your mouse over the smiley face, information of the payment plans will appear on your screen. Click on the smiley face to remove the information.
Get 70
Get 200
Daytime 40
Daytime 80
Daytime 200
Done
Internet


Re-Call - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/Recall.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Thank you for answering our questions...
To complete this research we need to know whether you've bought a WAP-capable mobile phone or exchanged your old mobile phone for a WAP-capable one. Would you agree to take part in a follow-up survey at a later date? This will be a short questionnaire, which can be completed on the Web.
Yes
No
Next
Done
Internet


Request for Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/Recall_Address.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Please enter your email address at which we can contact you for the second survey.
Email
Next
Done
Internet


Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/Summary.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Finally...
Would you like us to email you a copy of the survey summary?
Yes
No
Next
Done
Internet

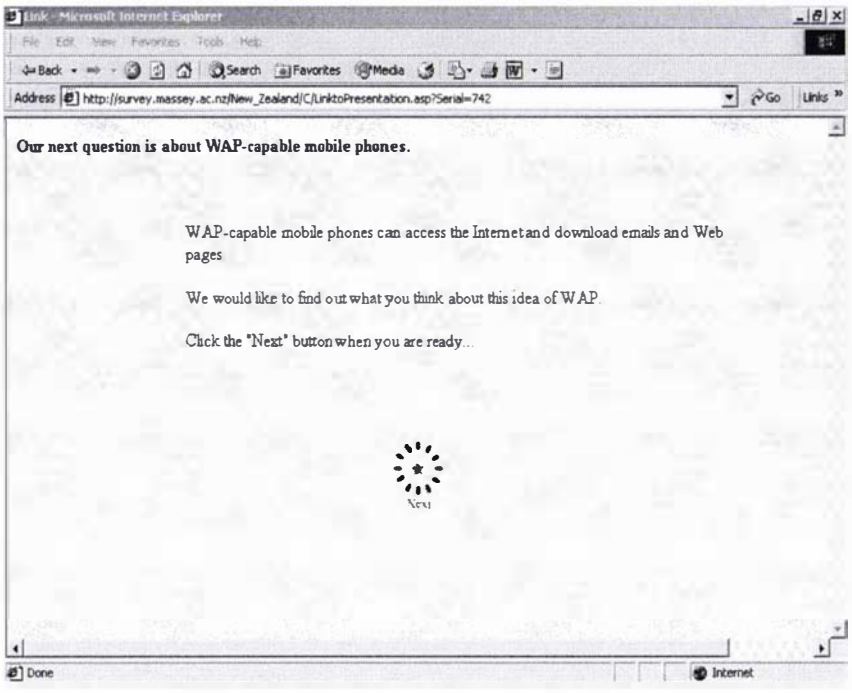

Request for Summary - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/Address.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Please enter your email address to receive a copy of the survey summary
Email
Next
Done
Internet

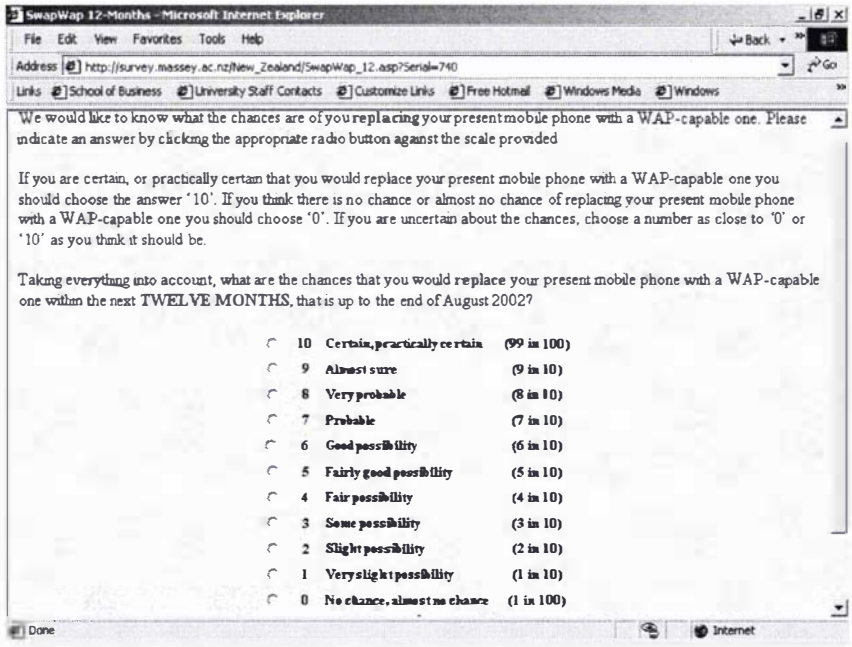

Thank You - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/ThankYou.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
Thank you for participating in our survey. Your help in carrying out this research is much appreciated.
You can now safely close this Web Browser or continue to another site.
Done
Internet

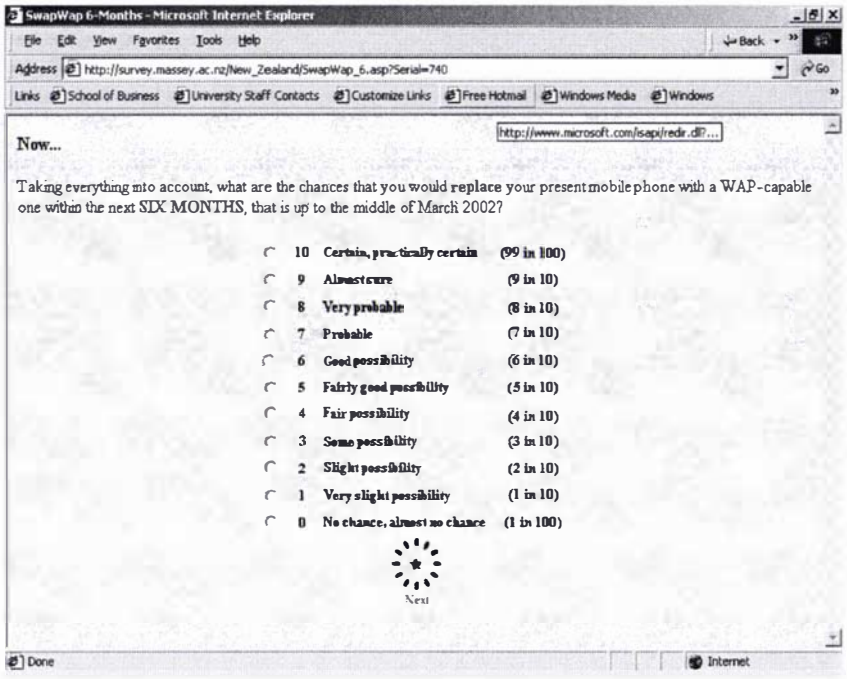
Appendix 11.3

Contextual Treatments

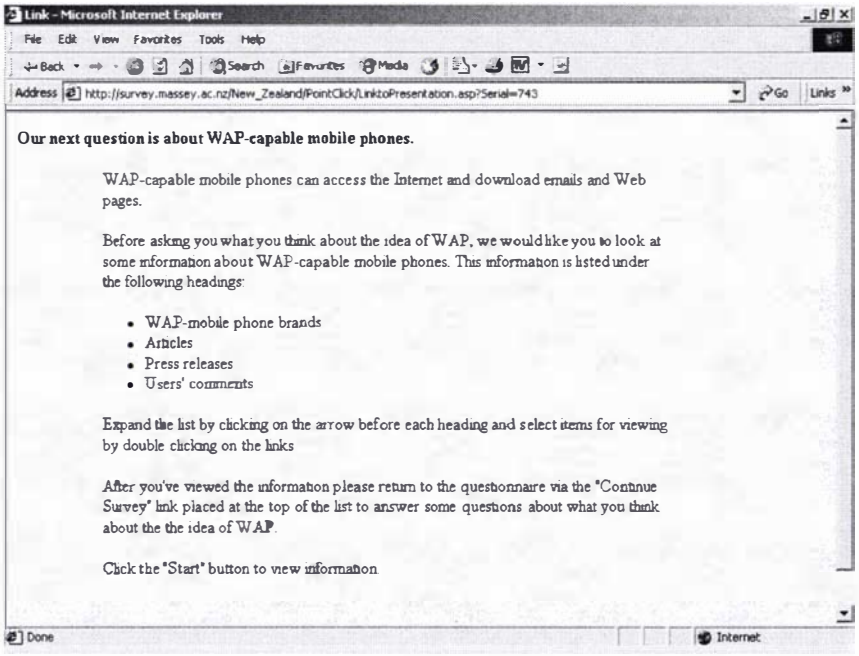
Appendix 11.3.1 Standard Treatment

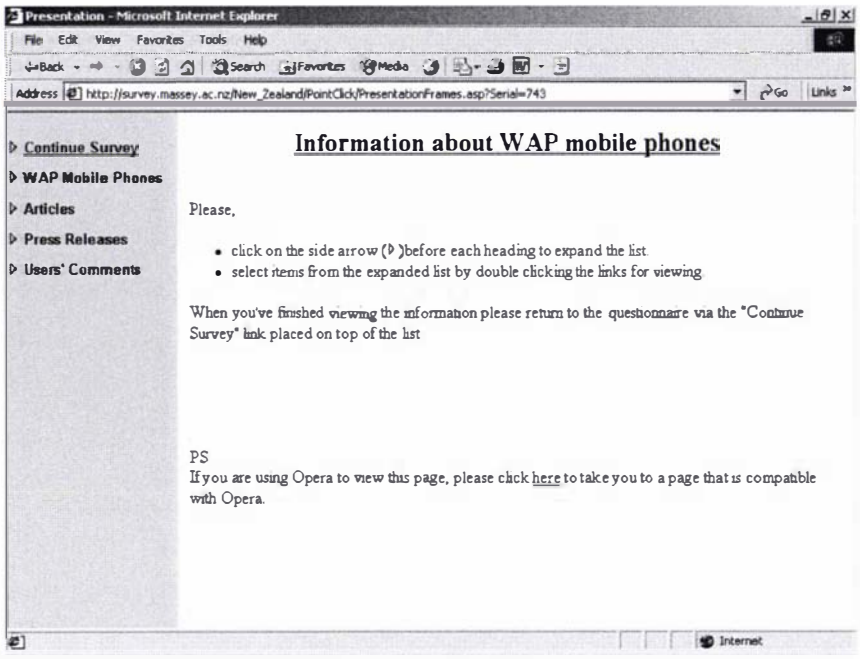


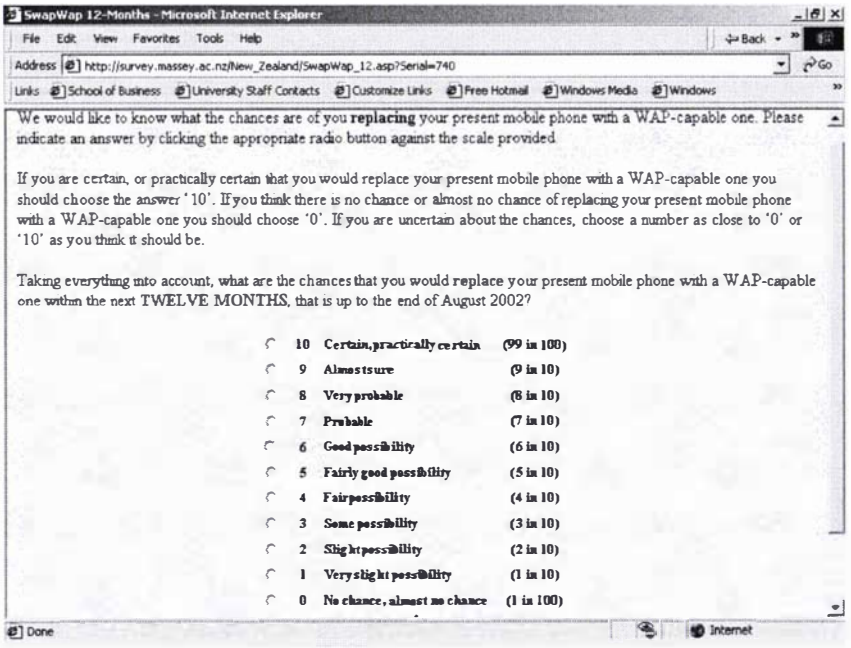


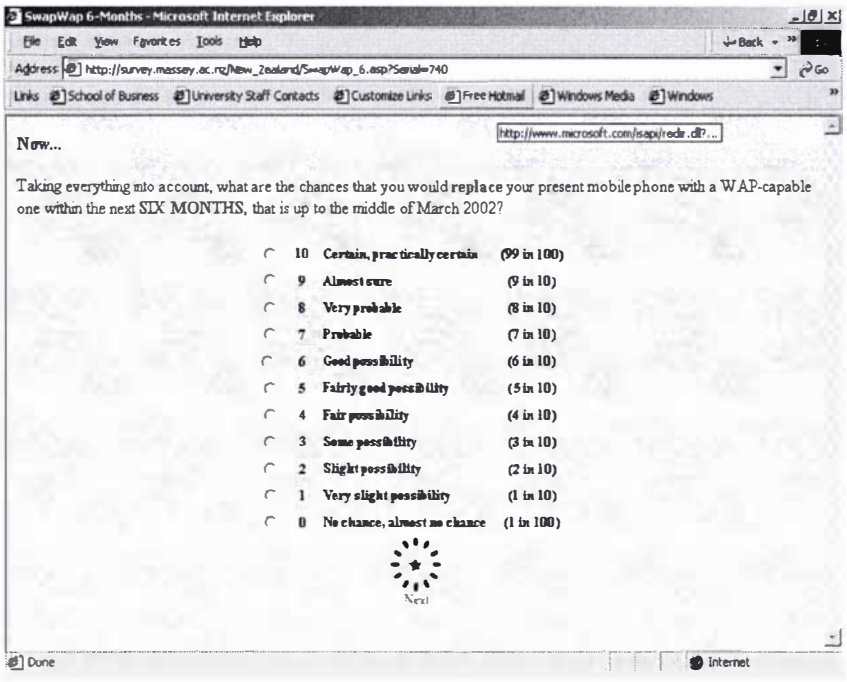


Appendix 11.3.2 Point & Click Treatment

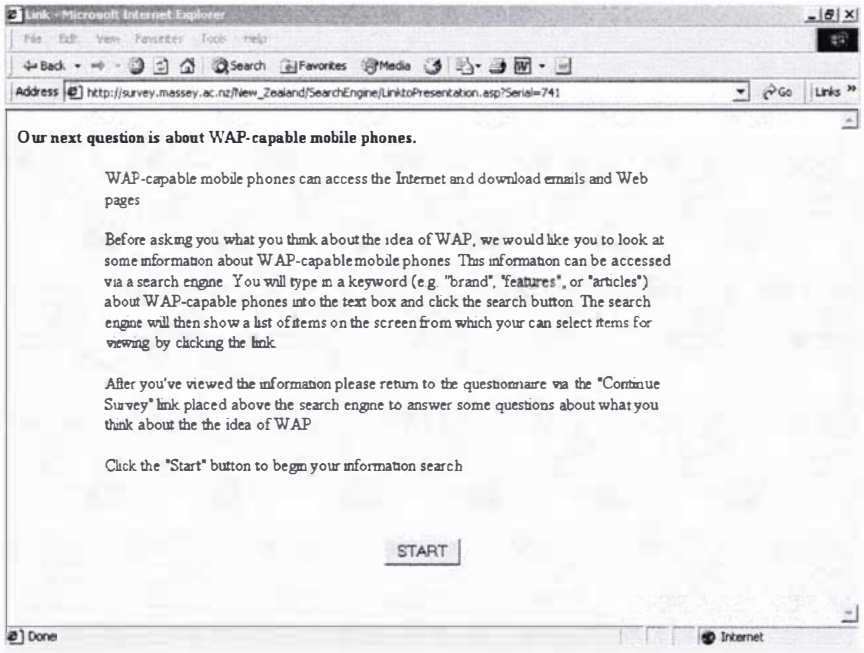


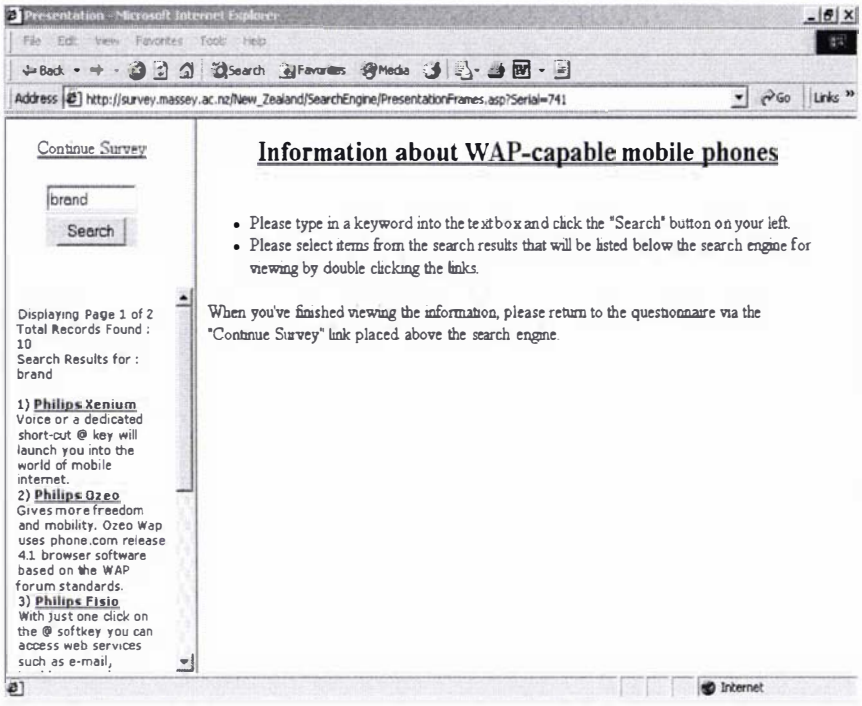

Presentation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://survey.massey.ac.nz/New_Zealand/PointClick/PresentationFrames.asp?Serial=743
Continue Survey
WAP Mobile Phones
Articles
Press Releases
Users' Comments
Information about WAP mobile phones
Please,
click on the side arrow (▷)before each heading to expand the list.
select items from the expanded list by double clicking the links for viewing.
When you've finished viewing the information please return to the questionnaire via the "Continue Survey" link placed on top of the list.
PS
If you are using Opera to view this page, please click here to take you to a page that is compatible with Opera.
Internet

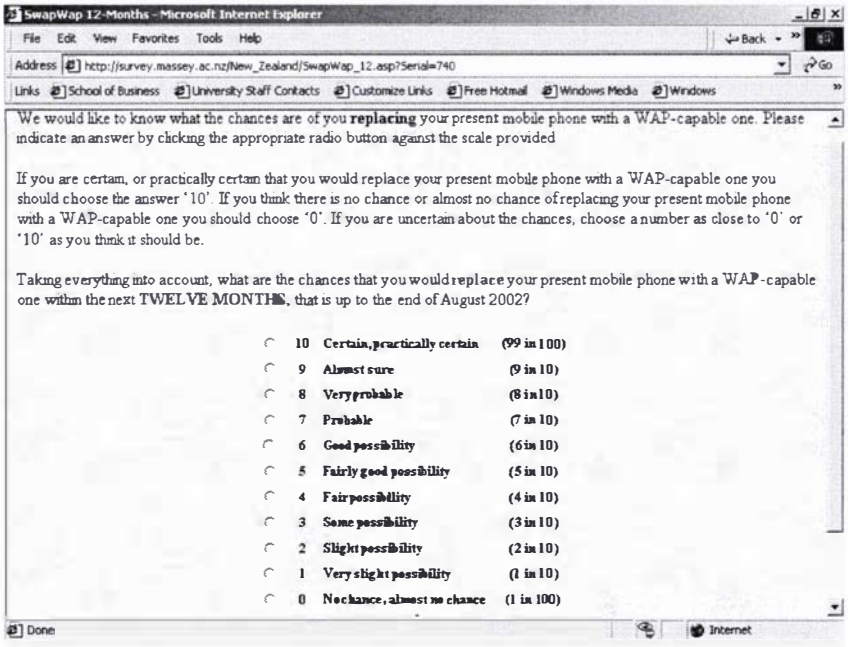

SwapWap 12-Months - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://survey.massey.ac.nz/New_Zealand/SwapWap_12.asp?Serial=740
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
We would like to know what the chances are of you replacing your present mobile phone with a WAP-capable one. Please indicate an answer by clicking the appropriate radio button against the scale provided.
If you are certain, or practically certain that you would replace your present mobile phone with a WAP-capable one you should choose the answer '10'. If you think there is no chance or almost no chance of replacing your present mobile phone with a WAP-capable one you should choose '0'. If you are uncertain about the chances, choose a number as close to '0' or '10' as you think it should be.
Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next TWELVE MONTHS, that is up to the end of August 2002?
10 Certain, practically certain (99 in 100)
9 Almost sure (9 in 10)
8 Very probable (8 in 10)
7 Probable (7 in 10)
6 Good possibility (6 in 10)
5 Fairly good possibility (5 in 10)
4 Fair possibility (4 in 10)
3 Some possibility (3 in 10)
2 Slight possibility (2 in 10)
1 Very slight possibility (1 in 10)
0 No chance, almost no chance (1 in 100)
Done
Internet

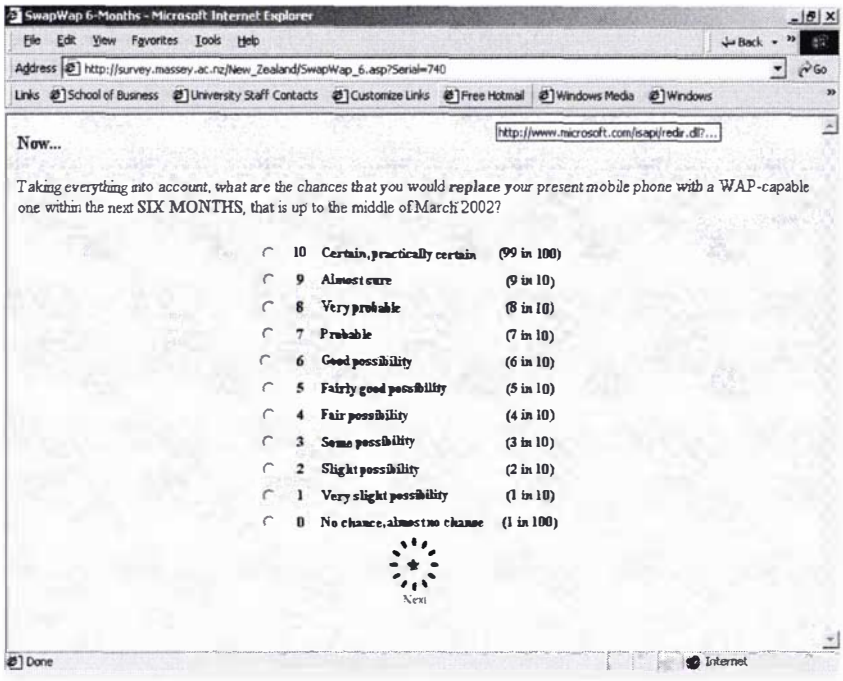


Appendix 11.3.3 Search Engine Treatment




Presentation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://survey.massey.ac.nz/New_Zealand/SearchEngine/PresentationFrames.asp?Serial=741
Go Links
Continue Survey
brand
Search
Displaying Page 1 of 2
Total Records Found : 10
Search Results for : brand
1) Philips Xenium
Voice or a dedicated short-cut @ key will launch you into the world of mobile internet.
2) Philips Ozeo
Gives more freedom and mobility. Ozeo Wap uses phone.com release 4.1 browser software based on the WAP forum standards.
3) Philips Fisio
With just one click on the @ softkey you can access web services such as e-mail,
Information about WAP-capable mobile phones
Please type in a keyword into the textbox and click the "Search" button on your left.
Please select items from the search results that will be listed below the search engine for viewing by double clicking the links.
When you've finished viewing the information, please return to the questionnaire via the "Continue Survey" link placed above the search engine.
Internet

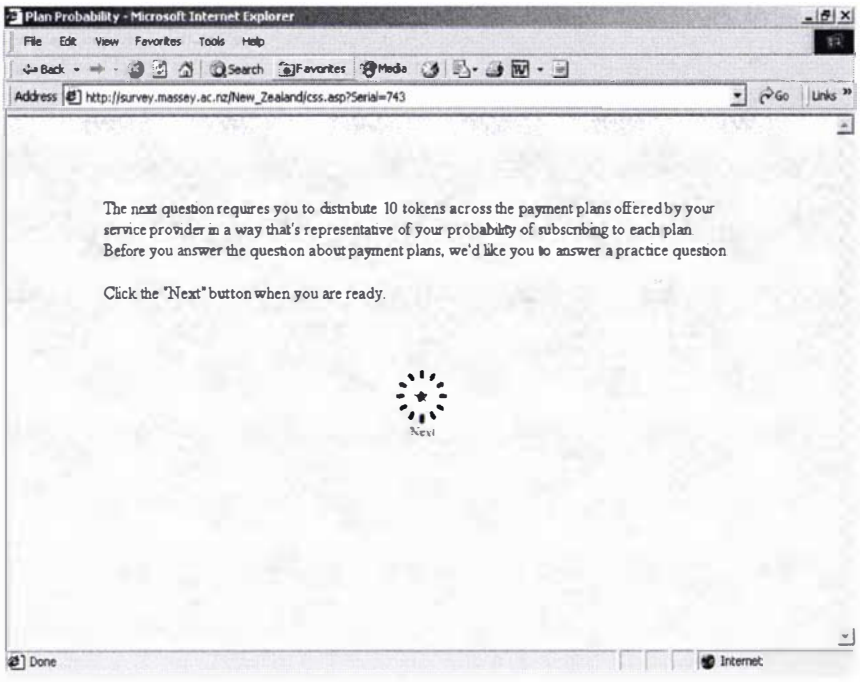

SwapWap 12-Months - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/SwapWap_12.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
We would like to know what the chances are of you replacing your present mobile phone with a WAP-capable one. Please indicate an answer by clicking the appropriate radio button against the scale provided
If you are certain, or practically certain that you would replace your present mobile phone with a WAP-capable one you should choose the answer '10'. If you think there is no chance or almost no chance of replacing your present mobile phone with a WAP-capable one you should choose '0'. If you are uncertain about the chances, choose a number as close to '0' or '10' as you think it should be.
Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next TWELVE MONTHS, that is up to the end of August 2002?
10 Certain, practically certain (99 in 100)
9 Almost sure (9 in 10)
8 Very probable (8 in 10)
7 Probable (7 in 10)
6 Good possibility (6 in 10)
5 Fairly good possibility (5 in 10)
4 Fair possibility (4 in 10)
3 Some possibility (3 in 10)
2 Slight possibility (2 in 10)
1 Very slight possibility (1 in 10)
0 No chance, almost no chance (1 in 100)
Done
Internet

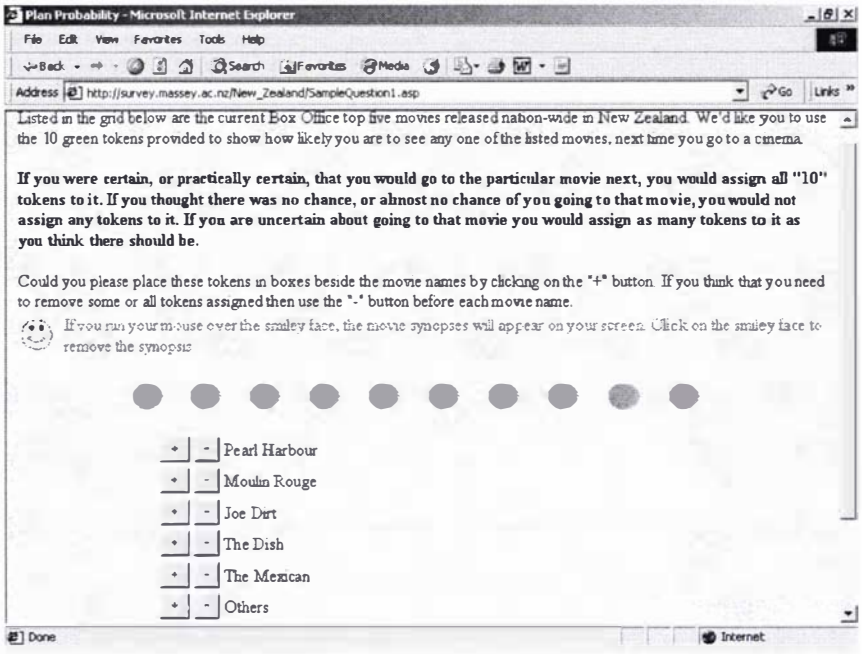

SwapWap 6-Months - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address http://survey.massey.ac.nz/New_Zealand/SwapWap_6.asp?Serial=740
Go
Links School of Business University Staff Contacts Customize Links Free Hotmail Windows Media Windows
http://www.microsoft.com/isapi/redir.dll?...
Now...
Taking everything into account, what are the chances that you would replace your present mobile phone with a WAP-capable one within the next SIX MONTHS, that is up to the middle of March 2002?
10 Certain, practically certain (99 in 100)
9 Almost sure (9 in 10)
8 Very probable (8 in 10)
7 Probable (7 in 10)
6 Good possibility (6 in 10)
5 Fairly good possibility (5 in 10)
4 Fair possibility (4 in 10)
3 Some possibility (3 in 10)
2 Slight possibility (2 in 10)
1 Very slight possibility (1 in 10)
0 No chance, almost no chance (1 in 100)
Next
Done
Internet

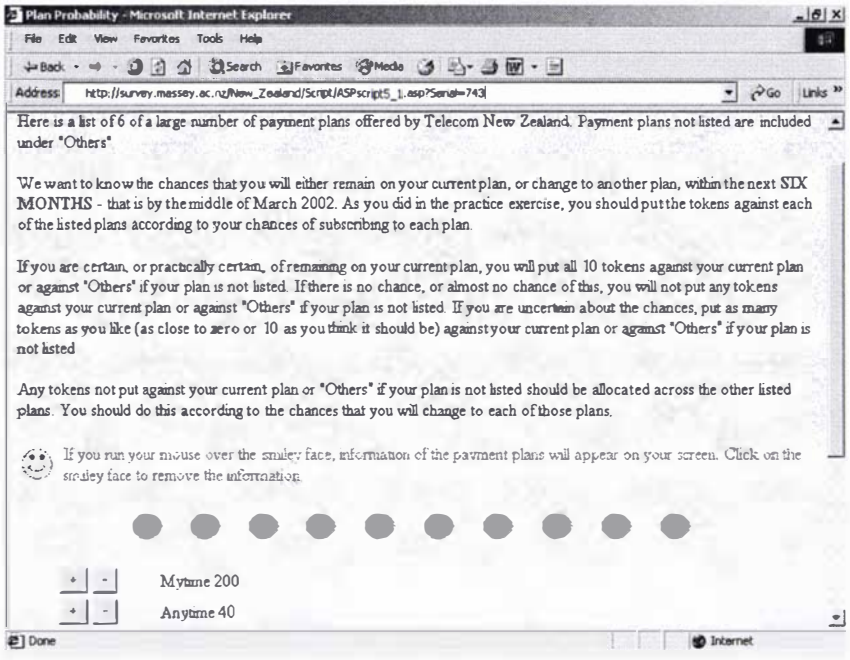
Appendix 11.4

Mutually Exclusive Behaviour Treatments

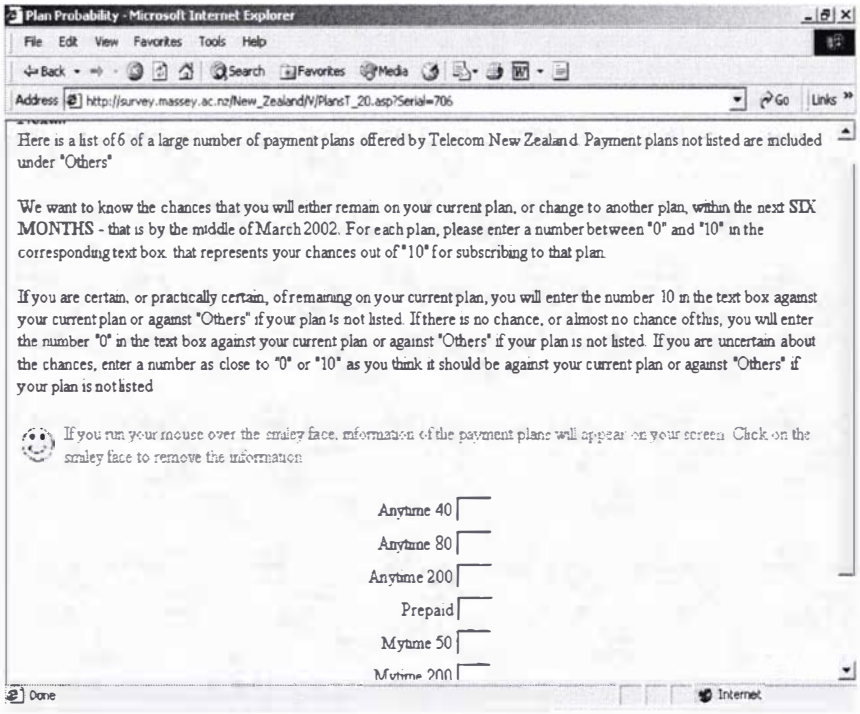
Appendix 11.4.1 Constant Sum Scale Treatment







Appendix 11.4.2 Weighted-Scores Treatment



Appendix 11.5

Internet-Based Surveys

Supplementary Tables

Supplementary T-test Tables of Table 6.9 (Mean Age and Proportion of Male Participants in the Actual and Final Samples)

Group Statistics

	Samples	N	Mean	Std. Deviation	Std. Error Mean
GENDER	Final Sample	696	.451	.4980	.0189
	Actual Sample	373	.461	.4992	.0258
AGE	Final Sample	703	48.659	15.4468	.5826
	Actual Sample	380	44.782	13.7839	.7071

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
GENDEF	Equal variance assumed	.360	.549	-.312	1067	.755	-.010	.0320	-.0727	.0528
	Equal variance not assumed			-.312	759.069	.755	-.010	.0320	-.0728	.0529
AGE	Equal variance assumed	9.285	.002	4.091	1081	.000	3.877	.9477	2.0174	5.7367
	Equal variance not assumed			4.232	855.390	.000	3.877	.9162	2.0788	5.6753

Supplementary T-test Tables of Table 6.13 (Comparisons of Mean Probability Scores of WAP-Capable Phones between Internet and Mail Participants for Non-Mobile Phone Users)

Group Statistics

	Mail survey and Control treatment	N	Mean	Std. Deviation	Std. Error Mean
12 months purchas probability data	Mail Survey	127	.0811	.18071	.01604
	Standard Treatme	66	.1091	.18952	.02333
6 months purchase probability data	Mail Survey	126	.0730	.19075	.01699
	Standard Treatme	66	.0788	.13532	.01666

Independent Samples Test

		Levene's Test for Equality of Variance		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	ig. (2-tailed	Mean Difference	Std. Error Difference	Lower	Upper
12 months purc probability data	Equal varian assumed	2.061	.153	-1.004	191	.317	-.0280	.02788	-.08299	.02701
	Equal varian not assumed			-.989	126.383	.325	-.0280	.02831	-.08401	.02803
6 months purch probability data	Equal varian assumed	.504	.479	-.219	190	.827	-.0058	.02641	-.05786	.04632
	Equal varian not assumed			-.243	173.170	.809	-.0058	.02380	-.05274	.04120

Supplementary T-test Tables of Table 6.14 (Comparisons of Mean Probability Scores of WAP-Capable Phones between Internet and Mail Participants for Mobile Phone Users)

Group Statistics

	Mail survey and Control treatmer	N	Mean	Std. Deviation	Std. Error Mean
12 months purcha probability data	Mail Survey	189	.1820	.27559	.02005
	Standard Treatm	206	.3252	.33546	.02337
6 months purchas probability data	Mail Survey	189	.1016	.21964	.01598
	Standard Treatm	206	.2204	.30359	.02115

Independent Samples Test

		Levene's Test for quality of Variance		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	ig. (2-tailed	Mean Difference	Std. Error Difference	Lower	Upper
12 months purcl probability data	Equal varian assumed	21.294	.000	-4.613	393	.000	-.1432	.03105	-.20428	-.08219
	Equal varian not assumed			-4.652	388.373	.000	-.1432	.03079	-.20377	-.08269
6 months purch probability data	Equal varian assumed	31.154	.000	-4.422	393	.000	-.1188	.02687	-.17162	-.06598
	Equal varian not assumed			-4.482	373.183	.000	-.1188	.02651	-.17092	-.06668

Appendix 11.6

Context of the Juster Scale

Supplementary Tables

Supplementary ANOVA Tables of Table 7.1 (Vodafone Survey: Comparison of Mean Probability Scores Obtained in the Treatments for the Twelve Months-Probability Data)

Descriptives

12 months purchase probability data

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Standard Treatment	167	.4072	.34512	.02671	.3545	.4599	.00	1.00
Point and Click Treatm	139	.3899	.33693	.02858	.3334	.4464	.00	1.00
Search Engine Treatm	154	.4117	.32297	.02603	.3603	.4631	.00	1.00
Total	460	.4035	.33475	.01561	.3728	.4341	.00	1.00

ANOVA

12 months purchase probability data

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.038	2	.019	.170	.844
Within Groups	51.396	457	.112		
Total	51.434	459			

Supplementary ANOVA Tables of Table 7.2 (Vodafone Survey: Comparison of Mean Probability Scores Obtained in the Treatments for the Six Months-Probability Data)

Descriptives

6 months purchase probability data

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Standard Treatment	167	.2587	.31379	.02428	.2107	.3066	.00	1.00
Point and Click Treatm	139	.2432	.31046	.02633	.1911	.2952	.00	1.00
Search Engine Treatm	154	.2370	.29193	.02352	.1905	.2835	.00	1.00
Total	460	.2467	.30510	.01423	.2188	.2747	.00	1.00

ANOVA

6 months purchase probability data

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.040	2	.020	.215	.807
Within Groups	42.685	457	.093		
Total	42.725	459			

Supplementary ANOVA Tables of Table 7.3 (Mobile Phone Users: Comparison of Mean Probability Scores Obtained in the Treatments for the Twelve Months-Probability Data)

Descriptives

12 months-purchase probaility data

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Standard Treatment	103	.3252	.33628	.03313	.2595	.3910	.00	1.00
Point & Click Treatme	85	.2718	.31419	.03408	.2040	.3395	.00	1.00
Search Engine Treatr	106	.2736	.27888	.02709	.2199	.3273	.00	1.00
Total	294	.2912	.31011	.01809	.2556	.3268	.00	1.00

ANOVA

12 months-purchase probaility data

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.184	2	.092	.958	.385
Within Groups	27.993	291	.096		
Total	28.177	293			

Supplementary ANOVA Tables of Table 7.4 (Mobile Phone Users: Comparison of Mean Probability Scores Obtained in the Treatments for the Six Months-Probability Data)

Descriptives

6 months-purchase probaility data

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Standard Treatment	103	.2204	.30433	.02999	.1609	.2799	.00	1.00
Point & Click Treatment	85	.1588	.25321	.02746	.1042	.2134	.00	1.00
Search Engine Treatmen	106	.1670	.24639	.02393	.1195	.2144	.00	1.00
Total	294	.1833	.27043	.01577	.1523	.2144	.00	1.00

ANOVA

6 months-purchase probaility data

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.221	2	.110	1.515	.222
Within Groups	21.208	291	.073		
Total	21.428	293			

Supplementary ANOVA Tables of Table 7.5 (Mobile Phone Users: Comparison of Mean Probability Scores Obtained in the Treatments after Applying the Information Viewing Criterion for the Twelve Months-Probability Data

Descriptives

12 months-purchase probaility data

	N	Mean	td. Deviatio	Std. Error	% Confidence Interval f Mean ower Boun	pper Boun	Minimum	Maximum
Standard Treatmen	103	.3252	.33628	.03313	.2595	.3910	.00	1.00
Point & Click Treatr	53	.2774	.31540	.04332	.1904	.3643	.00	1.00
Search Engine Trea	39	.2897	.27796	.04451	.1996	.3798	.00	.90
Total	195	.3051	.31895	.02284	.2601	.3502	.00	1.00

ANOVA

12 months-purchase probaility data

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.092	2	.046	.449	.639
Within Groups	19.643	192	.102		
Total	19.735	194			

Supplementary ANOVA Tables of Table 7.6 (Mobile Phone Users: Comparison of Mean Probability Scores Obtained in the Treatments after Applying the Information Viewing Criterion for the Six Months-Probability Data)

Descriptives

6 months-purchase probaility data

	N	Mean	td. Deviatio	Std. Error	% Confidence Interval f Mean ower Boun	pper Boun	Minimum	Maximum
Standard Treatmer	103	.2204	.30433	.02999	.1609	.2799	.00	1.00
Point & Click Treat	53	.1679	.25327	.03479	.0981	.2377	.00	1.00
Search Engine Tre	39	.1949	.24810	.03973	.1144	.2753	.00	1.00
Total	195	.2010	.28009	.02006	.1615	.2406	.00	1.00

ANOVA

6 months-purchase probaility data

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.098	2	.049	.623	.537
Within Groups	15.122	192	.079		
Total	15.220	194			

Supplementary Kruskal-Wallis H Tables of Table 7.7 (Mobile Phone Users: Comparison of Mean Ranks Obtained in the Treatments after Applying the Information Viewing Criterion for the Twelve Months-Probability Data)

Ranks

	Treatments	N	Mean Rank
12 months-purchase probaility data	Standard Treatment	103	100.61
	Point & Click Treatment	53	92.64
	Search Engine Treatment	39	98.40
	Total	195	

Test Statistics[a,b]

	12 months-purchase probaility data
Chi-Square	.724
df	2
Sig.	.696

a. Kruskal Wallis Test

b. Grouping Variable: Treatments

Supplementary Kruskal-Wallis H Tables of Table 7.8 (Mobile Phone Users: Comparisons of Mean Ranks after Applying the Information Viewing Criterion for the Six Months-Probability Data)

Ranks

	Treatments	N	Mean Rank
6 months-purchase probaility data	Standard Treatment	103	99.35
	Point & Click Treatment	53	93.14
	Search Engine Treatment	39	101.04
	Total	195	

Test Statistics[a,b]

	6 months-purchase probaility data
Chi-Square	.623
df	2
Sig.	.732

a. Kruskal Wallis Test

b. Grouping Variable: Treatments

Supplementary Kruskal-Wallis H Tables of Table 7.10 (Non-Mobile Phone Users: Comparison of Mean Ranks Obtained in the Treatments for the Twelve Months-Probability Data)

Ranks

	Treatments	N	Mean Rank
12 months-purchase probaility data	Standard Treatment	33	40.30
	Point & Click Treatment	24	32.56
	Search Engine Treatment	24	50.40
	Total	81	

Test Statistics[a,b]

	12 months-purchase probaility data
Chi-Square	9.146
df	2
Asymp. Sig.	.010
Exact Sig.	.009
Point Probability	.000

a. Kruskal Wallis Test

b. Grouping Variable: Treatments

Supplementary Kruskal-Wallis H Tables of Table 7.11 (Non-Mobile Phone Users: Comparison of Mean Ranks Obtained in the Treatments for the Six Months-Probability Data)

Ranks

	Treatments	N	Mean Rank
6 months-purchase probaility data	Standard Treatment	33	45.06
	Point & Click Treatment	24	32.96
	Search Engine Treatment	24	43.46
	Total	81	

Test Statistics[a,b]

	6 months-purchase probaility data
Chi-Square	6.636
df	2
Asymp. Sig.	.036
Exact Sig.	.034
Point Probability	.000

a. Kruskal Wallis Test

b. Grouping Variable: Treatments

Appendix 11.7

Mutually Exclusive Behaviours

Supplementary Tables

Supplementary T-test Tables of Table 8.3 (Telecom Subscribers: Comparisons of Mean Probability Scores of Payment Plans Obtained in the Treatments)

Group Statistics

	Treatments	N	Mean	Std. Deviation	Std. Error Mean
Mytime 50	Constant sum scale	84	.0417	.13007	.01419
	Weighted-scores	83	.0495	.16947	.01860
Mytime 200	Constant sum scale	84	.0726	.19657	.02145
	Weighted-scores	83	.0418	.14105	.01548
Anytime 40	Constant sum scale	84	.0310	.10639	.01161
	Weighted-scores	83	.0284	.12188	.01338
Anytime 80	Constant sum scale	84	.0226	.07660	.00836
	Weighted-scores	83	.0144	.05112	.00561
Anytime 200	Constant sum scale	84	.1262	.29004	.03165
	Weighted-scores	83	.0620	.21890	.02403
Prepaid	Constant sum scale	84	.3690	.45891	.05007
	Weighted-scores	83	.5074	.48305	.05302
Others	Constant sum scale	84	.3369	.44823	.04891
	Weighted-scores	83	.2965	.45329	.04976

Supplementary T-test Tables of Table 8.5 (Telecom Subscribers: Comparisons of Mean Ranks of Payment Plans Obtained in the Treatments)

Ranks

	Telecom subscribers:	N	Mean Rank	Sum of Ranks
Mytime 50	Contant sum scale	84	49.54	4161.00
	Weighted-scores	21	66.86	1404.00
	Total	105		
Mytime 200	Contant sum scale	84	50.57	4248.00
	Weighted-scores	21	62.71	1317.00
	Total	105		
Anytime 40	Contant sum scale	84	50.68	4257.00
	Weighted-scores	21	62.29	1308.00
	Total	105		
Anytime 80	Contant sum scale	84	50.62	4252.00
	Weighted-scores	21	62.52	1313.00
	Total	105		
Anytime 200	Contant sum scale	84	52.80	4435.50
	Weighted-scores	21	53.79	1129.50
	Total	105		
Prepaid	Contant sum scale	84	52.49	4409.00
	Weighted-scores	21	55.05	1156.00
	Total	105		
Others	Contant sum scale	84	53.67	4508.00
	Weighted-scores	21	50.33	1057.00
	Total	105		

Test Statistics[a]

	Mytime 50	Mytime 200	Anytime 40	Anytime 80	Anytime 200	Prepaid	Others
Mann-Whitney U	591.000	678.000	687.000	682.000	865.500	839.000	826.000
Wilcoxon W	4161.000	4248.000	4257.000	4252.000	4435.500	4409.000	1057.000
Z	-3.402	-2.222	-2.567	-2.712	-.177	-.375	-.507
Asymp. Sig. (2-tailed)	.001	.026	.010	.007	.859	.708	.612
Exact Sig. (2-tailed)	.001	.025	.011	.007	.834	.711	.617
Exact Sig. (1-tailed)	.001	.017	.011	.007	.405	.353	.313
Point Probability	.000	.000	.000	.000	.001	.002	.002

a. Grouping Variable: Telecom subscribers: Mann Whitney test

Supplementary T-test Tables of Table 8.7(Vodafone Subscribers: Mean Probability Scores and Ranks of Payment Plans Obtained in the Treatments)

Group Statistics

	Treatments	N	Mean	Std. Deviation	Std. Error Mean
Get 70	Constant sum scale	58	.0655	.19961	.02621
	Weighted-scores	46	.0470	.16322	.02407
Get 200	Constant sum scale	58	.1241	.25499	.03348
	Weighted-scores	46	.1419	.30601	.04512
Daytime 40	Constant sum scale	58	.0690	.23857	.03133
	Weighted-scores	46	.0384	.15564	.02295
Daytime 80	Constant sum scale	58	.0276	.13610	.01787
	Weighted-scores	46	.0589	.20493	.03021
Daytime 200	Constant sum scale	58	.0741	.21890	.02874
	Weighted-scores	46	.0531	.18169	.02679
Prepay	Constant sum scale	58	.5241	.44890	.05894
	Weighted-scores	46	.5290	.46399	.06841
Others	Constant sum scale	58	.1155	.29783	.03911
	Weighted-scores	46	.1318	.34010	.05014

Supplementary T-test Tables of Table 8.10 (Vodafone Subscribers: Comparisons of Mean Ranks of Payment Plans Obtained in the Treatments)

Ranks

	Code split into Constant	N	Mean Rank	Sum of Ranks
Get 70	Constant sum scale	58	36.12	2095.00
	Weighted-scores	16	42.50	680.00
	Total	74		
Get 200	Constant sum scale	58	35.83	2078.00
	Weighted-scores	16	43.56	697.00
	Total	74		
Daytime 40	Constant sum scale	58	35.58	2063.50
	Weighted-scores	16	44.47	711.50
	Total	74		
Daytime 80	Constant sum scale	58	35.23	2043.50
	Weighted-scores	16	45.72	731.50
	Total	74		
Daytime 200	Constant sum scale	58	36.42	2112.50
	Weighted-scores	16	41.41	662.50
	Total	74		
Prepay	Constant sum scale	58	39.28	2278.00
	Weighted-scores	16	31.06	497.00
	Total	74		

Test Statistics[a]

	Get 70	Get 200	Daytime 40	Daytime 80	daytime 200	Prepay
Mann-Whitney U	384.000	367.000	352.500	332.500	401.500	361.000
Wilcoxon W	2095.000	2078.000	2063.500	2043.500	2112.500	497.000
Z	-1.426	-1.532	-2.208	-2.790	-1.201	-1.408
Asymp. Sig. (2-tailed)	.154	.126	.027	.005	.230	.159
Exact Sig. (2-tailed)	.156	.125	.024	.007	.270	.161
Exact Sig. (1-tailed)	.091	.068	.024	.007	.137	.081
Point Probability	.002	.001	.001	.000	.001	.001

a. Grouping Variable: Code split into Constant sum scale treatment and weighted treatment

Appendix 11.8

Conference Papers

Appendix 11.8.1

DATABASE DRIVEN WEB-BASED SURVEY APPROACH FOR FORECASTING ADOPTION OF NEW TECHNOLOGY BASED PRODUCTS

Mathew Parackal

Sherly Parackal

Paper presented at the 26th CIRET Conference, Taipei, October 2002

Abstract

The prospect of reaching a wide audience in many countries has led researchers to develop various Internet based survey approaches. Internet based surveys such as email survey, disk by e-mail survey, Html-form survey, and Computer assisted Web-based survey have become popular among commercial and academic researchers. Although the Internet technology is efficient to gather information from a large number of individuals, methods of collecting information from probabilistic samples are still in their infancy. In this paper a Web-based survey approach designed for forecasting the adoption of new communication and technology-based product is explained. It also reports some results of a pilot-study that implemented this approach.

The survey program was developed using the Active Server Pages (ASP) technology. The paper describes this technology, which helped the researcher to control the survey procedure from an Access database. The various survey tasks performed by the program and their advantages are also discussed.

The results of the pilot-study suggest that the approach discussed in this paper could be used for collecting information from probabilistic sample of Internet users. Comparing the results with other similar studies (Brennan et al 1999, Parackal & Brennan 1999) showed that there was a distinct step forward in Web-based survey participation by respondents. The results also suggest that Web-based surveys and combination of Web-based and mail surveys could be used for collecting survey data.

APPENDIX 11.8.2

INTERNET-BASED & MAIL SURVEY: A HYBRID PROBABILISTIC SURVEY APPROACH

Mathew Parackal

Paper presented at the Ninth Australian World Wide Web Conference, Hyatt Sanctuary Cove, Gold Coast, from 5th to 9th July 2003 Paper presented at the The Ninth Australian World Wide Web Conference, Hyatt Sanctuary Cove, Gold Coast, from 5th to 9th July 2003.

Abstract

The Internet has become a popular survey medium among marketing researchers. The current coverage of the Internet however prevents it from being used as the sole medium in probability surveys. This paper explains a hybrid survey approach that used the Internet and the postal system to collect data from a probability sample. The paper presents the rationale of the approach and reports the results of a study that implemented the approach on two sub-groups. The paper also reports on the representativeness of the survey participants, the response rate received for the two survey media, and the overall response rate.

APPENDIX 11.8.3

FORECASTING MUTUALLY EXCLUSIVE BEHAVIOUR USING THE JUSTER SCALE

Mathew Parackal

Accepted for presentation at the 24th International Symposium on Forecasting, Sydney, Australia 4-7 July 2004

Abstract

Forecasting the adoption of mutually exclusive behaviours such as competing brands and election results has immense value to marketers. The challenging part of this Juster Scale application is in getting respondents to give probability scores in relation to the available alternatives. Respondents in general tended to give probability scores, treating alternatives as being independent. Consequently such probability scores failed to reflect the purchase behaviour of the sample logically.

Researchers in the past used a weighting procedure to fix the above discrepancy. While the weighting procedure allowed the logical interpretation of results, investigations into its accuracy was not satisfactory. Another method used to collect probability data for a set of alternatives was the Constant sum scale. This method forced respondents to give probability scores that added up to a constant number (usually 100 or 10) to convey their purchase behaviour. While the two methods have individual benefits, no studies have actually compared if the forecasts produced by the two methods are statistically similar or not. Hence a research was carried out to investigate the above. In this presentation first the Juster application will be explained followed by the results of the research carried out.

APPENDIX 11.8.4

A STUDY INVESTIGATING THE CONTEXTUAL REQUIREMENT OF THE JUSTER SCALE

Mathew Parackal, Ron Garland and Tony Lewis

In the proceedings of the 2004 Australia New Zealand Academic Conference, Victoria University, Wellington, New Zealand.

Abstract

Researchers have employed the Juster Scale to collect purchase probability data with good success. Reviewing the Juster Scale studies however revealed considerable variations in the scale's performance. Some of these variations appeared to be caused by its context, which are discussed in this paper. Prior to addressing the issues raised in the review of literature it was necessary to enquire whether the Juster Scale required additional inputs to collect purchase probability data in a purchasing context. In this paper results of a study that investigated the contextual requirement of the Juster Scale are reported.